UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science Master Thesis

Medical Intervention Recognition in the PICU

First examiner:

Prof. dr. A.P.J.M. (Arno) Siebes

Second examiner:

Dr. A.J. (Ad) Feelders

Daily Supervisors:

Ruben Zoodsma

PhD. L.M. van Loon

Candidate:

ing. T.H.G. Tran (4671554)

In cooperation with:

UMC Utrecht

June 27, 2024

Abstract

Monitoring and interpreting vital signs in pediatric intensive care is crucial for timely medical interventions. Analyzing the multivariate time series data of these vital signs with data science techniques can significantly improve patient monitoring. This study focuses on using clustering algorithms to detect and classify medical events in the Pediatric Intensive Care Unit (PICU).

The main problem addressed is identifying and classifying medical events from the vital sign measurements of cardiac infants. Here we show that using a combination of Mini Batch K-means clustering and Random Forest classifier validation, significant health events can be effectively identified. The "Difference Baseline Cubed (10m - 5m)" approach achieved the highest accuracy of 0.966, showing the best performance in clustering vital sign data.

These findings highlight the potential of advanced clustering techniques to improve patient monitoring and intervention strategies in the PICU. By involving medical experts to validate the clustering results, the findings can be made more relevant to clinical needs. This approach not only enhances current patient care but also opens the door for real-time monitoring and automatic event detection in critical care settings.

In a broader context, the methods developed in this study can be used in other fields where analyzing multivariate time series data is important. Future research could look into using more advanced models and real-time applications, which could significantly change how pediatric critical care and other areas operate.

Contents

1. Introduction

In the Pediatric Intensive Care Unit (PICU), the effective monitoring and timely medical intervention for critically ill children is paramount. This chapter sets the stage for understanding the critical role of data science in enhancing patient care. It outlines the motivation behind the study, provides a comprehensive literature review on the current state of medical event detection in healthcare, and details the specific objectives that this research aims to achieve. By leveraging advanced clustering algorithms, this study seeks to improve the detection and classification of medical events, ultimately contributing to better patient outcomes in the PICU.

1.1 Motivation and Context

In the PICU, critically ill children with a variety of underlying conditions receive intensive care and monitoring. An array of parameters, including heart rate, blood pressure, and respiration, are meticulously monitored to gauge the patients' physiological status. In today's era of data science and artificial intelligence, this wealth of data presents an opportunity for advanced analysis and insight.

However, a change is on the horizon for PICU practice: patients will soon have individual rooms for privacy, which is beneficial for families but poses challenges for monitoring. Understanding patterns in vital signs becomes crucial for this new setup to ensure patients receive the care they need.

In this study, the aim is to develop methods to recognize variations in patient data related to medical interventions. Specifically, common medical interventions such as adjustments in inotropic support, which refers to medication that enhances the contractility of the heart, and supplemental oxygen provided to patients to improve oxygenation levels, will be examined. By

analyzing retrospective data from pediatric patients, we seek to enhance our understanding of when these medical interventions take place. Currently, such medical interventions are not consistently annotated over time, limiting the utility of this information for clinical decision-making and model development. Addressing this challenge will be essential for the future development of models and smart alarming algorithms for PICU patients.

To address these challenges, this thesis explores the potential of clustering techniques in detecting significant medical events from time series vital measurements. Clustering, an unsupervised learning method, can identify natural groupings in data without predefined labels, making it well-suited for uncovering patterns in physiological data that correspond to medical interventions.

1.2 Literature review

1.2.1 Vital Signs and Data Analysis

Vital signs, such as heart rate, blood pressure, respiratory rate, and oxygen saturation, are critical indicators of a patient's health status and are continuously monitored over time. These measurements are interdependent, and their analysis is complex due to high dimensionality, temporal dependencies, and noise. Noise and artifacts, including movement, pain, and coughing (Malani, 2012); (Van Lieshout et al., 1989), further complicate the data. Advanced techniques are required to extract meaningful patterns from this data, enabling accurate monitoring and event detection, which are crucial for patient care and clinical decision-making in healthcare settings.

1.2.2 Event Detection in Healthcare

Event detection can significantly impact healthcare, aiding in early disease diagnosis, patient monitoring, and treatment optimization. For instance, Srinivasulu et al. (2021) detected apnea events using machine learning techniques for the clinical diagnosis of Sleep Apnea Syndrome by analyzing time, frequency, and statistical features of electrocardiogram signals. Fur-

ther literature reveals that medical events such as seizures (Potter et al., 2022), heart disease (Nagavelli et al., 2022), and brain infarcts (Van Hespen et al., 2021) have been detected using machine learning algorithms.

1.2.3 Supervised and Unsupervised Learning

Machine learning algorithms used for event detection and anomaly detection can be either supervised or unsupervised. Supervised methods require pre-modeling knowledge of when events or anomalies occur, necessitating consistent annotation or expert labeling, which is labor-intensive and prone to variability. Despite this, supervised learning has been successful in detecting and classifying events, as shown by Nagavelli et al. (2022) and Van Hespen et al. (2021). Chen et al. (2016) also used a supervised model to classify real and artifact alerts, employing a random forest model to discern relevant alerts from artifacts.

1.2.4 Unsupervised Learning in Healthcare

Unsupervised models, which do not rely on labeled data, have also been explored. Kavitha et al. (2021) applied unsupervised clustering techniques, such as K-means and K-medoids, to detect anomalies in healthcare data extracted from wearables. This approach increases the efficiency of health services. Potter et al. (2022) also employed unsupervised learning for seizure detection, using the first time series transformer-based model, which successfully identified seizures and outperformed other supervised learning models. Both of these studies provide can provide reliability and support in the health care system.

Clustering algorithms, such as K-means and DBSCAN, have shown promise in uncovering hidden patterns and identifying significant deviations in physiological data. These clustering methods can group similar time series data points based on their features, allowing for the identification of natural groupings and anomalies that may correspond to medical events. The application of these techniques to vital sign data can enhance the detection of critical medical events by recognizing patterns that might not be evident

through supervised learning approaches alone. According to Kanagala and Krishnaiah (2016), a comparative study demonstrated the effectiveness of K-means and DBSCAN in various contexts, including medical data analysis, highlighting their utility in detecting meaningful patterns and anomalies in physiological datasets.

1.2.5 Conclusion

Overall the current application of machine learning in the loads of data that a health care system can provide, has aimed to support physicians to provide more effective treatment to patients (Habehh & Gohel, 2021).

1.3 Objectives

Vital parameters represent multivariate time series data in the field of data science. The primary research question is: Can medical events be detected and classified by analyzing vital measurements in cardiac infants? This study aims to employ clustering techniques to identify and categorize significant health events, thereby enhancing patient monitoring and intervention strategies in the PICU.

2. Data

This chapter delves into the data used for this research, sourced from the PICU at Wilhelmina Children's Hospital. It provides a thorough description of the dataset, which includes vital sign measurements from cardiac infants. The chapter discusses the initial exploration and preparation of the data, addressing issues such as missing values and anomalies. Additionally, it highlights the ethical and legal considerations involved in handling sensitive patient data, ensuring compliance with privacy regulations and the protection of patient identities.

2.1 Description of the Dataset

The data utilized for this study were collected at the PICU of Wilhelmina Children's Hospital, part of the University Medical Center (UMC) Utrecht, The Netherlands. It comprises time-series recordings of vital measurements from 68 cardio patients, all of whom were infants less than one year of age, from the year 1999 to 2020. As mentioned earlier, the data provided is a time series, with measurements recorded in consecutive time frames in seconds. Each patient was assigned a unique pseudo-ID, and their timelines begin at the point of admission to the PICU. The patients were admitted at different times, and the duration of their stays in the PICU varied. Detailed descriptions of the vital measurements that were available in the provided dataset can be found in Appendix A.1. These measurements include a range of "normal" values, which differ based on the age of the patients. These are all patients diagnosed with different cardiac conditions, which can lead to varying "normal" vital measurements. Additionally, their vital measurements are influenced by the initiation of medical interventions and are more susceptible to movement-related noise. While all vital measurements were possible for every patient in the dataset, an exception was the Near-Infrared

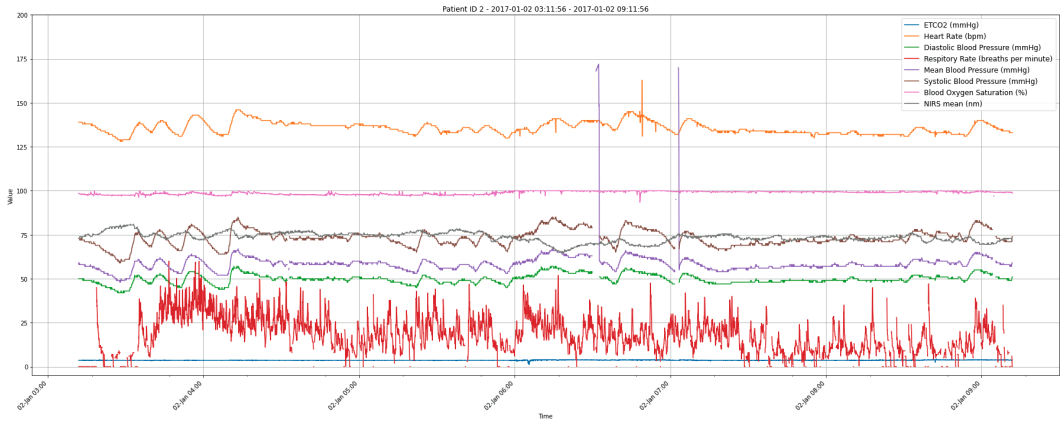

Figure 2.1: Example of Patients Vital Signs

Spectroscopy (NIRS) measurement. Due to the small size of the infants and the nature of the NIRS device, which involves a sticker placed on the head, it was sometimes feasible to fit only one sticker on smaller infants.

2.2 Data exploration

The dataset comprises high-frequency measurements recorded every second, resulting in a substantial volume of data. Specifically, it contains approximately 12 million rows and 9 columns, totaling around 108 million data points. Figure 2.1 illustrates the time series of measurements for three patients over a selected interval, highlighting daily variations and overall trends. During data analysis and visualization, numerous missing values (NAs) and anomalies were identified. Notably, anomalies in the blood pressure columns, where the mean blood pressure exceeded the systolic blood pressure, were classified as blood collection events.

2.3 Preparation of the Data

2.3.1 Missing Data Analysis

Missing data was observed sporadically throughout the dataset. Clustering, the main method used to detect events in this study, requires no missing values and a uniform number of features (vital sign measurements). Given that some patients were missing certain features, exclusion of these features

was neccesary

One of the NIRS measurements was sometimes limited to a single sensor on smaller infants, indicating that the missing data in NIRS measurements might be considered Missing at Random (MAR) due to its correlation with the size and condition of the infant. When assessing the number of patients missing both NIRS measurements, it was decided to exclude this measurement from the analysis. As shown in Table A.2, 12 patients lacked this NIRS measurement. Excluding the NIRS measurement allowed these patients to be included in the analysis.

Non-Invasive Blood Pressure (NIBP) measurements were also inconsistently recorded in short intervals, making them unsuitable for this study and thus excluded from the analysis. The inconsistent recording of NIBP measurements indicates that the missing data is likely Missing Not at Random (MNAR), possibly due to specific conditions or decisions made during data collection.

To ensure the clustering algorithms were applicable, only patients with the following measurements were included in the analysis: *mon_etco2*, *mon_hr*, *mon_ibp_dia*, *mon_ibp_mean*, *mon_ibp_sys*, *mon_rr*, and *mon_sat*. This selection resulted in a dataset comprising 39 patients. Although these measurements still had sporadic missing values, they were manageable and filling them in did not introduce significant bias. These were indicated by NaNs in the dataset. Since the measurements were continuous for each patient, no time points were missing. The NaNs were sporadic and varied in length, necessitating specific constraints to handle these data points efficiently. The conditions for handling missing data were as follows:

- If consecutive data points of a feature were missing for more than 10 seconds, the window was skipped.
- If fewer than 10 seconds of data were missing from a time window, the points were interpolated linearly.
- Exception: Time windows were adjusted for known events, such as blood collection, based on expert confirmation.

2.3.2 Windows and more preprocessing

After the missing data analysis, several additional steps were implemented to prepare the dataset for analysis. For clustering purposes, each patient's dataset was split into sliding windows. This approach captures a period of time rather than clustering the measurements per second. The windows were set to 10 minutes after experts opinions of the events being approximately this length of time.

Recognizable anomalies in the blood pressure feature, identified as blood collection events, were excluded from the windows to simplify the analysis. Other anomalies were retained to preserve as much information as possible, assuming some of these points could be medical interventions. The windows of measurements were then converted to arrays and scaled using MinMaxScaler() from scikit-learn. This scaling ensures all features are on a similar scale, which improves the performance of clustering algorithms. This approach reduces the impact of features with large ranges, enhances cluster separation, and makes the analysis easier to understand and more efficient.

2.3.3 Assumptions

- **Different interventions have different effects on a patient.** Each medical intervention is assumed to impact the patient's vital signs uniquely. For example, the administration of different medication or other treatments can cause distinct changes in measurements such as heart rate, blood pressure, and oxygen saturation.
- **Relative consistency of intervention effects across patients.** For the purposes of this study, it is assumed that the relative (percentage) change induced by a specific intervention is consistent across patients. This assumption is necessary to facilitate the clustering of similar events, such as the administration of epinephrine, based on their effects on vital signs. Without this assumption, the clustering algorithm would struggle to identify and group together similar events effectively. For instance, if an intervention typically results in a 10% increase in heart

rate, it is expected to cause a similar percentage change in all patients, regardless of their initial heart rate.

- **Baseline variability among patients.** It is assumed that each patient has a different "normal" range for vital signs due to inherent physiological differences and underlying health conditions. Therefore, transformations using the median of each specific vital measurement are employed to standardize data, making individual variations more comparable.
- **Independence of non-blood pressure features.** Except for the systolic, diastolic, and mean invasive blood pressure measurements, which are interdependent, all other features are assumed to be independent. This means that changes in one vital sign are not directly influencing changes in another, simplifying the analysis and interpretation of the data.
- **Impact of outliers and anomalies.** Recognizable anomalies, such as blood collection events, are excluded from the analysis to avoid skewing the results. However, other anomalies are retained to preserve the integrity of the data and ensure that significant variations and potential patterns are not overlooked.
- **Temporal consistency within sliding windows.** When splitting each patient's dataset into sliding windows for clustering, it is assumed that the vital sign patterns within each window are temporally consistent. It is assumed that an event can be captured in the time set window.
- **Scalability of features.** By scaling both transformed and non-transformed features using MinMaxScaler to a consistent range, typically [-1, 1], it is assumed that this will enhance the performance of clustering algorithms. This scaling approach is expected to mitigate the impact of features with large ranges, improve cluster separation, and make the analysis more interpretable and efficient.

2.4 Ethical and legal consideration of the data

Conducting research involving time series data of vital measurements from patients under the age of one year involves several critical ethical considerations. Ensuring the privacy and confidentiality of patient data is extremely important. All data used in this research was anonymized to protect the identities of the young patients, in compliance with the guidelines provided by UMC Utrecht.

Access to the data was strictly controlled and only permitted through a secure virtual machine environment to prevent unauthorized access and ensure data security. This measure is essential to protect sensitive health information from potential breaches.

Moreover, the potential implications of the research findings must be carefully considered. The aim is to contribute positively to the field of pediatric healthcare by identifying medical events in the vital measurements of infants, which could lead to improved monitoring and care. It is crucial to ensure that the outcomes of this research are used to enhance patient care and not to stigmatize or disadvantage any individuals.

3. Method

In this chapter, the methodologies employed in the study are meticulously detailed. It begins with a description of the chosen clustering techniques and the rationale behind their selection. The chapter then explains the original approach and the process of optimizing the number of clusters. Validation methods are also discussed to ensure the robustness of the clustering results. Furthermore, alternative approaches and specific event detection strategies are explored to enhance the accuracy and reliability of the findings.

3.1 Description of the method used

To capture the medical interventions within the patients measurements, clustering was chosen due to its proven success in event detection (Kanagala & Krishnaiah, 2016). However, clustering often involves extensive distance calculations, which can be computationally expensive, especially with large datasets. This computational burden made many clustering methods impractical for this study. In the context of data science, this involves determining how clustering algorithms can be utilized to detect and classify events within these multivariate time series.

Despite preprocessing, scaling, and windowing the data, methods such as DBSCAN, KMeansTimeSeries with dynamic time warping, HDBSCAN, and Gaussian Mixture Models were infeasible due to their high computational demands. The clustering method that performed efficiently was Mini-Batch K-Means. The effectiveness of Mini-Batch K-Means will be discussed in detail in the following chapter. With regards to expectations of clusters, we expect one cluster for the events, one for the artefarcts and one for no event.

Given that only one clustering method proved feasible, it became nec-

essary to explore alternatives within this clustering method. These alternatives will also be discussed in this chapter.

3.2 Original Approach

In the original approach, the features: end-tidal carbon dioxide, heart rate, diastole IBP, mean IBP, systole IBP, respiratory rate, and saturation were all included in the analysis. The main focus of this approach was to capture as much information as possible to make the data points distinguishable. The dataset was split into 10-minute windows, and for this approach, set to slide per minute of a patient's measurements. From here, the optimal number of clusters was calculated, which will be discussed later in this chapter. The algorithm used for clustering was Mini-Batch K-Means. After clustering, the clusters were validated both internally and externally. For visualization purposes, the windows were PCA-reduced to be shown in 3D. The validation process and clustering optimization provided statistics to compare the original approach to the alternative approaches.

3.2.1 Mini Batch K-means

The Mini-Batch K-Means algorithm is a variant of the standard K-Means algorithm. It employs the same clustering technique but with a key difference: instead of processing the entire dataset at once, it randomly selects small subsets (mini-batches) of the input data. These mini-batches are clustered to the nearest centroid, and the process continues iteratively until all data points are assigned to clusters. This approach significantly enhances computational efficiency and reduces memory usage compared to the standard K-Means algorithm, making it well-suited for large datasets.

However, the Mini-Batch K-Means algorithm has some limitations. It can be sensitive to the choice of mini-batch size: a batch size that is too large can negate the computational advantages, while a batch size that is too small can lead to noisy clustering results. Additionally, the algorithm may require more iterations to achieve convergence to a stable solution.

3.3 Optimizing the number of clusters

To determine the number of clusters for a model, three metrics were utilized: the Within-Cluster Sum Squares (WCSS), Silhouette Score and Davies-Bouldin Index. These metrics provided a comprehensive evaluation of clustering performance, supporting the identification of most suitable number of clusters.

3.3.1 Within-Cluster Sum of Squares (WCSS)

The WCSS measures the total variance within each cluster by calculating the sum of the squared distances between each point and its corresponding cluster centroid. It is used to assess the compactness of the clusters. The Elbow Method was applied to find the optimal number of clusters. This method involves plotting the WCSS against the number of clusters and identifying the "elbow" point, where the rate of decrease in WCSS significantly slows down, indicating diminishing returns from adding more clusters.

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where:

- K is the number of clusters.
- C_k is the k-th cluster.
- x_i is a data point in cluster C_k.
- μ_k is the centroid of cluster C_k.

3.3.2 Silhouette Score

The Silhouette Score evaluates the quality of the clustering by measuring how similar each point is to its own cluster compared to other clusters

(Rousseeuw, 1987). The score ranges from -1 to 1, with higher values indicating better-defined clusters. A higher Silhouette Score implies that the points are well-matched to their own cluster and poorly matched to neighboring clusters, signifying good separation between clusters.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ is the average distance between i and all other points in the same cluster.
- $b(i)$ is the minimum average distance between i and all points in any other cluster (the nearest cluster).

$$S = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

where:

- N is the total number of data points.

3.3.3 Davies-Bouldin Index

The Davies-Bouldin Index assesses the average similarity ratio of each cluster with its most similar cluster (Davies & Bouldin, 1979). A lower Davies-Bouldin Index indicates better clustering quality, with more distinct and well-separated clusters. This index is particularly useful for validating the compactness and separation of clusters.

$$\text{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \left(\frac{\sigma_k + \sigma_j}{d(\mu_k, \mu_j)} \right)$$

where:

- K is the number of clusters.
- σ_k is the average distance between each point in cluster k and the centroid of cluster k.
- σ_j is the average distance between each point in cluster j and the centroid of cluster j.
- $d(\mu_k, \mu_j)$ is the distance between the centroids of cluster k and cluster j.

3.4 Validation

3.4.1 Internal validation

For internal validation, a Random Forest classifier will be used to determine if the data points are easily distinguishable, thus assessing the effectiveness of the labeling process. This involves comparing accuracy measurements to ensure that data points are correctly clustered. To fairly classify the data, the number of labels needed to be balanced, so the dataset was undersampled to achieve an even distribution of labels. The dataset was then split into a training set (70%) and a test set (30%). The classifier was trained on the training set and tested on the test set, with accuracy measured. Additionally, a confusion matrix and a classification report were generated to validate the classifier's performance.

3.4.2 External validation

To understand the characteristics of the clusters, the mean and the standard deviations were calculated of each cluster for each measurement. Additionally, to assess the realism of the clustering, the distribution of clusters per patient was analyzed. The labeled windows were also visualized within individual patient measurements to inspect their accuracy and consistency.

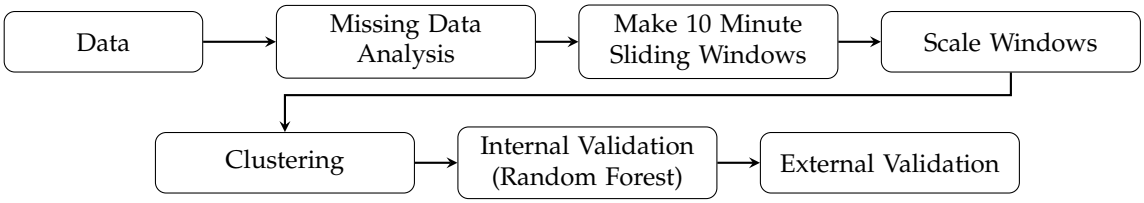

Figure 3.1: Methodology Flow Chart

3.5 Alternative approach

3.5.1 Feature Selection and Transformation

Given the large and dense set of data points in the original approach, it was necessary to reduce computational costs and prevent model overfitting. To address the density of data points and find more variability, different options were considered to achieve more concise clusters. Most features were independent, except for the systolic, diastolic, and mean invasive blood pressure (IBP) measurements. Due to their interdependence and the redundancy of information they provided, only the mean IBP was included in the analysis.

To improve the separability of data points, the focus shifted to examining the difference of measurements relative to a baseline of each patient (median). This approach accounts for individual patient "normal" values. The window sliding time was also adjusted from 1 minute to 5 minutes to reduce overlap and enhance the distinction between data points.

Two transformation approaches were used:

- Percentage error:

$$\frac{\text{measurement} - \text{median}}{\text{median}} \times 100$$

- Cubed difference:

$$(\text{measurement} - \text{median})^3$$

Cubing the difference emphasized outliers and keeps negative values negative, making these data points more distinct and easier to cluster. The internal and external validation processes remained consistent with those mentioned earlier in this chapter.

4. Results

This chapter presents the findings of the study, comparing the results of different clustering approaches. It provides a detailed analysis of the original approach and the alternative methods, highlighting their respective performances. The chapter also includes a comprehensive comparison of the approaches and discusses the external validation of the model. During this chapter only a few results will be shown, in the appendix from A.2 until A.5 the rest has been attached.

4.1 Original approach

In figure 4.1 you can see the distributions of the points and the optimal amount of clusters for this model. The sliding windows in this model are very close to each other, resulting in significant overlap and many identical points. This overlap explains the dense clustering, with little distinction between data points. Additionally, using three blood pressure measurements (systolic, diastolic, and mean) with similar variance and interdependence adds redundancy. Including all three does not contribute to distinguishing between the data points, as they provide similar variability.

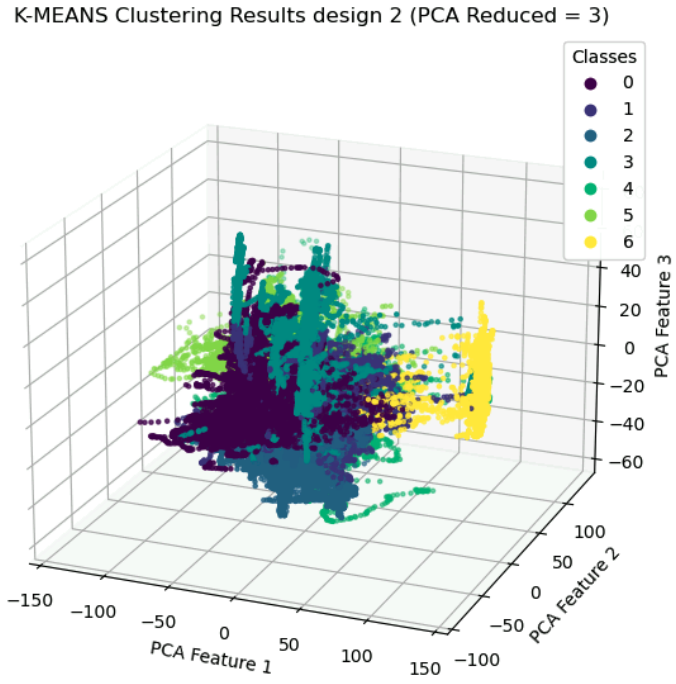

Figure 4.1: Plot of the original approach

4.2 Alternative approach

4.2.1 Second model: difference % median window 10m - 1m

Figure 4.2 presents the clustering results for the 10-minute window with 1-minute steps using the percentage difference baseline approach, reduced to three principal components (PCA). The plot illustrates five distinct clusters, each represented by a different color. The clustering shows a clear separation between the different classes, with Class 1 (blue) and Class 4 (yellow) exhibiting more compact and distinct groupings, suggesting homogeneity within these clusters. Classes 0 (purple), 2 (teal), and 3 (green) display more spread distributions, indicating greater variability within these clusters.

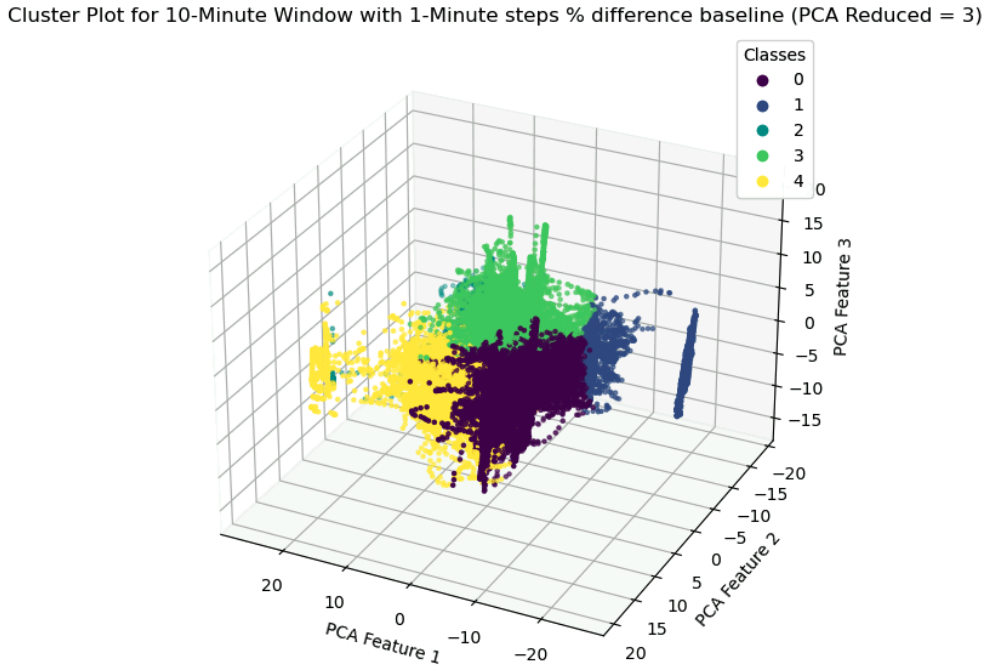

Figure 4.2: Plot of the Alternative (% difference) 1 minute sliding approach

4.2.2 Third model: difference % median window 10m - 5m

Figure 4.3 presents the clustering results for the 10-minute window with 5-minute steps using the percentage difference baseline approach, reduced to three principal components (PCA). The plot illustrates five distinct clusters, each represented by a different color. The clustering shows clear separation between the different classes, with Class 0 (purple) and Class 4 (yellow). The Classes: 1, 2 and 3 have less variability in their data points.

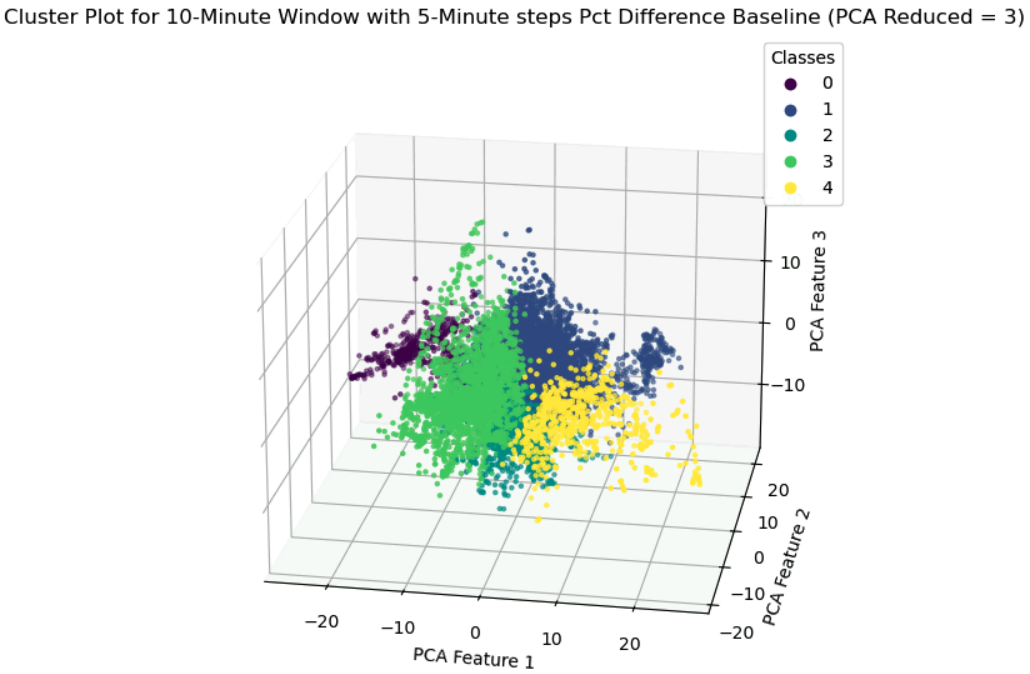

Figure 4.3: Plot of the Alternative (% difference) 5 minutes sliding approach

4.2.3 Fourth model: difference^3 median window 10m - 5m

Figure 4.4 presents the clustering results using the cubed difference baseline approach with 10-minute windows and 5-minute steps, reduced to three principal components (PCA). The plot illustrates three distinct clusters, each represented by a different color. The clustering shows an even dispersion of the different classes, with Classes 0 (purple), 1 (teal), and 2 (yellow) spread uniformly across the PCA space. This even dispersion suggests that the cubed difference baseline approach captures a balanced representation of the data, indicating consistent variability within each cluster. The PCA features on the axes captures the significant variance in the data, showing the that the clustering algorithm can differentiate between various patterns.

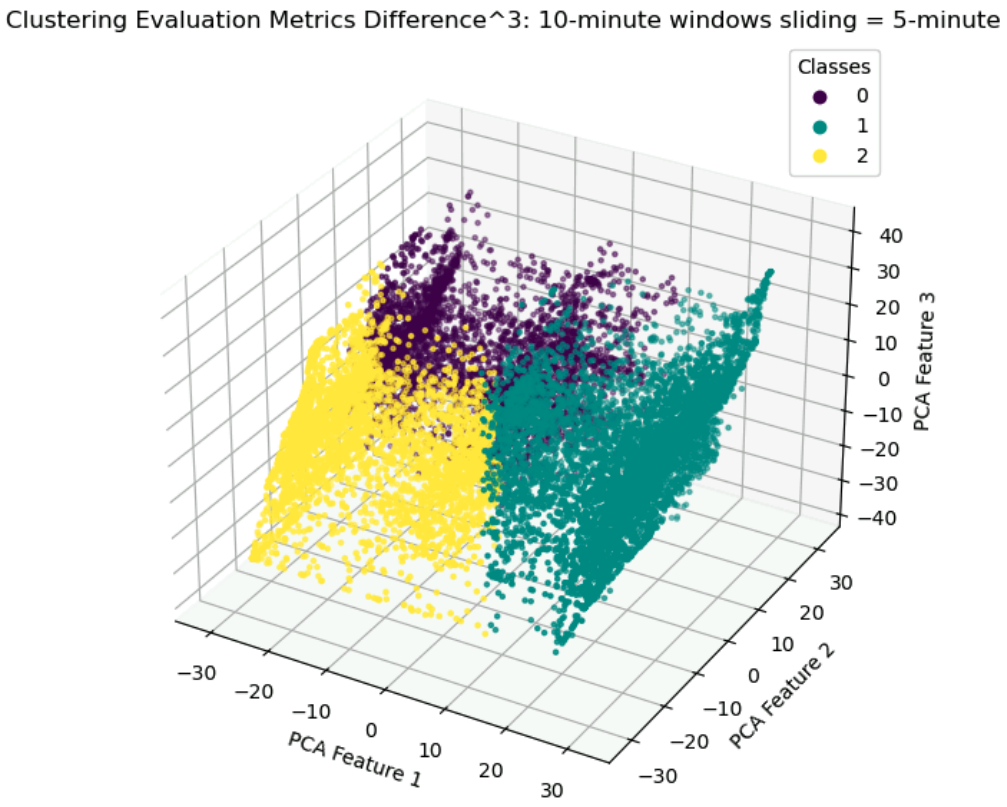

Figure 4.4: Plot of the Alternative (cubed) approach

4.3 Comparison of the approaches

Table 4.1 presents the internal validation results for the different clustering approaches using a Random Forest classifier to measure accuracy. The "Difference Baseline Cubed (10m - 5m)" model achieved the highest accuracy of 0.966, indicating the most reliable cluster differentiation. This model also demonstrated the best clustering performance with a Davies-Bouldin Index of 1.560 and a Silhouette score of 0.21, suggesting well-separated and cohesive clusters. In comparison, the original model, despite having more clusters (7), showed lower performance with an accuracy of 0.958, a Davies-Bouldin Index of 1.63, and a Silhouette score of 0.16. The "Difference Baseline %" models, using 10-minute windows with 1-minute and 5-minute steps, achieved accuracies of 0.959 and 0.941, respectively, with moderate Davies-Bouldin and Silhouette scores. These results highlight the effective-

ness of the cubed difference baseline approach combined with larger step sizes in capturing meaningful patterns in the data and achieving a better clustering quality.

Table 4.1: Internal Validation Results

Model	Window	n Clusters	Davies-Bouldin	Silhouette	WCSS	Accuracy
						Random Forest
Original	10m - 1m	7	1.63	0.16	110597494	0.958
Difference baseline %	10m - 1m	5	1.79	0.13	4761626	0.959
Difference baseline %	10m - 5m	5	1.71	0.13	1061595	0.941
Difference baseline^3	10m - 5m	3	1.56	0.21	10392273	0.966

4.4 External validation of the model

In this section we will discuss results for the cubed model.

4.4.1 Statistics of each cluster

Table 4.2: Centroid Difference to the Median (10-minute Window): Mean Difference and Standard Deviation per Measurement

Cluster	ETCO2		Heart Rate		Respitory Rate		Blood Pressure		Saturation	
	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
0	0.00	0.60	4.61	1.44	0.82	2.95	-0.19	3.60	-0.35	2.87
1	0.04	0.55	-3.30	3.00	-0.61	2.05	-3.10	2.84	-1.05	2.84
2	-0.07	0.58	-2.54	2.52	-0.02	3.08	2.83	2.17	-0.67	2.81

Table 4.3: Cluster Difference to the Median per Centroid (10-minute Window): Mean and Standard Deviation per Measurement (%)

Cluster	ETCO2 (%)		Heart Rate (%)		Respiratory Rate (%)		Blood Pressure (%)		Saturation (%)	
	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
0	-0.01	0.72	1.79	2.08	-0.10	1.45	-0.16	1.11	2.13	1.45
1	0.13	1.07	-1.46	1.13	-1.70	1.44	-0.47	1.13	2.04	1.44
2	-0.20	1.11	-1.06	1.28	-1.60	1.41	0.37	1.41	2.06	1.41

Table 4.4: Mean of the Median for each Cluster

Cluster	ETCO2	Heart Rate	Respiratory Rate	Blood Pressure	Saturation
0	4.36	146.82	31.14	52.87	87.98
1	4.28	145.48	30.98	53.69	88.59
2	4.41	146.02	30.98	52.40	87.74

4.4.2 Statistics clusters per patient

In this section, we discuss and visualize the clustering of events per patient to evaluate the realism of the clusters. Table 4.5 presents the mean percentage of windows that were clustered into each cluster for each patient. The average time clustered in Cluster 0 is 41.12%, in Cluster 1 is 25.80%, and in Cluster 2 is 33.08%. Given that Clusters 1 and 2 exhibit the most variability, they are likely to correspond to distinct events or artifacts, which aligns with our expectations.

Table 4.5: Mean % Windows in each Cluster

Cluster	Mean % windows for all patients
0	41.12
1	25.80
2	33.08

5. Discussion and Conclusion

The concluding chapter summarizes the key findings and contributions of the study. It offers a discussion on the implications of the results, drawing conclusions on the effectiveness of the clustering methods used. The chapter also addresses the limitations of the study and proposes recommendations for future research. By reflecting on the study's impact on patient care in the PICU, this chapter emphasizes the importance of continued advancements in data science for healthcare applications.

5.1 Discussion

Within this study, the aim was to detect and classify medical interventions (events) within the patient dataset provided. There were four different approaches to reach this goal, and these approaches differ in parameters but were all clustered using the same algorithm: Mini Batch K-means. All of the models were optimized to find the best number of clusters and validated internally and externally. For the internal validation, the Random Forest classifier was used to measure the accuracy of the predicted clusters. The results in Table 4.1 show that the model with the difference to the baseline cubed is the best-separated one. The results are close to each other, but when looking at the Davies-Bouldin Index and the Silhouette score, the cubed model scores the best well.

The hypothesis for the general clustering is that one cluster will represent no event, one will capture artifacts, and one will represent events. Given the assumptions, the most important thing to keep in mind is that we expected to cluster the same kind of change within one cluster and assume that these changes were the same kind of event.

5.1.1 Internal Validation

The internal validation results indicated that the "Difference baseline cubed" model performed the best across several metrics. Specifically, this model had the highest accuracy (0.966) when validated using the Random Forest classifier. This high accuracy suggests that the clusters formed are distinct and the model can reliably differentiate between different types of events. Additionally, the Davies-Bouldin Index (1.560) and Silhouette score (0.21) for this model were the most favorable, further indicating well-separated clusters with a reasonable level of cohesion.

5.1.2 External Validation

In the external validation of the model, the statistics of each cluster, as shown in Tables 4.2 and 4.3, provide a deeper understanding of the characteristics of each cluster. The analysis of the cluster differences to the median per centroid over a 10-minute window reveals distinct patterns for each cluster.

For instance, in Table 4.2, Cluster 0 shows relatively stable values across measurements, suggesting that this cluster might represent periods with no significant events. Cluster 2, on the other hand, exhibits higher standard deviations in respiratory rate and blood pressure, which might indicate the presence of events or interventions. Cluster 1 shows noticeable differences in heart rate and saturation levels, which could suggest artifacts or specific types of events.

In Table 4.3, where the difference to the median per centroid is expressed as a percentage, similar patterns emerge. Cluster 0 has smaller deviations, reinforcing the idea that this cluster might represent normal, event-free periods. Cluster 1 and Cluster 2 show larger deviations, indicating more variability likely due to medical events or artifacts.

5.1.3 Comparison of the Approaches

The alternative approaches to clustering provided notable benefits, particularly in terms of efficiency. By using fewer data points with more variation,

the clustering algorithm was faster and got more accurate results. The minimum expectation was to identify clusters corresponding to no events, artifacts, and actual events. This was successfully achieved using the cubing model, and the percentage difference model also effectively clustered the data into meaningful groups.

However, it is crucial for experts to externally validate these clusters to determine what the events represent. While some models appeared visually better separated, their accuracy scores were comparable. This suggests that, although it is possible to distinguish data points based on measurements alone, the true meaning of these clusters requires validation. Expert evaluation is necessary to confirm the types of events each cluster represents.

5.1.4 Objective

Despite the promising results, the objective of the study has not been fully achieved. One significant limitation is the inability to label the clusters accurately. As a non-expert, it is challenging to determine if a particular measurement belongs to a specific medical event. This limitation proves the necessity of collaboration with medical professionals who can provide expert annotations for the data, ensuring that the clusters are correctly interpreted and validated.

5.2 Conclusion

In conclusion, the "Difference baseline cubed" model proved to be the most effective in clustering the dataset into meaningful groups. The internal validation metrics demonstrated high accuracy and well-separated clusters, while the external validation provided insights into the characteristics of each cluster. These results support the hypothesis and confirm that the clustering approach can successfully differentiate between normal periods, artifacts, and medical events.

However, it is important to note that the objective of the study has not been fully achieved due to the inability to accurately label the clusters. Col-

laboration with medical professionals is essential to provide expert annotations for the data, ensuring that the clusters are correctly interpreted and validated. This clustering framework could potentially enhance patient monitoring and intervention strategies in clinical settings.

5.3 Limitations

5.3.1 Unsupervised Nature

Being unsupervised, there is no way to exclusively cluster events. Some clusters may represent a combination of patient deterioration and medication effects, without clear confirmation of specific events.

5.3.2 Event Intervals and Duration

There is no clear indication of the number and duration of interventions. The chosen windows might not capture precise events. Due to time constraints, finding the "optimal" window size was not feasible. Different window lengths could yield better results, as events like blood collection (5 minutes) versus oxygen adjustment can vary significantly. The study's windows might capture multiple events, complicating clustering. Shorter windows, might better recognize specific events.

5.3.3 Variability of Patient Response

The initial assumption was that similar events would produce uniform effects across all patients, resulting in consistent changes in measurements. However, the variability in individual patient responses to medication means that identical events can express differently in different patients. This variability complicates the task of clustering these events based solely on measurement data. Similarly, different events that produce similar measurement patterns in a single patient cannot be easily distinguished based solely on these measurements.

5.3.4 Computational Constraints

The use of sensitive data necessitated a secure virtual machine, limiting the use of computationally intensive methods. This time constraint meant only faster models were viable, potentially missing more complex patterns. Better models might be achievable with higher processing power or a smaller dataset.

5.4 Future Work and Recommendations

Given the high accuracy achieved using the relatively simple K-Means clustering algorithm, future work could focus on validating and possibly refining this approach further. However, exploring more computationally intensive models could still offer benefits in terms of robustness and generalization. Advanced techniques such as deep learning methods might provide improvements in handling edge cases and outliers, thereby enhancing the overall reliability of the clustering process.

Additionally, experimenting with various window sizes and step sizes could provide deeper insights and help identify specific events within individual clusters. Tailoring the windowing approach to different types of medical events might further enhance detection accuracy.

Another important aspect of future research is the validation of the clustered data points by medical professionals. Expert annotations would ensure that the clusters are accurately interpreted and that the events identified are clinically relevant. Collaboration with healthcare experts would not only validate the current findings but also refine the clustering approach to better suit medical needs.

Moreover, focusing on a specific event approach could enhance the detection and classification of medical events. This involves targeting certain types of events by assigning different weights to the measurements. Specifically, research could look for three kinds of events: pulmonary, cardiac, and blood pressure-related events, by giving more weight to the measurements most relevant to these events. This targeted approach could improve the ac-

curacy and relevance of the event detection algorithms, providing more precise monitoring and intervention strategies tailored to these distinct physiological responses.

Lastly, implementing real-time clustering and anomaly detection systems in clinical settings could offer significant advancements in patient monitoring and intervention strategies. Developing user-friendly tools and dashboards for healthcare providers to visualize and interact with the clustering results would facilitate practical application and improve patient outcomes.

Bibliography

Malani, P. N. (2012). Harrison's Principles of Internal Medicine. *JAMA*, *308*(17), 1813. https://doi.org/10.1001/jama.308.17.1813-b

Van Lieshout, E. J., Van Lieshout, J. J., Harkel, A. D. J. T., & Wieling, W. (1989). Cardiovascular Response to coughing: Its value in the assessment of autonomic nervous control. *Clinical science*, *77*(3), 305–310. https://doi.org/10.1042/cs0770305

Srinivasulu, A., Mohan, S. M., Harika, T., Srujana, P., & Revathi, Y. (2021). Apnea Event Detection Using Machine Learning Technique for the Clinical Diagnosis of Sleep Apnea Syndrome. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. https://doi.org/10.1109/icspc51351.2021.9451743

Potter, İ. Y., Zerveas, G., Eickhoff, C., & Duncan, D. (2022). Unsupervised Multivariate Time-Series Transformers for Seizure Identification on EEG. *IEEE Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla55696.2022.00208

Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of healthcare engineering*, *2022*, 1–9. https://doi.org/10.1155/2022/7351061

Van Hespen, K. M., Zwanenburg, J. J., Dankbaar, J. W., Geerlings, M. I., Hendrikse, J., & Kuijf, H. J. (2021). An anomaly detection approach to identify chronic brain infarcts on MRI. *Scientific reports*, *11*(1). https://doi.org/10.1038/s41598-021-87013-4

Chen, L., Dubrawski, A., Wang, D., Fiterau, M., Guillame-Bert, M., Bose, E., Kaynar, A. M., Wallace, D. J., Guttendorf, J., Clermont, G., Pinsky, M. R., & Hravnak, M. (2016). Using supervised machine learning to classify real alerts and artifact in online multisignal vital sign monitoring data*. *Critical care medicine*, *44*(7), e456–e463. https://doi.org/10.1097/ccm.0000000000001660

Kavitha, M., Srinivas, P. V. V. S., Kalyampudi, P. S. L., F, C. S., & Srinivasulu, S. (2021). Machine learning techniques for anomaly detection in smart healthcare. *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. https://doi.org/10.1109/icirca51532.2021.9544795

Kanagala, H. K., & Krishnaiah, V. J. R. (2016). A comparative study of K-Means, DBSCAN and OPTICS. *International Conference on Computational Collective Intelligence*. https://doi.org/10.1109/iccci.2016.7479923

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current genomics*, *22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/tpami.1979.4766909

A. Appendices

A.1 Vital Measurements

- **ETCO2 (mon_etco2)**: End-tidal carbon dioxide (mmHg) - The maximum concentration of carbon dioxide at the end of an exhaled breath, which indicates how well CO2 is being eliminated by the lungs.
- **Heart Rate (mon_hr)**: Heart rate (beats per minute) - The number of heartbeats per minute.
- **Diastole IBP (mon_ibp_dia)**: Invasive Diastolic Blood Pressure (mmHg) - The pressure in the arteries when the heart is at rest between beats, measured invasively.
- **Mean IBP (mon_ibp_mean)**: Invasive Mean Blood Pressure (mmHg) - The average arterial pressure during a single cardiac cycle, measured invasively.
- **Systole IBP (mon_ibp_sys)**: Invasive Systolic Blood Pressure (mmHg) - The pressure in the arteries when the heart beats, measured invasively.
- **Diastole NIBP (mon_nibp_dia)**: Non-Invasive Diastolic Blood Pressure (mmHg) - The pressure in the arteries when the heart is at rest between beats, measured non-invasively.
- **Mean NIBP (mon_nibp_mean)**: Non-Invasive Mean Blood Pressure (mmHg) - The average arterial pressure during a single cardiac cycle, measured non-invasively.
- **Systole NIBP (mon_nibp_syst)**: Non-Invasive Systolic blood pressure (mmHg) - The pressure in the arteries when the heart beats, measured non-invasively.
- **Respiratory Rate (mon_rr)**: Number of breaths per minute - The number of breaths taken per minute.
- **Saturation (mon_sat)**: Blood oxygen saturation (%) - The percentage of oxygen-saturated hemoglobin relative to total hemoglobin in the blood.

- **Near-infrared spectroscopy Left (mon_nirs_l)**: Near-Infrared Spectroscopy measurement on the left side (nm)- Used to assess tissue oxygenation and hemodynamics on the left side of the body.
- **Near-infrared spectroscopy Right (mon_nirs_r)**: Near-Infrared Spectroscopy measurement on the right side (nm)- Used to assess tissue oxygenation and hemodynamics on the right side of the body.

A.2 Missing Values table

NA table

Key	Value
('mon_etco2', 'mon_hr', 'mon_ibp_dia', 'mon_ibp_mean', 'mon_ibp_sys', 'mon_rr', 'mon_sat')	17
('nirs_mean',)	12
('mon_etco2', 'mon_ibp_dia', 'mon_ibp_mean', 'mon_ibp_sys')	4
('mon_rr',)	3
('mon_rr', 'nirs_mean')	2
('mon_etco2', 'mon_ibp_dia', 'mon_ibp_mean', 'mon_ibp_sys', 'nirs_mean')	1
('mon_etco2', 'nirs_mean')	1
('mon_ibp_dia', 'mon_ibp_mean', 'mon_ibp_sys', 'nirs_mean')	1

Table A.1: Count of Patients with Different Combinations of Missing Measures

A.3 Original Design Results

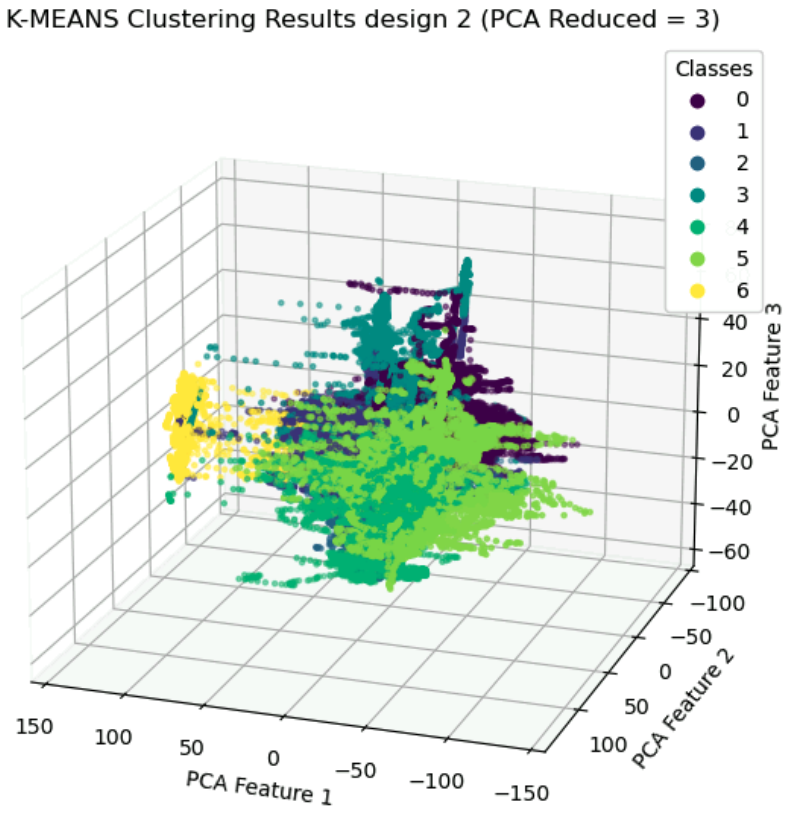

Figure A.1: Plot of the Original approach

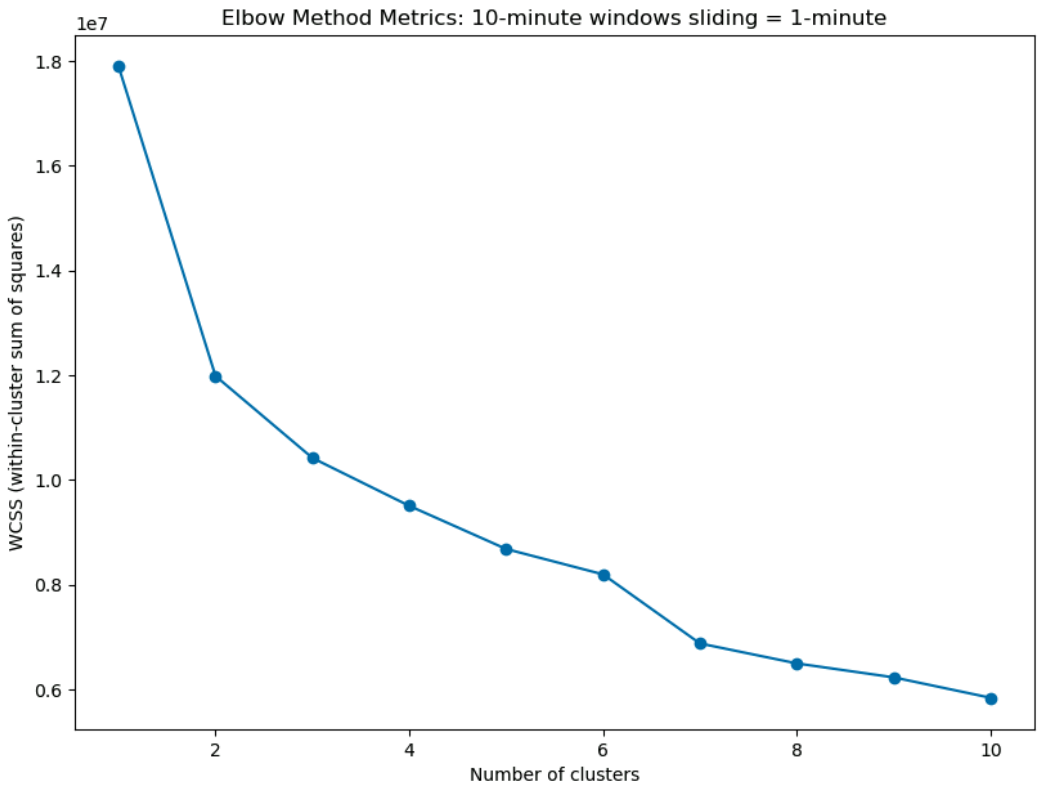

Figure A.2: Elbow Plot of the Original approach

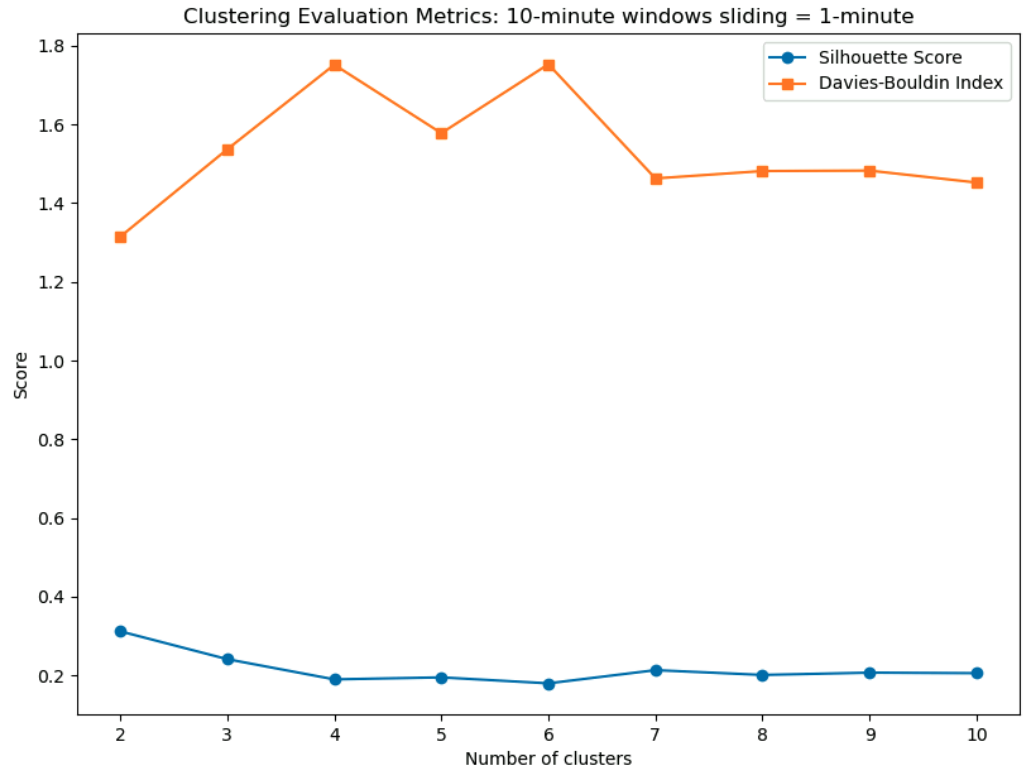

Figure A.3: Silhouette Score of the Original approach

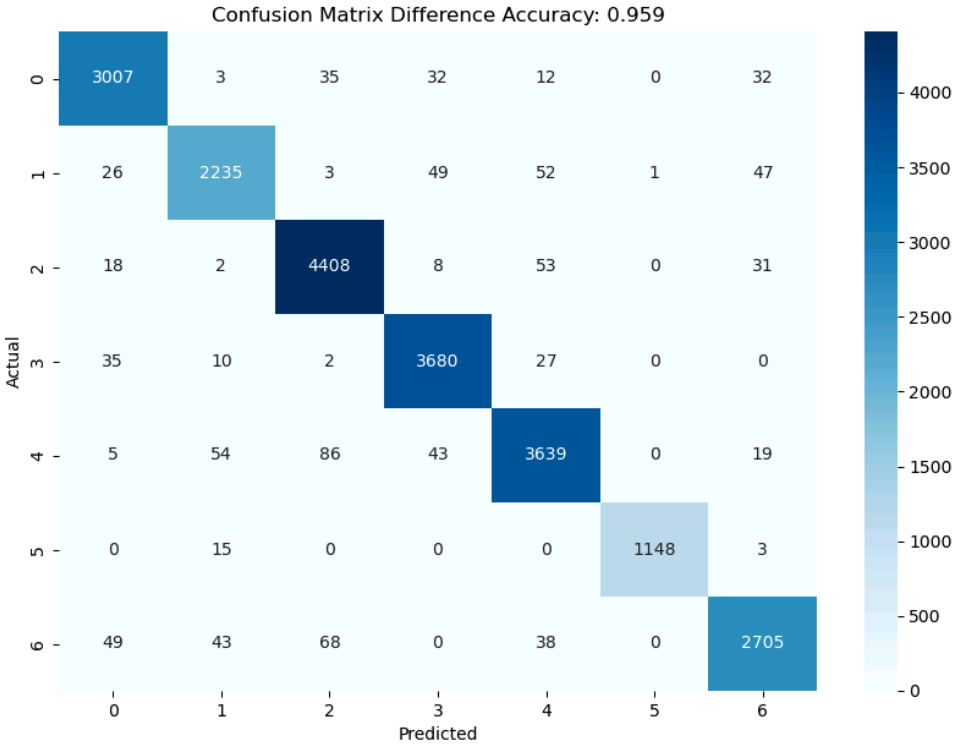

Figure A.4: Confusion Matrix Original approach

Table A.2: Classification Report: Original approach

Class	Precision	Recall	F1-Score	Support
0	0.96	0.96	0.96	3121
1	0.95	0.93	0.94	2413
2	0.96	0.98	0.97	4520
3	0.97	0.97	0.97	3754
4	0.95	0.95	0.95	3846
5	1.00	0.99	0.99	1166
6	0.95	0.93	0.94	2903
Accuracy			0.96	21723
Macro Avg	0.96	0.96	0.96	21723
Weighted Avg	0.96	0.96	0.96	21723

Table A.3: Cluster Centroids Means and Standard Deviations for Vital Measurements

Cluster	ETCO2		Heart Rate		IBP Dia		Respiratory Rate		IBP Mean		IBP Sys		Saturation	
	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
0	4.46	0.79	170.24	15.63	44.66	4.17	36.72	5.75	54.37	5.04	72.66	8.29	89.05	9.68
1	3.78	0.91	143.55	12.36	41.68	4.92	24.90	6.65	55.75	5.06	81.44	9.78	80.33	10.74
2	5.21	0.81	143.79	12.95	36.46	4.47	36.36	6.26	46.35	4.79	62.16	8.14	95.00	5.39
3	4.43	0.82	139.98	15.65	53.22	7.18	37.17	6.70	66.67	6.92	92.43	10.46	91.15	8.13
4	3.88	0.92	128.30	13.45	39.00	4.77	32.45	5.12	51.12	5.24	72.01	9.94	83.70	10.06
5	3.38	0.76	147.48	8.81	40.41	5.35	0.62	3.51	49.71	5.92	63.09	9.76	94.31	6.20
6	4.16	0.71	159.19	13.53	35.74	4.85	31.02	4.47	46.31	4.83	64.78	7.58	77.87	8.80

Table A.4: Cluster Distribution and Window %

Pseudo ID	0	1	2	3	4	5	6	# Windows	Approx. Time (h)
1	0	0	12.53	0	87.47	0	0	359	59.83
2	0	12.86	0	2.06	2.83	82.24	0	1695	282.5
3	0	0.29	0.39	0	99.12	0	0	1021	179.17
5	16.08	4.77	19.65	4.41	26.30	0.26	1.54	7396	1232.67
7	13.46	0	0	36.81	18.96	1.65	0	364	60.67
8	0	0	0	0.87	0	0	0	18	3
9	8.97	4.43	25.50	0.15	11.78	44.43	4.74	4090	681.67
10	9.94	5.06	0	1.86	6.11	0	2.54	3272	545.33
12	0	77.07	0	22.95	0	0	0	458	76.33
15	76.66	0.99	3.47	2.58	0.50	4.50	0	2018	336.33
16	36.52	0	0	28.19	0	0	0	1703	283.83
18	0.24	0	41.23	0	0	1.44	7.10	6212	1035.33
19	0	95.75	0	0	31.44	0	3.94	636	106
22	58.21	6.83	0.57	4.86	0.21	3.76	25.57	3355	559.17
24	2.13	0.07	54.89	29.79	13.12	0	0	3048	508
27	12.69	4.57	37.26	9.14	56.35	0	0	197	32.83
28	37.16	62.64	0	0	0	0.19	0	2061	343.5
29	66.11	0	0	0	0	0	0	33	30
30	0.61	0	13.66	61.43	15.30	0	0	28	4.67
34	6.35	87.30	0	0	0	6.35	0	63	10.5
35	0	81.52	0	22.64	0	0	0	18	15.33
38	0	10.99	20.32	18.24	46.75	0	0	92	15.33
40	0.78	60.94	14.06	5.47	0	18.75	0	128	21.33
42	4.94	0	0	93.13	0	0	0	1862	310.33
44	1.25	1.21	82.84	0	4.93	0	9.77	2477	412.83
46	6.04	5.28	71.89	15.28	0.19	1.32	0	1060	176.67
47	0	1.04	0	0	0	0	0	0	2.66
50	0.94	0	12.15	24.65	0	0	0	58	1.53
52	4.86	0	0.09	39.66	0	0.04	0	5380	896.67
53	2.82	0	15.39	34.66	0	7.44	7.49	1988	331.33
55	11.75	16.23	15.32	0.32	0.19	1.67	1.67	2537	422.83
61	2.12	90.92	0	0	6.56	0.39	0	1509	251.5
63	8.40	23.36	0.32	1.10	0.07	66.57	0	8181	1363.5
65	7.35	0.57	7.19	2.69	0	0	42.19	3155	525.83
66	5.81	0.08	91.99	1.42	0	0	0	2537	422.83
67	11.27	8.17	20.28	0	57.75	0	2.54	355	59.17

Table A.5: Average Window Distribution for each Patient

Pseudo ID	**Mean % Windows**
0	32.3846
1	14.1487
2	14.246
3	22.7658
4	15.6818
5	18.5032
6	5.123
# Windows	1856.59
Approx. Time (h)	309.432

A.4 Cubed Design

Table A.6: Cluster Distribution and Window %

Pseudo ID	0	1	2
1	41.33	30.67	28.00
2	45.58	20.23	34.19
3	37.02	7.69	55.29
5	38.47	39.61	22.46
7	43.99	19.51	36.59
9	100.00	0.00	0.00
10	37.76	36.20	26.85
12	41.55	14.14	39.04
16	18.09	39.36	42.55
19	49.40	24.82	25.79
22	41.83	20.63	37.54
24	40.66	35.65	23.69
27	46.51	14.73	38.76
29	44.82	21.39	34.02
30	41.79	31.05	28.16
35	39.02	26.83	34.15
38	36.65	40.86	22.49
42	26.83	12.20	60.98
43	60.00	0.00	40.00
46	42.09	40.45	17.45
47	21.88	46.88	31.25
52	47.62	28.57	23.81
55	26.19	42.86	30.95
61	26.09	19.06	54.84
63	32.14	10.71	57.14
66	36.24	39.68	24.07
67	48.48	36.36	15.15
44	43.81	22.20	33.99
46	40.15	27.27	32.58
47	15.79	31.58	52.63
52	47.62	33.33	19.05
53	46.49	26.14	27.37
55	36.72	26.05	37.22
56	41.38	24.14	34.48
61	40.51	36.33	23.15
63	44.43	22.04	33.53
65	39.50	28.21	32.29
66	41.95	20.14	37.93
67	35.62	23.29	41.10

A.5 % Difference Median 10m-1m Results

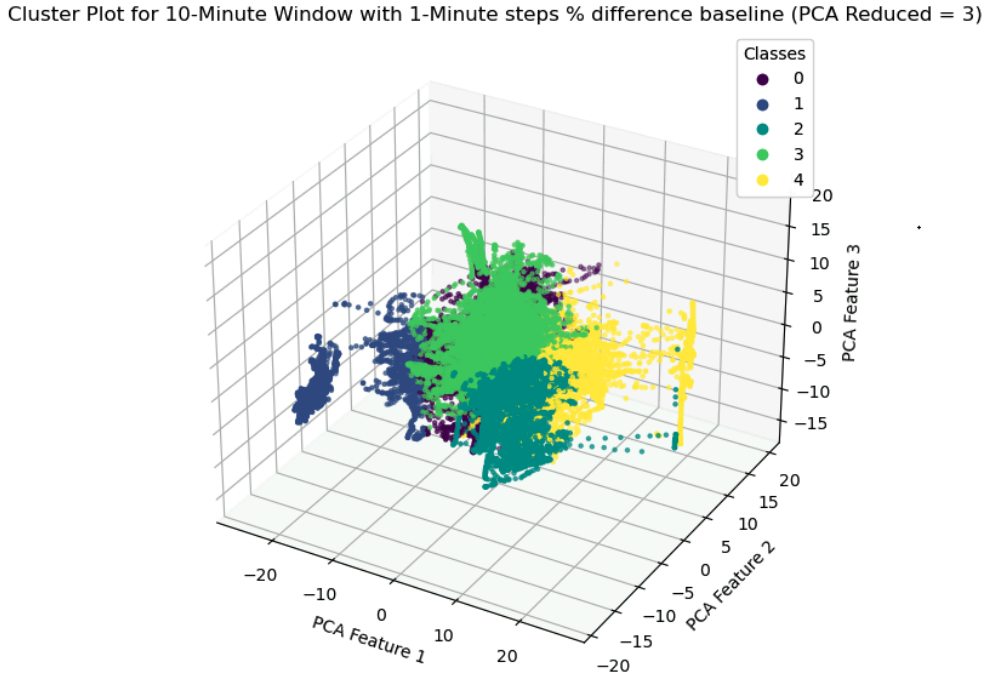

Figure A.5: Plot of the Alternative 10m-1m Baseline Difference approach

Table A.7: Classification Report: Alternative 10m-1m Baseline Difference approach

Class	Precision	Recall	F1-Score	Support
0	0.96	0.98	0.97	8274
1	0.99	0.93	0.96	952
2	0.97	0.96	0.97	2531
3	0.95	0.97	0.96	7186
4	0.98	0.89	0.93	2568
Accuracy			0.96	21723
Macro Avg	0.96	0.96	0.96	21723
Weighted Avg	0.96	0.96	0.96	21723

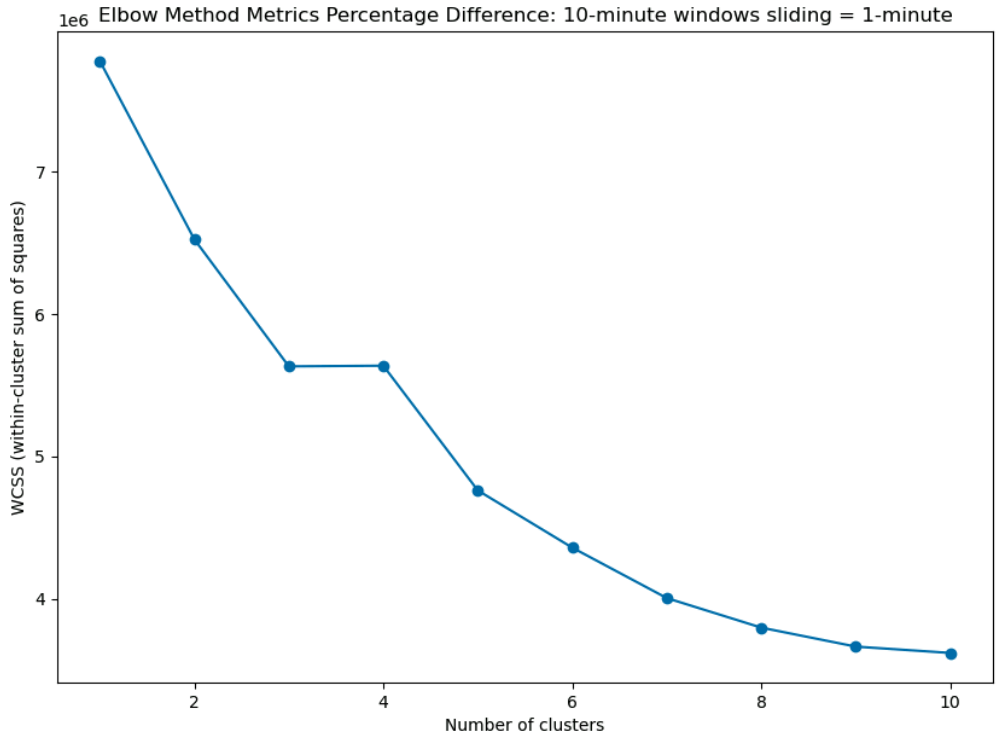

Figure A.6: Elbow Plot Alternative 10m-1m Baseline Difference approach

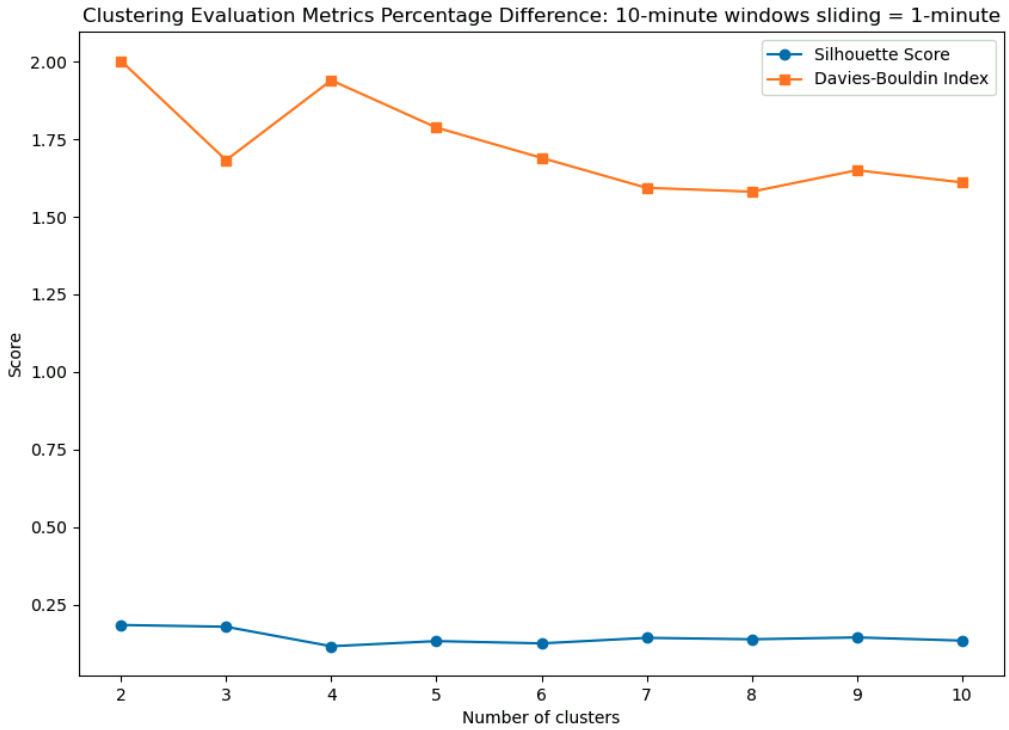

Figure A.7: Silhouette & Davies Bouldin Index Alternative 10m-1m Baseline Difference approach

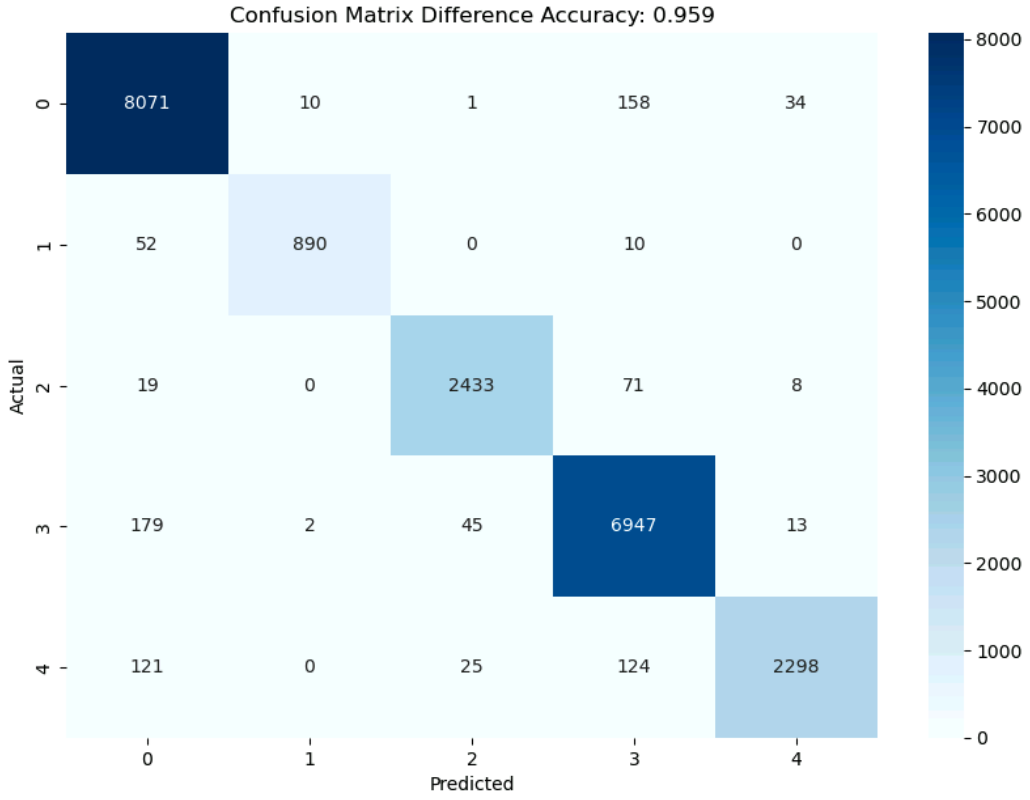

Figure A.8: Confusion Matrix Alternative 10m-1m Baseline Difference approach

A.6 % Difference Median 10m-5m Results

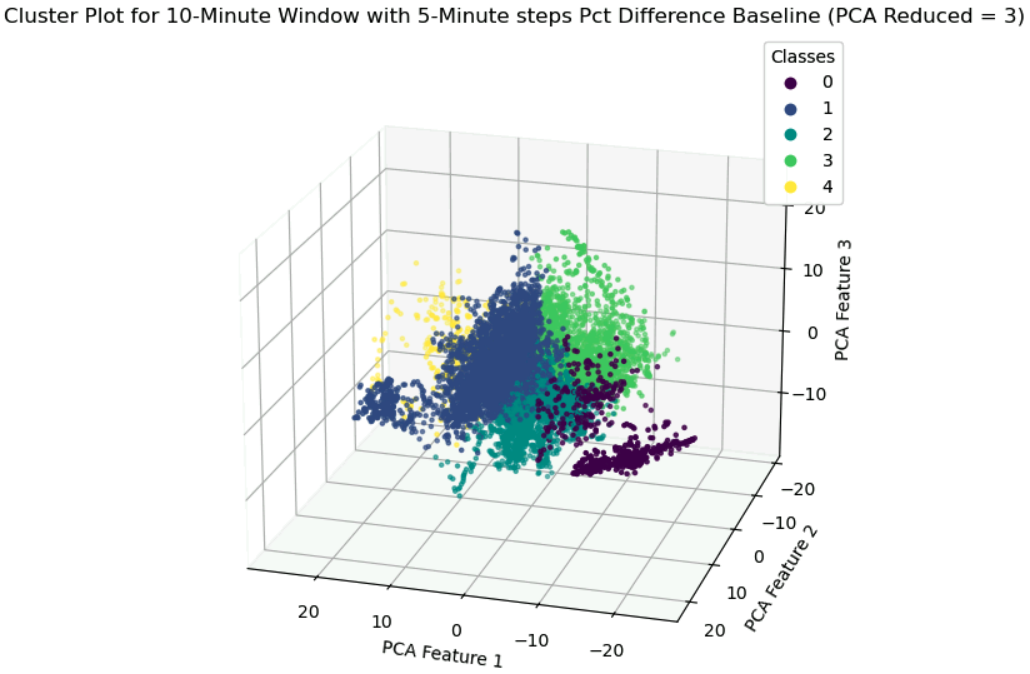

Figure A.9: Plot of the Alternative 10m-5m Baseline Difference approach

Table A.8: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.95	0.96	0.95	1027
1	0.96	0.96	0.96	1412
2	0.98	0.88	0.93	213
3	0.93	0.88	0.90	444
4	0.92	0.93	0.92	1206
Accuracy			0.96	21723
Macro Avg	0.96	0.96	0.96	21723
Weighted Avg	0.96	0.96	0.96	21723

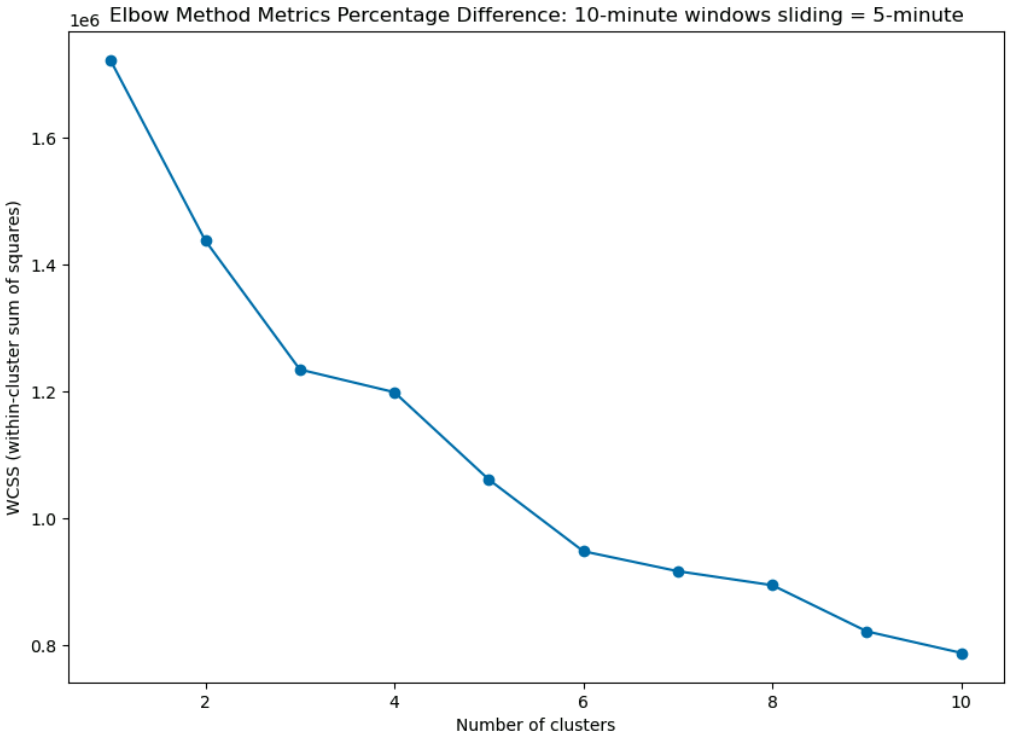

Figure A.10: Plot of the Alternative 10m-5m Baseline Difference approach

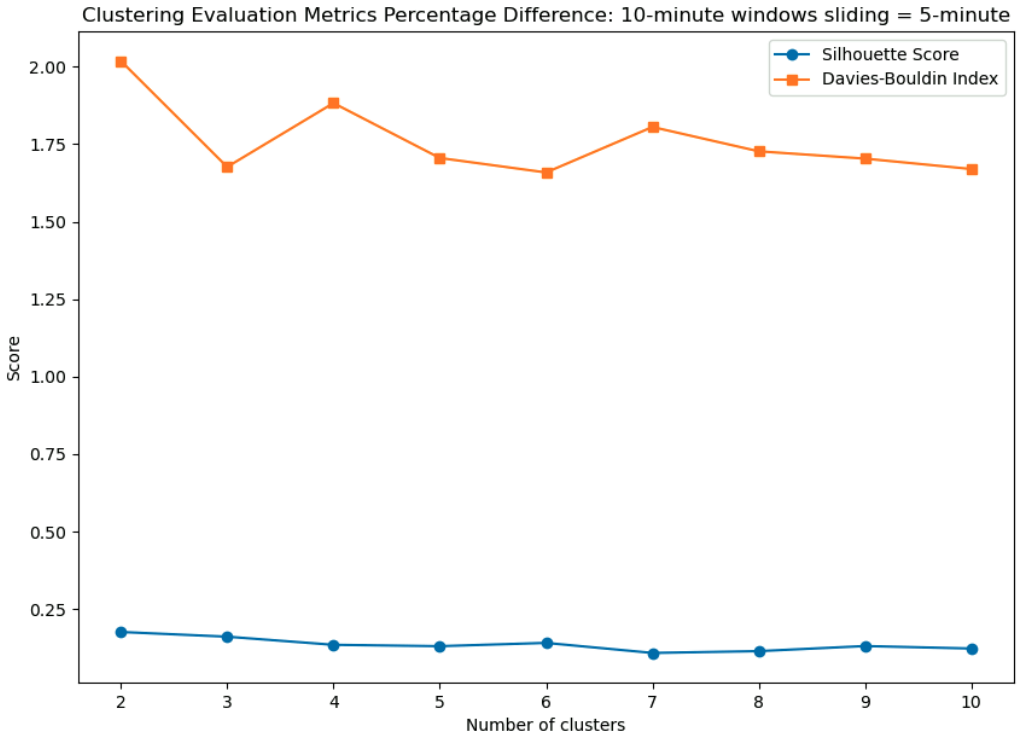

Figure A.11: Silhouette and Davies Bouldin Index of the Alternative 10m-5m Baseline Difference approach

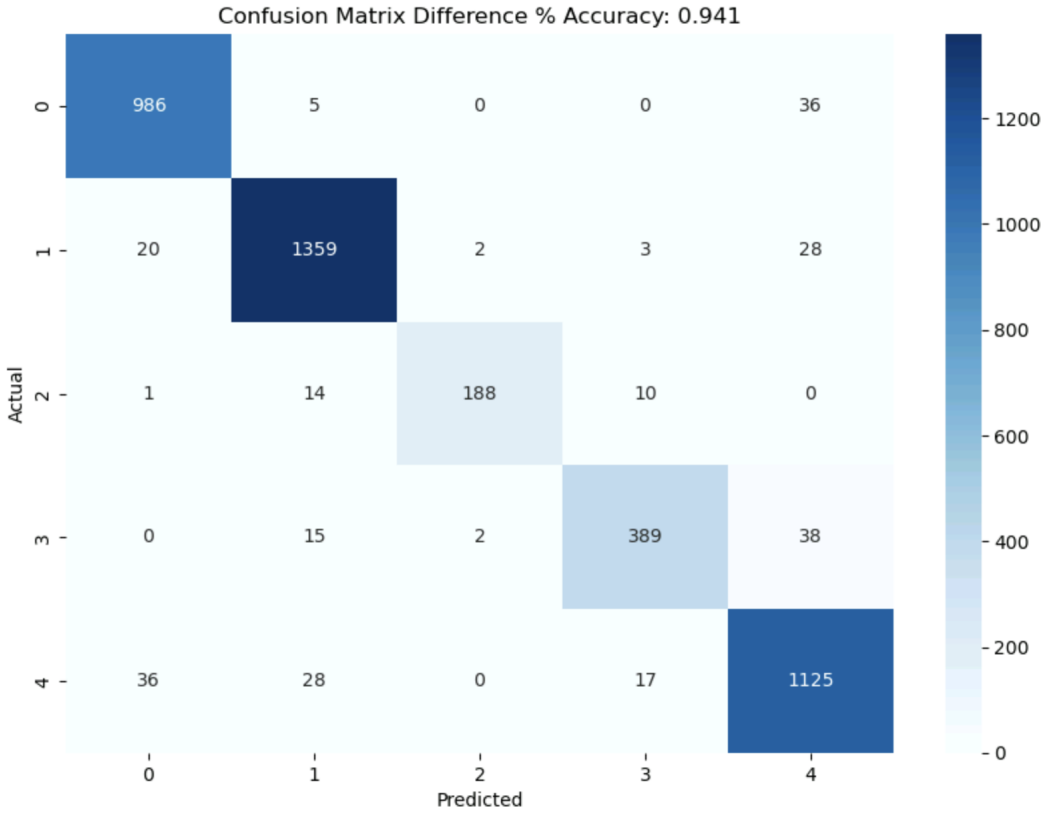

Figure A.12: Confusion Matrix Alternative 10m-5m Baseline Difference approach