Processing MIMIC-III for Evaluation of Various Blood Pressure Estimation Models

Magdalena Kostoska[1], Ivan Kuzmanov[1], Bojana Koteska[1], Fedor Lehocki[2], and Ana Madevska Bogdanova[1]

[1] University Ss. Cyril and Methodius,
Faculty of Computer Science and Engineering,
Rugjer Boskovikj 16, 1000 Skopje, N. Macedonia
[2] Institute of Measurement Science,
Slovak Academy of Sciences and Faculty of Informatics and Information Technologies,
STU, Slovakia
magdalena.kostoska@finki.ukim.mk
ivan.kuzmanov@students.finki.ukim.mk
bojana.koteska@finki.ukim.mk
fedor.lehocki@stuba.sk
ana.madevska.bogdanova@finki.ukim.mk

Abstract. The development of non-invasive easily available blood pressure estimation methods using electrocardiogram - ECG and/or photoplethysmogram - PPG signals has gained increasing attention. However, there is a lack of consistency in the evaluation of these methods due to variations in the size and availability of data in published datasets. Our research involves retrieving, cleaning, and storing a portion of the MIMIC-III database for utilization in model training and testing. This paper outlines our methodology for processing the MIMIC-III database, along with the challenges encountered during the process.

Keywords: MIMIC-III · electrocardiogram · photoplethysmogram · blood pressure estimation · artificial neural network · deep learning.

1 Introduction

Noninvasive and continuous monitoring of blood pressure (BP) has gained significant research interest due to the high prevalence of hypertension and the increased availability of low-cost sensors. These techniques and methods measure blood pressure without the need for invasive procedures, such as inserting a catheter into an artery. The models aim to provide accurate and convenient alternatives to invasive blood pressure monitoring, making them suitable for routine clinical use and home monitoring. Some of these models utilize artificial intelligence techniques to develop a model for estimating BP using manually extracted features from photoplethysmogram (PPG) and electrocardiogram (ECG) signals.

To train and evaluate the aforementioned models, substantial datasets containing the signals they utilize, along with corresponding recorded blood pressure

values, are essential. Collection of such data is a big challenge, since it requires continuous monitoring and recording of signals and data using medically approved devices, patient/person consents, ethical committee consent, server storage and management methods, etc. A lot of papers rely on using datasets created by other studies. There are several available public datasets that contains some of the required signals and values.

In this paper we present the met challenges while using the publicly available MIMIC-III database [4], the processing methods we use, our supporting infrastructure, lessons learned and some directions for future utilization of this dataset.

The rest of the paper is organized as follows. An overview on similar papers or related researches is presented in Section 2. Section 3 gives overview of publicly available datasets often used for BP estimation methods. Section 4 briefly describes the MIMIC-III database. Section 5 describes the used process we use. The results and the discussion are presented in Section 6. The conclusion is presented in Section 7.

2 Related Work

Previous studies have investigated the relationship between blood pressure and other physiological signals, such as ECG and PPG. There are several papers that discuss public datasets for blood pressure estimation and utilize these datasets for research.

The paper by by S.G. Khalid et al. (2018) [7] compares different machine learning approaches for blood pressure estimation using publicly available PPG datasets. The raw PPG signals and their corresponding reference systolic and diastolic blood pressures (BPs) were extracted from the University of Queensland vital sign dataset, an online database. This database contained PPG waveforms from 32 cases, from which a total of 8133 high-quality signal segments (each lasting 5 seconds) were extracted. These segments were then subjected to preprocessing, including normalization of both width and amplitude. It evaluates the performance of various algorithms and discusses the challenges and opportunities in PPG-based blood pressure estimation.

The paper by W. Wang et al. (2023) [12] discusses a dataset called "PulseDB" that is derived from two existing datasets: MIMIC-III and VitalDB. PulseDB comprises three main components:

- A collection of 5,245,454 meticulously selected high-quality 10-second segments of Electrocardiogram (ECG), Photoplethysmogram (PPG), and Arterial Blood Pressure (ABP) waveforms. These segments were sourced from 5,361 subjects, carefully retrieved from both the MIMIC-III waveform database matched subset and the VitalDB database.
- Comprehensive subject identification and demographic details are included, enriching the dataset with valuable supplementary input features. These particulars can be harnessed to enhance the performance of Blood Pressure (BP) estimation models. Additionally, they serve the purpose of assessing

the models' applicability to subjects not previously encountered, gauging their generalizability.
– Precise positional information regarding the characteristic points within the ECG and PPG signals is also integrated into PulseDB. This data provides essential context for waveform analysis and interpretation, contributing to a more comprehensive understanding of the cardiovascular dynamics under consideration.The purpose of this dataset is to provide a resource for benchmarking and evaluating cuff-less blood pressure estimation methods.

The paper by C. E. Hajj et al. (2020) [2] use the Multiparameter Intelligent Monitoring in Intensive Care database to validate models using PPG signals. Their innovative method leverages PPG signals in combination with advanced recurrent network models, specifically Long Short-Term Memory and Gated Recurrent Units. Part of the process aimed to streamline the input feature vector by minimizing redundancy and complexity. Consequently, a refined set of features was meticulously identified, demonstrating superior performance in the precise estimation of blood pressure values.

The review paper by Gholamhosseini et al. (2018) [3] focuses on blood pressure estimation using photoplethysmography and discusses the publicly available datasets used in related studies. It provides an overview of the dataset characteristics and highlights the challenges and advancements in this field.

The paper by J. Lee et al. (2019) [10] uses part of the VitalDB data bank and analyzed to assess the efficacy of PAT, PTT, and confounding factors on the estimation of BP. The dataset employed in this study underwent a stringent selection process guided by specific exclusion criteria. Initially, the recordings were meticulously scrutinized to ensure the presence of Electrocardiogram (ECG), Photoplethysmogram (PPG), and Arterial Blood Pressure (ABP) data. Subsequently, recordings with durations less than 30 minutes were excluded from consideration. Furthermore, a thorough visual inspection was carried out to identify instances of saturation and other distortions. Such identified anomalies were either excised from the recordings or led to the removal of the entire recording. To maintain data integrity, recordings showcasing an average heart rate below 40 beats per minute (BPM) or exceeding 200 BPM, as determined by the Pan-Tompkins algorithm, were also eliminated. Lastly, a minimum threshold of 100 cardiac cycles for analyzable data was set after the extraction of features. Consequently, a total of 2309 recordings were meticulously chosen from an initial pool of 6388 recordings, forming the basis of the study's dataset.

3 Public datasets for blood pressure estimation

There are several public datasets available for blood pressure estimation, which have been widely used for research and development of blood pressure prediction models. Here are a few notable examples:

– PhysioNet's MIMIC-II Dataset [9]: PhysioNet, a platform dedicated to sharing physiological signal data, hosted a older version of the MIMIC database.

It is also a publicly available dataset that contains de-identified electronic health record (EHR) data from patients admitted to the intensive care units (ICUs) of a large academic medical center.
- MIMIC-III Dataset [4]: The MIMIC-III dataset contains electronic health records from ICU patients. It includes blood pressure measurements, along with other physiological signals and clinical information. Researchers have utilized this dataset for blood pressure estimation studies and to develop predictive models for monitoring and managing hypertension.
- UCI Machine Learning Repository [6]: The UCI Machine Learning Repository hosts various datasets that can be used for blood pressure estimation. For example, the "Physiological Dataset for Wearable Blood Pressure Estimation" includes physiological signals, such as photoplethysmogram (PPG) and electrocardiogram (ECG), along with corresponding blood pressure measurements. Other interesting dataset is "Cuff-Less Blood Pressure Estimation Data Set" [5].These datasets are useful for developing wearable device-based blood pressure estimation algorithms.
- PPG-DaLiA Dataset [11]: The PPG-DaLiA dataset is a publicly available dataset for photoplethysmogram-based blood pressure estimation and it is a part of the Monash, UEA & UCR time series regression repository. It contains PPG signals collected from participants using smartphone-based sensors, along with reference blood pressure measurements. The dataset facilitates the development and evaluation of PPG-based blood pressure estimation algorithms.
- Framingham Heart Study Dataset [1]: The Framingham Heart Study is a long-term study focused on cardiovascular health. The dataset includes blood pressure measurements, as well as a wealth of other clinical and demographic information from participants. It has been used in various studies related to cardiovascular health, including blood pressure estimation and prediction.
- VitalDB Dataset [8]: is an openly accessible public dataset comprising intraoperative vital signs and biosignals collected by the Department of Anesthesia at Seoul National University Hospital. The data is captured using the Vital Recorder program and includes raw waveforms of arterial blood pressure (ABP), electrocardiogram (ECG), and photoplethysmogram (PPG) obtained from a commercial patient monitor device.

These datasets provide researchers with opportunities to develop and evaluate Machine Learning or Deep Learning models for blood pressure estimation using different inputs, such as physiological signals, clinical data, or a combination of both. It's important to review the specific documentation and terms of use for each dataset to ensure compliance with any restrictions or guidelines.

4 MIMIC-III

MIMIC-III (Medical Information Mart for Intensive Care III) [4] is a publicly available dataset that contains de-identified electronic health record (EHR) data

of patients admitted to the intensive care units (ICUs) of a large academic medical center. It is widely used by researchers and data scientists in the medical field for various studies and projects.

Scope: MIMIC-III includes data from over 46,000 adult patients admitted to the ICUs at Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA, between 2001 and 2012. The dataset covers a wide range of clinical information, including demographic details, vital signs, laboratory measurements, medications, procedures, diagnoses, and survival outcomes.

Data Categories: The dataset is organized into several categories, including patients, admissions, diagnoses, procedures, prescriptions, chart events, and more. These categories capture different aspects of a patient's medical history, allowing researchers to explore and analyze various dimensions of intensive care.

Data Types: MIMIC-III contains structured data, such as tables and relational databases, as well as unstructured data in the form of clinical notes and free-text reports. This combination of structured and unstructured data provides a rich source of information for researchers interested in natural language processing and text mining applications.

De-Identification: To protect patient privacy, the dataset has undergone a rigorous de-identification process. Personally identifiable information, such as names, addresses, and social security numbers, has been removed or anonymized. However, it still retains the clinical details necessary for research purposes.

Research Opportunities: The MIMIC-III dataset offers extensive research opportunities across a wide range of medical disciplines. Researchers can use the dataset to investigate clinical outcomes, develop predictive models, study disease patterns, assess treatment effectiveness, and explore healthcare utilization patterns, among other applications.

Access and Usage: MIMIC-III is freely available to researchers, provided they meet certain requirements and complete the required data use agreement. Access to the dataset helps ensure responsible use and adherence to privacy regulations and ethical guidelines.

5 Materials and Methods

In this section we describe the dataset and our process for data retrieval, cleaning and storage.

5.1 Dataset

The MIMIC-III dataset is organized into several data categories, each capturing different aspects of a patient's medical history. It is quite large database containing of 67,830 sets of records corresponding to around 30,000 patients who were admitted to intensive care units. Here are the main data categories found in the MIMIC-III dataset:

Patients: This category provides demographic information about each patient, including their unique identifier, gender, date of birth, date of death (if applicable), and other relevant details.

Admissions: The admissions category contains information related to each patient's hospital admission, such as admission and discharge timestamps, admission type (emergency, elective, etc.), and details about the admission source and location.

Diagnoses: This category includes diagnostic information for each patient, such as International Classification of Diseases, Ninth Revision (ICD-9) codes for primary and secondary diagnoses. It helps researchers explore disease patterns and study the impact of specific diagnoses on patient outcomes.

Procedures: The procedures category records the medical procedures performed on patients during their hospital stay. It includes procedure codes and descriptions, allowing researchers to analyze treatment patterns and interventions.

Prescriptions: This category provides information about the medications prescribed to patients. It includes details such as drug codes, start and end timestamps, dosage, and administration routes.

Chart Events: Chart events contain time-stamped measurements and observations recorded during a patient's ICU stay. These include vital signs (such as heart rate, blood pressure, and respiratory rate), laboratory values, fluid balance, ventilator settings, and other clinical measurements.

Laboratory Measurements: This category focuses specifically on laboratory test results, such as blood tests, urine analyses, microbiology reports, and other diagnostic tests. It includes information about the specific tests performed, their timestamps, and the corresponding values.

Microbiology: The microbiology category provides details about microbiological cultures and sensitivities performed on patients, helping researchers investigate infectious diseases and antibiotic management.

Note Events: This category contains free-text clinical notes, including physician notes, nursing notes, discharge summaries, and other medical documents. These notes provide additional context and qualitative information about a patient's medical history.

5.2 Methods

In order to create clean database we can further use in our research, we retrieve, clean and store data from MIMIC-III database. We use Python and our supporting infrastructure in the process.

Having set the goal of our research, our interest is to primarily obtain the ECG, PPG and ABP signals available in the database, as well as demographic data related to the signals and patients.

Due to the size of the database (approximately 6.7 TB and more than 67000 records) we process parts of the datasets (range of records) in few parallel processes. Due to our infrastructure we cannot run more that 4 or 5 parallel processes and each of them process up to 100 records. We manually run these processes (with set range of records) and monitor the execution.

Records processing Initially we have processed and stored whole records. We have realised that this is not practical of useful for our research, so we started process each of the records in segments. The record processing workflow is presented in Figure 1. For each segment we check where the data types we seek are available (ECG, PPG and ABP signals) and if the data is available then we further process the segment and data. If the data signals are of our minimum defined length (125 points at the moment, will increase to 1250 points in future), then part of the signals are further processed for null values and flat lines. Then each segment will be recorded to our server, along with other physiological parameters available in the segments (we may find further use of this parameters in the future). We also further process the data and extract features from the ECG and PPG signals and store this data in separate files (with name variation of the name of the files where we store the original data).

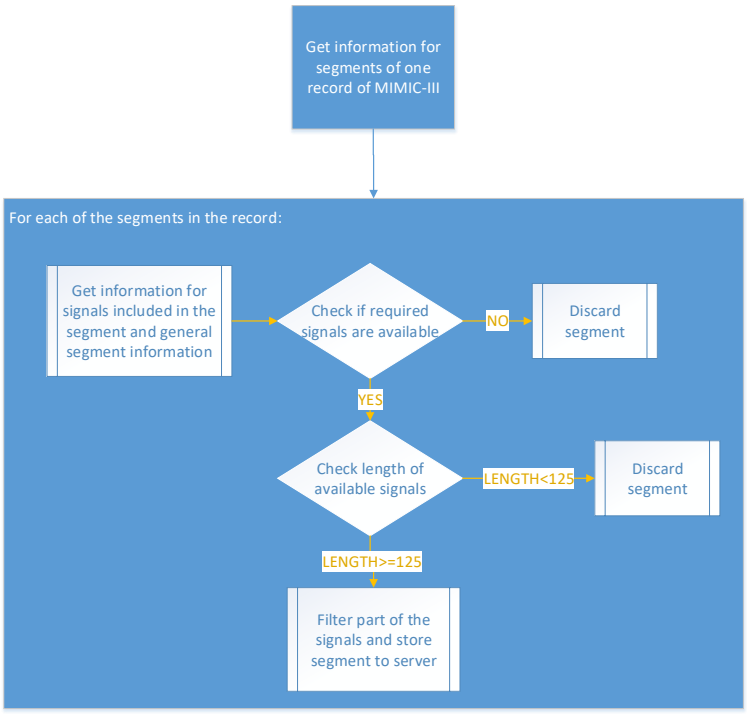

Fig. 1. Record processing workflow

We store the data on our server using python pickle. For practical reasons we have chosen to use Pickle format, for its ability to serialize and deserialize

almost any object using the Python programming language, its speed and size compression features.

Regarding the demographic data, in the MIMIC-III the names of the records are matched with the patients from whom the signals are retrieved. In our subset, each segment is named with a subject ID and and the number of the segment.

Supporting infrastructure We use two different instances of our local supporting infrastructure in this process.

We have dedicated storage server installed in our faculty, that is used by all the partners in the project, where we keep retrieved and processed data. We use owncloud platform for this server. The ownCloud software is installed on virtual machine. We use the ownCloud Community Edition, as we are small to medium organization and we 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.

Also, we use another virtual machine dedicated for the project. The machine currently has 32GB of RAM, 500GB storage space and 8 cores. The system is Linux based and we use it to run python scripts uninterruptedly to retrieve and process data from MIMIC-III and then to store the data to our storage server.

6 Results and Discussion

The popularity of MIMIC-III stems from its extensive data, large patient cohort, real-world relevance, and open accessibility, making it a valuable resource for researchers in the medical field.

Due to the scope of our research, only part of the database is relevant to us, but we cannot directly extract only the data we need. Thus, we needed the above described process, in order to obtain cleaned and high-quality data we can further use in our research.

Filtering and processing 6.7 TB of data is a big challenge and requires a lot of effort, time and infrastructure.

So far we have managed to process over 2500 records and around 1.5TB of data. Initially we have processed whole records and stored 510 records with 441 GB data. We have also created data dictionary regarding data quality of the signals, so the other teams can choose whether or not to work with certain records.

We have also realised that keeping full record is not practical or useful for our research. After detailed analysis we concluded that full length signal data is not usable in the given format, since a lot of the signal had missing or null values in given segments. We needed more focused and filtered value. So we started process each of the records in segments. The database is now transformed into segments. Thus we extracted the viable and high quality segments and we obtained 165.3 GB of data in 6045 data segments, from 208 different patients.

Figure 2 depicts the amount of data we have already processed (in GB), the initial amount of data we have kept for each record and the final amount of data

we are now keeping for segments we have filtered. We can conclude that so far we have kept only 11% of the processed data.

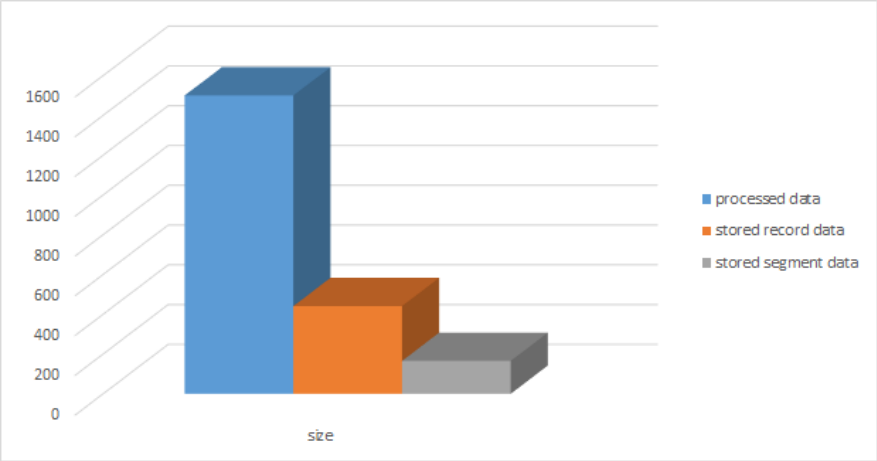

Fig. 2. Amount of data processed and stored

We have faced several challenges in the MIMIC-III processing that would be of interest for future users of MIMIC-III and should plan accordingly:

- **Detailed analysis of MIMIC-III structure**: While initial analysis of the structure of MIMIC-III in necessary to begin with processing, and there are publicly available codes and examples, we would recommend in-depth analysis, including statistical analyses and visualisation/plotting of downloaded data. Working with other publicly available codes requires there detailed testing of the code functionality.
- **Definition of data of interest**: Selection of data that will be of interest for the research and focusing on processing that type of data.
- **Resources**: Provisioning for resources, especially processing resources. The processing of data is a lengthy and demanding process (some of the records and segments are very big). If a further processing of data is needed (beside initial features extractions and similar processing) a large enough available storage will be needed.

7 Conclusion

MIMIC-III represents one of the most popular public databases due to rich and comprehensive data, large patient cohort, longitudinal nature and real-world clinical data, as well as the open access.

This paper discusses the encountered challenges in utilizing the publicly available MIMIC-III database, the processing methods we have employed and the

supporting infrastructure we have used. It relates the challenges we have met and gives and potential directions for future utilization of this dataset.

Our research is focused on blood pressure category estimation, given ECG and PPG signals, and we use ABP signal for validation. Our goal was to obtain the records containing all of these signals. We have started with records collecting and later found that keeping segments is more practical and less storage consuming.

So far we have processed around 1.5TB of data and we have found useful, for our project, only 11% of the data. We have fully utilized our dedicated resources and we have paralyzed the process.

Our goal for the future is to continue with processing the rest of the database and to publish the cleaned dataset, along with the computed features for the signals.

Acknowledgment

This paper has been written thanks to the support of the "Smart Patch for Life Support Systems" - NATO project G5825 SP4LIFE and by the National project IBS4LIFE of Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

References

1. Bhardwaj, A.: Framingham heart study dataset (2022). https://doi.org/10.34740/KAGGLE/DSV/3493583, `https://www.kaggle.com/dsv/3493583`
2. El Hajj, C., Kyriacou, P.A.: Cuffless and continuous blood pressure estimation from ppg signals using recurrent neural networks. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 4269–4272 (2020). https://doi.org/10.1109/EMBC44109.2020.9175699
3. Gholamhosseini, H., Baig, M.M., Rastegar, S., Lindén, M.: Cuffless blood pressure estimation using pulse transit time and photoplethysmogram intensity ratio. In: pHealth. pp. 77–83 (2018)
4. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific Data **3**(1), 160035 (May 2016). https://doi.org/10.1038/sdata.2016.35, `https://doi.org/10.1038/sdata.2016.35`
5. Kachuee, M., Kiani, M.M., Mohammadzade, H., Shabany, M.: Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. In: 2015 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1006–1009 (2015). https://doi.org/10.1109/ISCAS.2015.7168806
6. Kachuee, M., Kiani, M.M., Mohammadzade, H., Shabany, M.: Cuff-less blood pressure estimation data set (2015), `https://archive.ics.uci.edu/ml/datasets/Cuff-Less+Blood+Pressure+Estimation`

7. Khalid, S.G., Zhang, J., Chen, F., Zheng, D., et al.: Blood pressure estimation using photoplethysmography only: comparison between different machine learning approaches. Journal of healthcare engineering **2018** (2018)
8. Lee, H.C., Jung, C.W.: Vital recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. Scientific reports **8**(1), 1–8 (2018)
9. Lee, J., Scott, D.J., Villarroel, M., Clifford, G.D., Saeed, M., Mark, R.G.: Open-access MIMIC-II database for intensive care research. Annu Int Conf IEEE Eng Med Biol Soc **2011**, 8315–8318 (2011)
10. Lee, J., Yang, S., Lee, S., Kim, H.C.: Analysis of pulse arrival time as an indicator of blood pressure in a large surgical biosignal database: Recommendations for developing ubiquitous blood pressure monitoring methods. Journal of Clinical Medicine **8**(11) (2019). https://doi.org/10.3390/jcm8111773, `https://www.mdpi.com/2077-0383/8/11/1773`
11. Tan, C.W., Bergmeir, C., Petitjean, F., Webb, G.I.: Time series extrinsic regression. Data Mining and Knowledge Discovery pp. 1–29 (2021). https://doi.org/https://doi.org/10.1007/s10618-021-00745-9
12. Wang, W., Mohseni, P., Kilgore, K.L., Najafizadeh, L.: Pulsedb: A large, cleaned dataset based on mimic-iii and vitaldb for benchmarking cuff-less blood pressure estimation methods. Frontiers in Digital Health **4** (2023). https://doi.org/10.3389/fdgth.2022.1090854, `https://www.frontiersin.org/articles/10.3389/fdgth.2022.1090854`