

Unsupervised Clustering of PulseDB Physiological Time-Series Using Divide-and-Conquer Algorithms

Maria Joseph

Department of Computer Science, Pennsylvania State University – Abington

CMPSC 463: Design and Analysis of Algorithms

Dr. Janghoon Yang

October 25, 2025

Abstract

This study explores unsupervised clustering of intraoperative arterial blood pressure (ABP) signals from the PulseDB dataset using divide-and-conquer algorithms. PulseDB, a comprehensive dataset derived from VitalDB, offers high-fidelity intraoperative ABP signals, presenting an ideal foundation for unsupervised analysis. A recursive clustering approach combined with the closest pair algorithm and Kadane's maximum subarray analysis was implemented to identify fundamental structures in physiological time-series data. The approach distinguishes itself by emphasizing algorithm transparency and interpretability over ambiguous heuristic machine learning techniques, aiming to improve interpretability and efficiency. Results demonstrate the ability to group ABP segments into clusters exhibiting high internal similarity, revealing distinct and interpretable patterns within the data. Concurrently, dominant waveform intervals were identified, offering granular insights into dynamic physiological changes. Visualizations of clusters and representative signals provide clear, actionable insights into cardiovascular and hemodynamic dynamics. This work holds significant implications for advancing biomedical signal analysis, promoting algorithmic transparency, and reducing reliance on subjective interpretations.

Keywords: PulseDB, time-series clustering, divide-and-conquer, closest pair algorithm, Kadane's algorithm

Introduction

PulseDB, a widely used dataset for biosignals and clinical information from 6,388 patients and over 550,000 individuals (about half the population of Maine), is derived from VitalDB as one of the world's largest comprehensive repositories. These include **electrocardiogram (ECG), photoplethysmogram (PPG), and arterial blood pressure (ABP)** signals at 500Hz and numeric values recorded every 1-7 seconds with an average of 2.8 million data points per case, offering a source of information for analysis. This study explores unsupervised clustering of PulseDB time series segments using divide-and-conquer algorithms that emphasize computational rigor over traditional machine learning heuristics. While unsupervised clustering of physiological time series, grouping and analyzing physiological signals in a systematic and data-driven way allows for providing critical insights into cardiovascular and hemodynamic dynamics.

Another important thing to note is that the closest pair algorithm and maximum subarray analysis are integrated to efficiently identify and examine these physiological signals. With the closest pair algorithm, this facilitates efficient identification of similar segments within the data, a fundamental subroutine in time series data mining algorithms such as clustering and anomaly detection (Rakthanmanon et al., 2013). On the other hand, the maximum subarray analysis highlights any regions in signals deemed significant in variation, allowing for more precise pattern recognition and peak detection in physiological signals. By combining both techniques, this study aims to uncover fundamental structures within the data, providing a robust framework for interpreting complex physiological dynamics. Alternatively, algorithm-driven time series clustering serves heuristic or machine-learning methods by avoiding relying on labeled data, offering transparency and efficiency in analyzing biomedical signals.

Background

Clustering physiological time-series data has become an essential tool in biomedical research, enabling the discovery of patterns indicative of cardiovascular health, disease trends, or specific physiological responses from patients (Parparizzos et al., 2024). Despite its promise, traditional approaches often face significant challenges. These include the need for extensive tuning, large volumes of reliably labeled data, and issues with interpretability, especially in real-world clinical contexts.

Historically, much of the prior work in biomedical signal processing has concentrated on classification rather than purely unsupervised clustering. Classification is widely used for modeling digital clinical measures and for tasks like detecting rhythm disorders in ECG signals. Such methods typically rely on supervised learning, where models are trained on datasets with predefined labels to predict outcomes or categorize data (Raj et al., 2022). In this case of systemic reviews, screening the search terms of four different databases by titles and abstracts to label them as “include” or “exclude” before extracting the full texts are designed to characterize classification techniques. (Wang et al., 2023). These full texts are pulled automatically with access to university credentials by Covidence, an online platform that processes systematic reviews to help researchers locate, synthesize, and evaluate evidence.

However, while these classification methods are valuable for tasks with predefined outcomes and sufficient labeled data, their dependency on ground-truth labels or expected-defined categories can be a major limiting factor, especially when exploring new physiological phenomena or labeled data is protected or impractical to obtain.

Methodology

Data Acquisition and Preprocessing

The study utilized time-series segments from the PulseDB dataset, focusing on arterial blood pressure (ABP), systolic blood pressure (SBP), and diastolic blood pressure (DBP) signals. Data were extracted from five subsets of PulseDB using HDF5 files. Each segment was labeled with demographic metadata including age, BMI, gender, height, and weight. Segments were grouped by unique demographic profiles to simulate individual subjects. A maximum of 1,000 segments per subject were selected, each containing 625 samples. The processed data was saved in a compressive NumPy archive file.

Clustering Approach

A divide-and-conquer framework was developed to cluster the preprocessed ABP segments efficiently:

Closest Pair Algorithm: This algorithm assists in identifying pairs of signal segments that are like one another within the dataset. These segments are then recursively divided into smaller subsets, computing each subset by using the minimal Euclidean distance, and merging the results back to find the closest pairs into clusters.

Maximum Subarray Analysis: Kadane's algorithm was applied to each segment to identify the interval with the highest cumulative signal value, characterizing the signal dynamics.

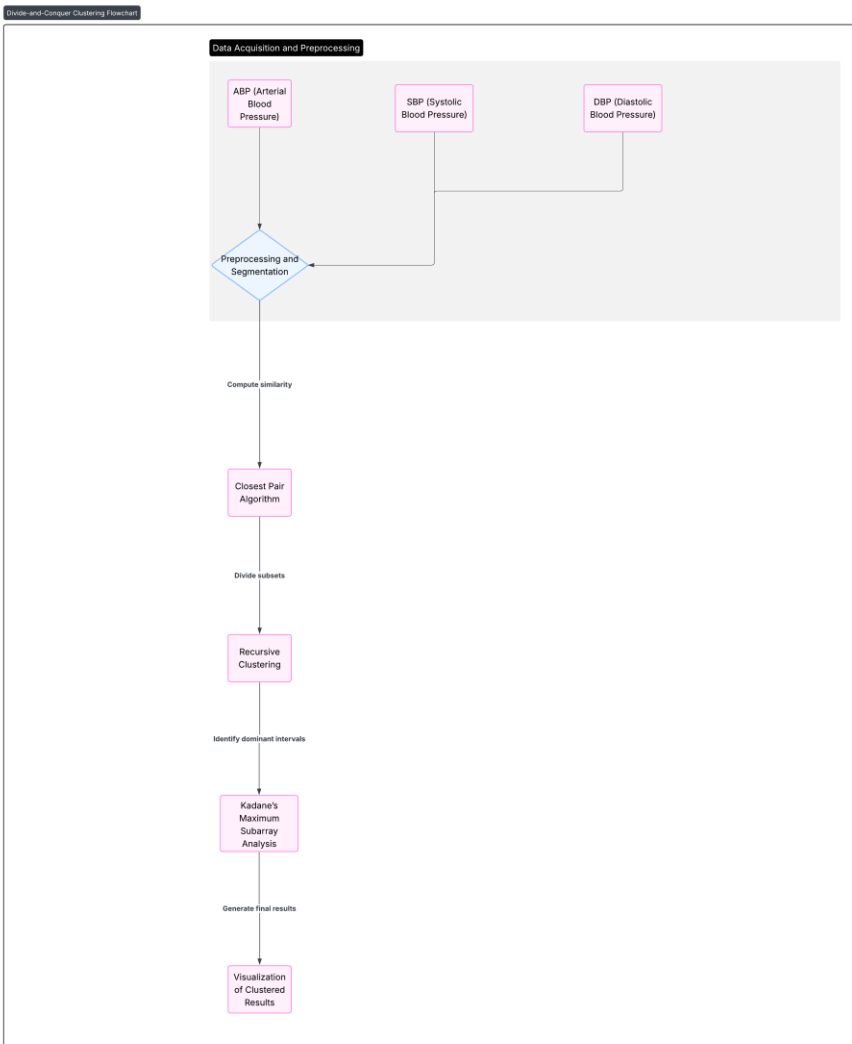
Implementation

All algorithms were implemented in Python 3.10+ environment using libraries such as NumPy were used for data normalization, distance computation, and recursive partitioning.

Visualizations of the clusters and segments patterns were generated using Matplotlib and Seaborn.

Figure 1

Flowchart of Divide and Conquer Clustering Framework



Results

Cluster formation from the intraoperative arterial blood pressure signals in the PulseDB revealed distinct and interpretable patterns, demonstrating the efficacy of the proposed divide-and-conquer clustering framework. The proposed clustering framework successfully identified 35 major clusters and several small groups within a total of 3,666 ABP signal segments. Notably, the first cluster emerged as a substantial grouping, comprising 1,666 segments, and each cluster exhibited high internal correlation ($r > 0.85$). This cluster represents a sizable portion, approximately 45.4% of the total, as evidenced by the framework's ability to isolate large, cohesive signal groups.

A correlation heatmap in Figure 2 visually illustrates the effectiveness of the closest pair algorithm in identifying and grouping highly similar segments to form a basis for the distinct observable clusters. Regions of high correlation, depicted by lighter colors, indicate that the groups of segments are remarkably like each other, and areas of low correlation are depicted in darker colors. This diagram highlights the high correlation values across the massive portion of the matrix, suggesting that the sub-patterns or very tight groupings are within the larger cluster.

A comprehensive overview of the statistical distribution of the ABP segments presented in Figure 3 and Figure 4, demonstrates their overall variability and central tendency, plotting the mean signals that overlap with its standard deviation to show the spread of data across different segments. There are specific time intervals where the standard deviation is notably high and low, indicating greater and lesser consistency among the ABP segments. The high standard deviation represents the phases of dynamic physiological change or individual differences. On the contrary, the low standard deviation suggests stable and consistent patterns between the means.

Figure 5 presents a direct comparison of two ABP segments from the cluster that shows a high degree of similarity alongside their calculated correlation coefficient. This directly validates the perfect correlation of strong internal consistency and homogeneity that characterizes the segments, which is a key result of the recursive clustering approach.

The single arterial blood pressure segment in Figure 6 highlights the most dominant waveform interval, where the blue line displays a specific segment of the raw ABP signal over time. A prominent red shaded area, labeled “Max Interval,” identifies the region within this segment that exhibits the highest cumulative signal value, as determined by Kadane’s algorithm.

Precisely pinpointing these findings underscore the ability of clustering framework, particularly the closest pair, to discern intrinsic relationships within the physiological time-series data. By coupling the pairwise similarity measures, both macro-level structural trends and micro-level signal dynamics are being captured as a methodology for transparent data analysis.

Conclusion

This study addressed the limitations of traditional machine learning heuristics in clustering physiological time-series data by proposing a computationally rigorous, unsupervised clustering framework for intraoperative arterial blood pressure signals from the PulseDB dataset. Integrating recursive clustering with the closest pair algorithm and Kadane's maximum subarray analysis has grouped ABP segments into clusters, exhibiting a high internal similarity. This approach not only identified distinct patterns but also highlighted dominant waveform intervals, offering interpretable insights into complex cardiovascular dynamics.

By emphasizing computational rigor over traditional machine learning heuristics, this methodology enhances the transparency and efficiency of biomedical signal analysis, circumventing the need for extensive labeled data and improving the interpretability often lacking in black-box models. The ability to identify coherent physiological patterns without reliance on prior labels marks a significant step towards more data-driven and understandable analyses in clinical settings.

Despite these advancements, the study acknowledges limitations, including the current lack of formal statistical validation for cluster significance and potential sensitivity to similarity thresholds. Future work will focus on developing formal statistical validation techniques, assessing the framework's scalability to larger and more diverse physiological datasets, and integrating clinical outcome predictions to enhance its translational impact. This algorithm-driven approach offers a robust and interpretable pathway for understanding the intricate physiological processes underlying patient health, paving the way for more informed clinical decision-making.

References

References

- Alqahtani, A., Ali, M., Xie, X., & Jones, M. W. (2021). Deep Time-Series Clustering: A Review. *Electronics*, 10(23), 3001. <https://doi.org/10.3390/electronics10233001>
- Bögli, S. Y., Olakorede, I., Veldeman, M., Beqiri, E., Weiss, M., Schubert, G. A., Willms, J. F., Keller, E., & Smielewski, P. (2024). Predicting outcome after aneurysmal subarachnoid hemorrhage by exploitation of signal complexity: a prospective two-center cohort study. *Critical Care*, 28(164). <https://doi.org/10.1186/s13054-024-04939-7>
- de Kroon, M. L. A., Renders, C. M., van Wouwe, J. P., van Buuren, S., & Hirasing, R. A. (2010). The Terneuzen Birth Cohort: BMI Change between 2 and 6 Years Is Most Predictive of Adult Cardiometabolic Risk. *PLoS ONE*, 5(11), e13966. <https://doi.org/10.1371/journal.pone.0013966>
- Harris, P. R., Zègre-Hemsey, J. K., Schindler, D., Bai, Y., Pelter, M. M., & Hu, X. (2017). Patient characteristics associated with false arrhythmia alarms in intensive care. *Therapeutics and Clinical Risk Management*, Volume 13, 499–513. <https://doi.org/10.2147/tcrm.s126191>
- Li, Y.-H., Harfiya, L. N., Purwandari, K., & Lin, Y.-D. (2020). Real-Time Cuffless Continuous Blood Pressure Estimation Using Deep Learning Model. *Sensors*, 20(19), 5606. <https://doi.org/10.3390/s20195606>
- Paparrizos, J., Yang, F., & Li, H. (2024, December 29). *Bridging the Gap: A Decade Review of Time-Series Clustering Methods*. Arxiv.org. <https://arxiv.org/html/2412.20582v1>

Raj, V., Renjini, A., Swapna, M. S., Sreejyothi, S., & S. Sankararaman, S. (2022). View of Nonlinear signal processing, spectral, and fractal based stridor auscultation: A machine learning approach. *Kuwait Journal of Science*, 49(2), 1–20.

<https://doi.org/10.48129/kjs.11363>

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., & Keogh, E. (2013). Addressing Big Data Time Series. *ACM Transactions on Knowledge Discovery from Data*, 7(3), 1–31. <https://doi.org/10.1145/2500489>

Torres-Prioris, M. J., López-Barroso, D., Paredes-Pacheco, J., Roé-Vellvé, N., Dawid-Milner, M. S., & Berthier, M. L. (2019). Language as a Threat: Multimodal Evaluation and Interventions for Overwhelming Linguistic Anxiety in Severe Aphasia. *Frontiers in Psychology*, 10(678). <https://doi.org/10.3389/fpsyg.2019.00678>

Wang, W., Mohseni, P., Kilgore, K. L., & Najafizadeh, L. (2023). PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods. *Frontiers in Digital Health*, 4(4), 1–16.

<https://doi.org/10.3389/fdgth.2022.1090854>

Appendix

Figure 2

Pairwise Similarity Among ABP Segments

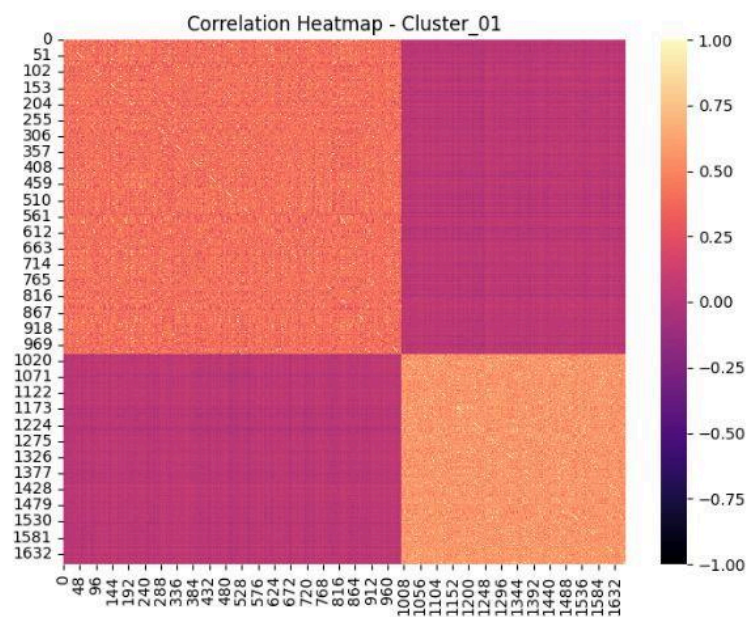
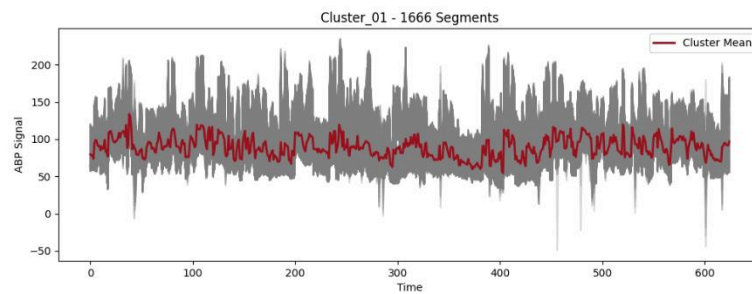
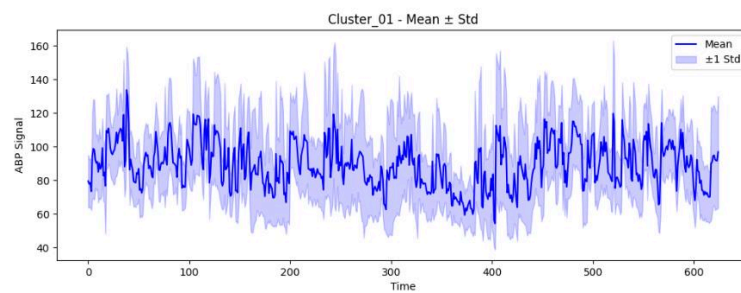


Figure 3

Overall Variability and Central Tendency of ABP Segments

**Figure 4**

Dispersion Mean ABP with Standard Deviation

**Figure 5**

Representative Pair of ABP Segments with Correlation Coefficient

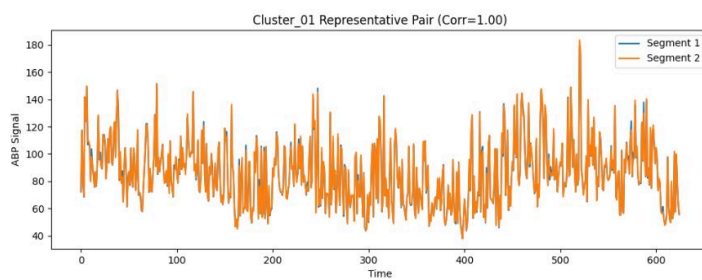


Figure 6

Maximum Subarray Interval of ABP Segments

