Applied Medical Informatics
Original Research Vol. 46, No. 4 /2024, pp: 129-139

Time-series of White Blood Cells and Cancer Antigen 15-3 in a Metastatic Breast Cancer Patient, Referencing Clustering Algorithm Assessment

Alexandros CLOUVAS *

Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Egnatia Street, GR-54124 Thessaloniki, Greece.
E-mail : clouvas@ece.auth.gr

*Author to whom correspondence should be addressed.

Received: 18 November 2024/Accepted: 23 December 2024/ Published online: 26 December 2024

Abstract
Clustering is a powerful tool for analyzing time-series data related to cancer markers. Cancer Antigen 15-3, a glycoprotein commonly associated with breast cancer, was a key focus of this study. Time-series data of White Blood Cells (WBC) and Cancer Antigen 15-3 from a metastatic breast cancer patient were utilized as references to assess the clinical relevance of nine clustering algorithms. The WBC time-series measurements were particularly noteworthy because the specific therapy administered to the patient clearly defined all clustering parameters, including the number of clusters and their boundaries. This provided a unique opportunity to rigorously evaluate the performance of the nine clustering algorithms. Remarkably, seven to nine of these algorithms achieved perfect alignment between the observed clusters and the expected results based on the therapy. The application of these clustering methods to the time-series data revealed that K-means with Euclidean distance, K-means with Manhattan distance, and K-medoids were the most effective algorithms for analyzing both Cancer Antigen 15-3 and WBC time-series data.

Keywords*:* Cluster analysis; K-means; Time-series; Cancer Antigen CA 15-3 (CA 15-3); Neutrophils; White Blood Cells (WBC); Male breast cancer

Introduction

The application of unsupervised machine learning algorithms presents unprecedented opportunities to uncover patterns in clinical biomarkers, enhancing healthcare outcomes and providing new insights into disease mechanisms [1]. Cancer Antigen 15-3 (CA 15-3) is a glycoprotein commonly associated with benign and malignant breast diseases, particularly in cases of breast cancer involving liver or bone metastases. Identified as a potential breast cancer marker in the 1980s [2], CA 15-3 has since been extensively studied, ultimately becoming a routine marker in clinical practice for breast cancer ([3] and references therein). Notably, this study [3] demonstrated that systematic monitoring of CA 15-3 facilitated the early detection of metastatic disease in 37% of cases through rising CA 15-3 levels. However, current guidelines [4–6] advise against serial CA 15-3 measurements during early breast cancer follow-up due to limited evidence of a survival benefit. Despite these recommendations, many oncologists continue to perform serial assessments of CA 15-3 levels in asymptomatic patients with early-stage breast cancer.

In a recent study [7], the authors of the current work monitored plasma CA 15-3 levels over time in a male patient with metastatic breast cancer. This study was the first to apply cluster analysis to CA 15-3 time-series data,

revealing two distinct clusters, one corresponding to the pre-recurrence phase and the other to the metastatic stage. The transition between these clusters provides valuable insight into the onset of metastasis.

The current study marked a significant advancement, as time-series measurements of both White Blood Cells (WBC) and CA 15-3 from the same patient (as in the previous study) were used to evaluate the clinical relevance of various clustering algorithms. WBC time-series measurements were particularly valuable; due to the specific therapy administered, clustering parameters such as the number and boundaries of clusters were well-defined, enabling rigorous testing of different clustering algorithms. Additionally, applying various clustering methods to CA 15-3 time-series data identified the most suitable algorithms for analyzing these specific time-series.

Materials and Methods

Medical History of the Patient

The medical history of this 69-year-old male patient is detailed in a previous study [7]. In brief, the patient underwent right mastectomy on April 27, 2020, due to a malignant tumor. Thirty-seven months later, on May 18, 2023, elevated CA 15-3 levels (37.2 U/mL) were detected for the first time. As the CA 15-3 levels continued to rise, a Positron Emission Tomography/Computed Tomography (PET/CT) scan was performed on August 11, 2023, revealing metastatic breast cancer in the trochanteric region of the left leg. This metastasis was classified as grade 3 according to the Elston and Ellis grading system [8] and showed Estrogen Receptor (ER) and Progesterone Receptor (PR) positivity, Human Epidermal Growth Factor Receptor 2 (HER2) expression at 20%, and a Ki-67 (MIB1) proliferation index of 30%. Following this diagnosis, the medical team at the "Theagenio" Cancer Hospital in Thessaloniki, Greece initiated the following treatment regimen for the patient:

1. Letrozole (Letrozole Teva 2.5 mg) in combination with palbociclib (Ibrance 125 mg).
2. Two monthly injections (XGEVA 120 mg/1.7 ml 1 VIAL) (ARVEKAP 3.75 mg INJ).
3. Radiotherapy (10 sessions, with each session delivering 3 Gy of radiation, totaling an absorbed dose of 30 Gy). Additionally, prior to commencing radiotherapy, prophylactic threading with a long-gamma nail was performed on September 13, 2023.

Palbociclib, in combination with Letrozole, is a widely used first-line treatment for patients with hormone receptor-positive (HR+), HER2-negative advanced or metastatic breast cancer. However, one of the most common side effects of palbociclib is a reduction in the white blood cell (WBC) count due to a decrease in neutrophils, which are a key component of WBCs.

Time-series measurement of WBC and plasma CA 15-3

The patient's WBC count and plasma CA 15-3 levels were monitored using regular blood tests. Over the past four years, 28 blood tests for WBC levels were conducted between October 15, 2020, and September 19, 2024, while 37 blood tests for CA 15-3 levels were performed between April 14, 2020, and September 19, 2024, across two microbiology laboratories. CA 15-3 levels were measured using an enhanced chemiluminescence immunoassay in the first laboratory and a microparticle chemiluminescence immunoassay in the second [7]. To ensure accuracy, measurements were cross-validated three times by conducting tests at both laboratories with no more than a two-day interval between each paired test. Discrepancies in results between the two laboratories occurred at rates of 11.3%, 1.9%, and 3.6%, respectively [7]. White blood cell (WBC) count was measured using a complete blood count test, a standard procedure that provides information on various types of cells in the blood.

Because time-series measurements of WBC and plasma CA 15-3 were collected at irregular intervals, a combination of linear interpolation and moving average smoothing techniques was employed to preprocess the data and convert it into regularly spaced intervals. A 30-day interval was used for the WBC time-series, whereas a 15-day interval was applied to the plasma CA 15-3 series.

Cluster Analysis

Cluster analysis of the WBC and plasma CA 15-3 time-series was conducted on the Wolfram Cloud platform provided by Wolfram Research [9], using the Wolfram Language for computational tasks. The processed data were sampled at regular intervals of 30 days for WBC and 15 days for CA 15-3, and structured as pairs (x,y), where

x represents the number of days since the initial measurement (October 15, 2020) for WBC and April 14, 2020, for CA 15-3 and y corresponds to the WBC values (in k/μL) and CA 15-3 values (in U/mL), respectively This analysis utilized the built-in "Find Cluster" function. The time-series processed data (x,y) are provided as input to the "Find Clusters" function. The function internally treats each data point (x,y) as an entity to be clustered. The "Find Clusters" function offers a variety of clustering algorithms. In this study, the following clustering algorithms were applied:

1. **K-means algorithm** [10] (with Euclidean distance): The K-means algorithm aims to minimize within-cluster variance, which is the sum of the squared Euclidean distances between each point in a cluster and its centroid. This minimization process is achieved through iterative refinement until convergence, where the assignments of data points to clusters and the positions of the cluster centroids stabilize. In the context of clustering, similarity refers to the degree of resemblance or proximity between the data points within the same cluster. Data points that are similar to each other are grouped together into the same cluster, whereas dissimilar data points are placed into different clusters. The number of clusters should be predefined.
2. **K-means algorithm** [10] (with Manhattan distance): The K-means algorithm with Manhattan distance is a variation of the standard K-means clustering algorithm, where the metric used to measure the "closeness" between data points is the Manhattan distance instead of the Euclidean distance. While Euclidean distance measures the straight-line distance between two points, Manhattan distance, also known as taxicab or city block distance, measures the distance between two points along the axes at right angles, following a grid-like path that moves only vertically and horizontally. Similar to standard K-means, this variation partitions data into K clusters; however, the use of the Manhattan distance affects both point assignments to clusters and centroid updates. As with K-means clustering, the number of clusters must be predefined.
3. **K medoids** [11]: The K-medoids algorithm is a clustering method similar to K-means, but with one key difference: instead of using the mean of the data points to represent the centroid of each cluster, K-medoids uses an actual data point (called a medoid) as the cluster center. This makes K-medoids more robust to outliers than K-means, because medoids are less affected by extreme values**.** The number of clusters should be predefined**.**
4. **DBSCAN** [12]: DBSCAN is a density-based clustering algorithm that groups points that are closely packed together while marking points that lie alone in low-density regions as outliers (noise). Unlike K-means, DBSCAN does not require that the number of clusters be specified beforehand. It is well suited for datasets with clusters of similar density and is robust to noise and outliers.
5. **Neighborhood Contraction** [13]: The Neighborhood Contraction method is a density-based clustering algorithm that typically merges clusters iteratively by contracting the neighborhood of a point to find the most representative center or centroid for each cluster. This method is generally more suitable for data with complex structures, where clusters may not have spherical or compact shapes.
6. **Jarvis-Patrick** [14]: Jarvis-Patrick clustering is a density-based clustering algorithm that groups data points based on their shared nearest neighbors. This algorithm is primarily used to identify clusters in datasets that might not be spherical or regularly shaped, and it can effectively handle noisy or sparse data. This is particularly useful when the goal is to find clusters based on the similarity of data points, where each data point is clustered with others that have a high number of nearest neighbors. Jarvis-Patrick does not require that the number of clusters be specified beforehand.
7. **Mean Shift** [15]: Mean Shift is a non-parametric clustering algorithm that does not require the number of clusters to be specified in advance. It is based on the idea of shifting a window or "kernel" iteratively to find the mode (or peak) of the data distribution that represents the center of a cluster. This algorithm is often used for image segmentation, object tracking, and density estimation tasks, particularly when the number of clusters is unknown or when the clusters have arbitrary shapes.
8. **Spectral** [16]: Spectral clustering is a popular clustering technique that uses the properties of graphs and linear algebra to identify clusters. Unlike traditional clustering methods such as K-means, which often rely on distance-based measures, spectral clustering is especially effective for datasets with non-convex or complex cluster structures.
9. **Gaussian Mixture** [17]: The Gaussian Mixture Model (GMM) clustering algorithm is a probabilistic model that assumes that data points are generated from a mixture of several Gaussian distributions, each representing a cluster. Unlike K-means clustering, which assigns points to clusters by minimizing the

Euclidean distance, GMM clustering assigns points probabilistically to multiple clusters, based on their likelihood.

Clustering methods 4–9 do not require the number of clusters to be specified in advance, whereas methods 1–3 require a predefined cluster count. In this study, the optimal number of clusters for methods 1–3 was determined using the elbow method [18]. To identify this optimal number, the K-means algorithm was run for various values of K (the number of clusters). For each K, the Within-Cluster Sum of Squares (WCSS), which quantifies the average squared distance of all points in a cluster from the cluster centroid, is calculated. A plot of WCSS versus K was then generated, and the optimal cluster number corresponds to the point where the rate of decrease in WCSS notably slows, creating an "elbow" shape in the plot. After determining the optimal number of clusters, the final computations were conducted using clustering algorithms 1 to 3. In contrast, for algorithms 4–9, the computations were more straightforward, as these methods do not require the number of clusters to be specified beforehand.

The quality of clustering was evaluated using the Silhouette Score [19], which is a valuable metric for assessing clustering performance. This score quantifies how well each data point fits into its assigned cluster and how separate it is from the other clusters. The silhouette score (s) ranges from -1 to 1, with higher values indicating better-defined and more appropriate clusters. A score close to +1 indicates that the data points are well clustered. Specifically:

1. $s \geq 0.7$, indicating a strong clustering. The clusters were well defined, with a clear separation between them. Data points are closely grouped within their own clusters and are far from the neighboring clusters.
2. $0.5 \leq s < 0.7$: reasonably good clustering. The clusters are well separated, although there might be a slight overlap or proximity between neighboring clusters.
3. $0.25 \leq s < 0.5$: moderate clustering. Clustering is acceptable but not very distinct. Some points may be close to other clusters, suggesting some overlap or less clear cluster boundaries.
4. $s \approx 0$: The data points are near or on the decision boundary between clusters, being equally close to two or more clusters, implying overlapping or ambiguous boundaries.
5. s close to -1: Misclassified data points. These points are closer to the center of another cluster than their assigned cluster, typically indicating poor clustering, where data points do not naturally fit into the assigned cluster.

Results

Time-Series of WBC and CA 15-3 Results

Figure 1 shows the time-series measurements of WBC. Open circles represent raw WBC measurements, while filled circles represent measurements processed using a combination of linear interpolation and moving average smoothing techniques to transform the irregularly sampled data into regularly spaced 30-day intervals. The vertical dashed line marks September 15, 2023, when the patient began treatment with palbociclib (Ibrance, 125 mg). As noted previously, palbociclib commonly causes several side effects, with a reduction in WBC being the most frequent. Therefore, it is important to understand the patterns shown in Figure 1. The group of values before September 15, 2023, is expected to be significantly higher than the group of values after this date. This suggests that in a cluster analysis of the WBC time-series measurements, two clusters should be anticipated, with September 15, 2023, marking the boundary between them. Therefore, the pattern in Figure 1 allows for the rigorous testing of different clustering algorithms, as both the expected number of clusters and the boundaries between them are clearly defined.

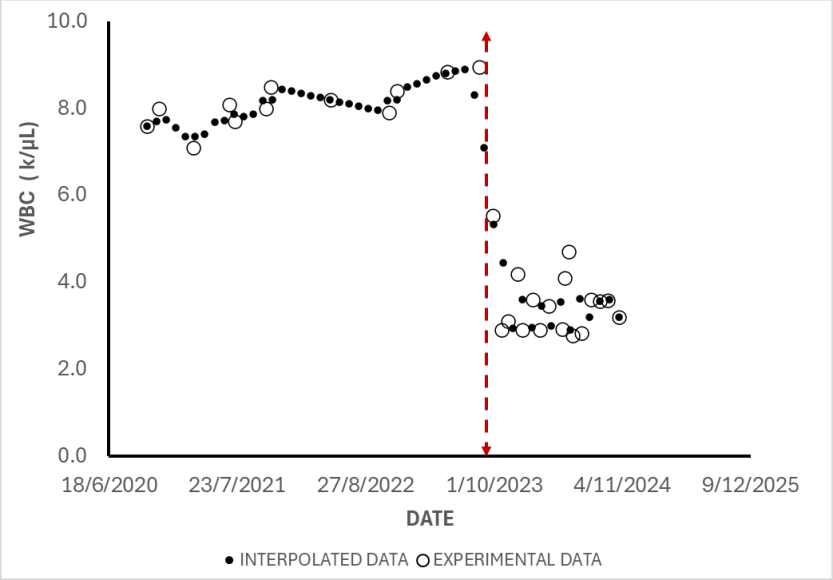

Figure 1. Open circles represent the raw WBC measurements, while filled circles represent measurements processed using a combination of linear interpolation and moving average smoothing techniques to transform the irregularly sampled data into regularly spaced 30-day intervals. The vertical dashed line marks September 15, 2023, the date when the patient began treatment with palbociclib (Ibrance, 125 mg).

Figure 2 shows the time-series measurements of plasma CA 15-3. Open circles represent raw CA 15-3 measurements, while filled circles represent measurements processed using a combination of linear interpolation and moving average smoothing techniques to transform the irregularly sampled data into regularly spaced 15-day intervals.

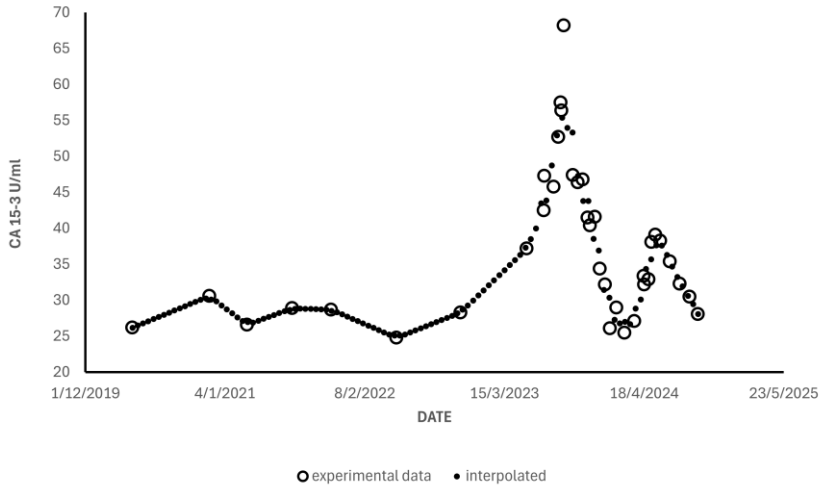

Figure 2. Open circles represent the raw plasma CA 15-3 measurements, while filled circles represent measurements processed using a combination of linear interpolation and moving average smoothing techniques to transform the irregularly sampled data into regularly spaced 15 -day intervals.

The marked increase observed in July–August 2023 was attributed to the presence of metastatic cancer in the trochanteric region of the patient's left leg. Conversely, a significant decrease reflected the effectiveness of the treatment administered by oncologists. The continuous rise in CA 15-3 levels from 30 U/mL began at the end of 2022. This was approximately 7.5 months before the PET/CT scan revealed metastatic breast cancer in the trochanteric region of the left leg. Finally, there is no clear explanation from oncologists for the smaller second peak in CA 15-3 levels shown in Figure 2. According to them, several factors could cause temporary false-positive

results, including infection, trauma, and other conditions.

Cluster Analysis Results

To enhance readability, Table 1 lists the code numbers (C/N) of the nine clustering algorithms used in this study.

Table 1. Code Numbers (C/N) and Names of the 9 clustering algorithms used in this work.

C/N	Name	C/N	Name	C/N	Name
1	K-means (with Euclidean distance)	**4**	DBSCAN	**7**	Mean Shift
2	K-means (with Manhattan distance)	**5**	Neighborhood Contraction	**8**	Spectral
3	K medoids	**6**	Jarvis-Patrick	**9**	Gaussian Mixture

The number of clusters (K) must be specified in advance for the 1-3 clustering algorithms. In this study, the optimal number of clusters was determined using the elbow method. Figures 3 and 4 show the Within-Cluster Sum of Squares (WCSS) values for WBC and CA 15-3, respectively, plotted against the number of clusters. The "elbow" point in these plots indicates the optimal number of clusters, occurring where the rate of decrease in WCSS significantly slows, forming an "elbow" shape. Figures 3 and 4 clearly show that the optimal number of clusters is K=2 for both WBC and CA 15-3.

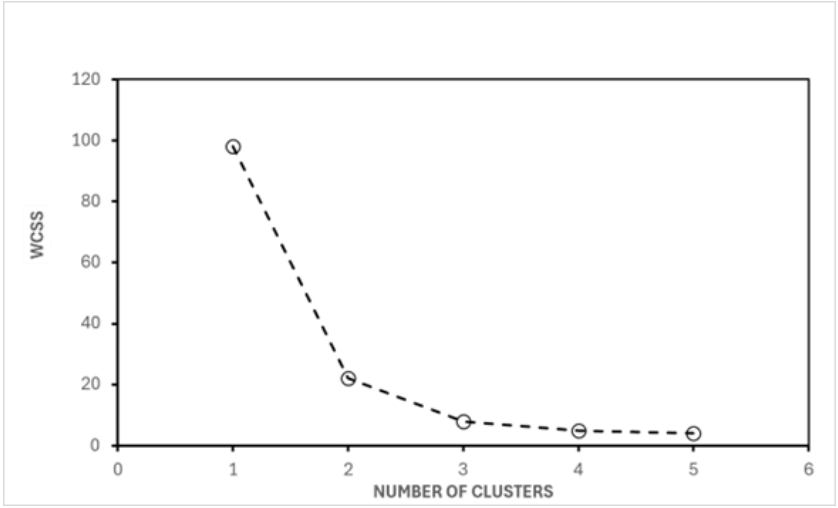

Figure 3. The WCSS values against the number of clusters for WBC.

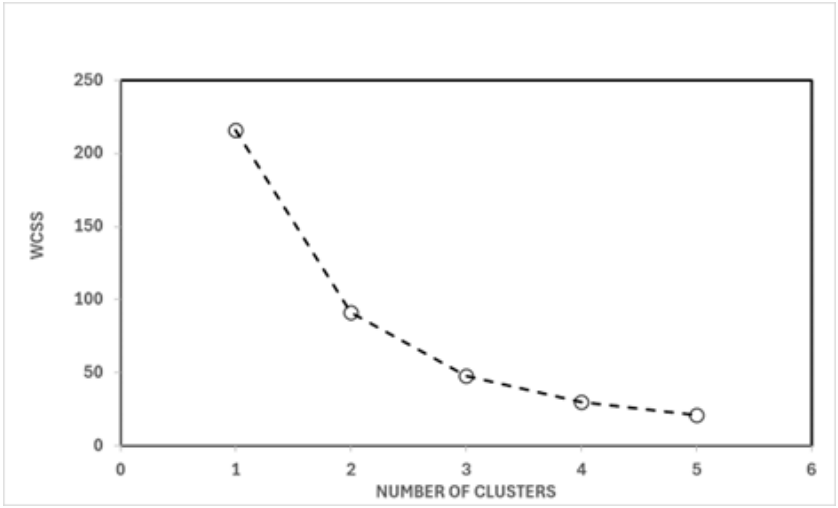

Figure 4. The WCSS values against the number of clusters for CA 15-3

Figure 5 presents the clustering analysis of WBC-processed data sampled at 30-day intervals using clustering algorithms with Code Numbers 1, 2, 3, 4, 5, 6, and 9. All these algorithms consistently identified two distinct clusters. Closed circles denote WBC values in cluster 1, whereas filled circles represent values in cluster 2. The vertical dashed line marks September 15, 2023, the start date of the patient's treatment with palbociclib (Ibrance 125 mg). This clear distinction between clusters shows that the algorithms with Code Numbers 1, 2, 3, 4, 5, 6, and 9 effectively separated the WBC time-series data into two periods: period A, corresponding to the pre-treatment phase, and period B, representing the post-treatment phase with palbociclib. Additionally, the Silhouette Score for the WBC time-series cluster analysis using these clustering algorithms was 0.509, indicating reasonably good clustering. This score suggests that the clusters are fairly well separated, although there may be a slight overlap or close proximity between neighboring clusters.

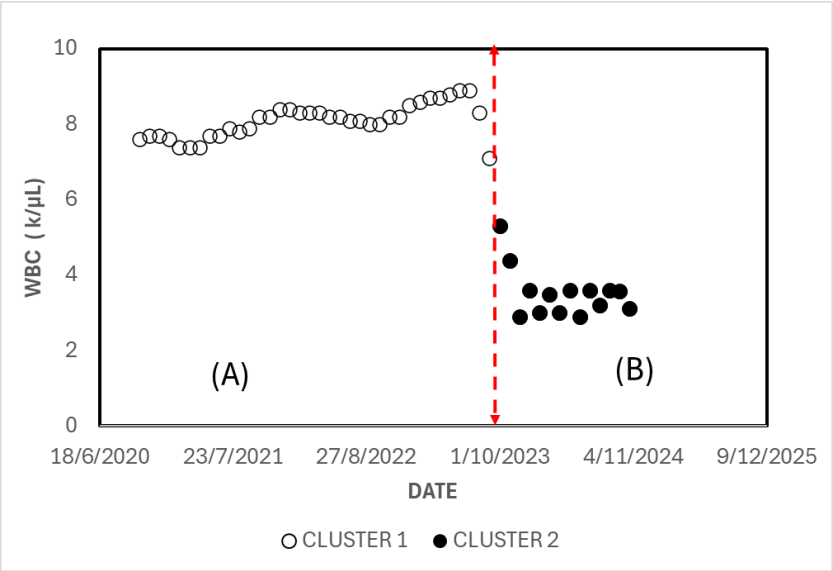

Figure 5. Identification of two clusters in the WBC time-series data, processed and sampled at 30-day intervals using clustering algorithms with Code Numbers 1, 2, 3, 4, 5, 6, and 9. Open circles represent measurements associated with cluster 1, while filled circles denote those in cluster 2. The vertical dashed line marks September 15, 2023, the start date of the patient's treatment with palbociclib (Ibrance, 125 mg).

Cluster analysis of the WBC time-series using clustering algorithms with code numbers 7 and 8 failed to identify the two expected clusters in the data. Specifically, Algorithm 7 detected 10 clusters, whereas Algorithm 8 identified nine clusters.

Table 2 shows the number of clusters identified in the CA 15-3 time-series data, which were processed and sampled at 15-day intervals using the nine clustering algorithms. Notably, for clustering algorithms 1–3, the number of clusters was predetermined using the elbow method (Figure 4).

Figures 6 and 7 present the clustering analysis of the processed CA 15-3 data, sampled at 15-day intervals, using algorithms designated by code numbers 1–3 and 9, respectively. Closed circles represent CA 15-3 values assigned to Cluster 1, while filled circles denote values in Cluster 2. The analysis achieved a Silhouette Score of 0.594, indicating reasonably good clustering performance. Clustering algorithms with code numbers 1–3 effectively identified two distinct clusters, corresponding to the pre-recurrence phase and the metastatic stage [7]. Similarly, the clustering algorithm with code number 9 also identified two clusters. Although the feature highlighted in Figure 7 is intriguing and will be discussed in the following section, it lacks clinical relevance.

Table 2. Number of clusters identified in the CA 15-3 time-series data using the nine clustering algorithms.

C/N	Name	Number of Clusters
1	K-means (with Euclidean distance)	2
2	K-means (with Manhattan distance)	2
3	K medoids	2
4	DBSCAN	7
5	Neighborhood Contraction	4
6	JarvisPatrick	25
7	Mean Shift	15
8	Spectral	9
9	Gaussian Mixture	2

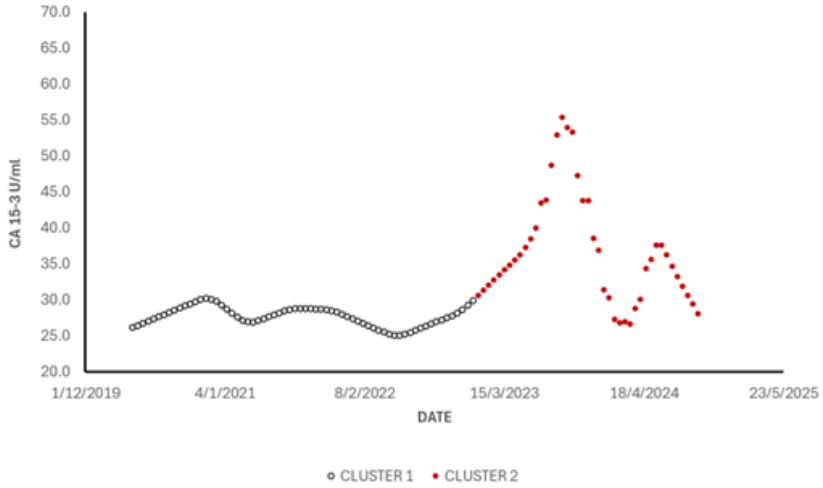

Figure 6. Clustering analysis of CA 15-3 processed data, sampled at 15-day intervals, using clustering algorithms with Code Numbers 1, 2, and 3. Open circles represent CA 15-3 values in cluster 1, while filled circles denote values in cluster 2.

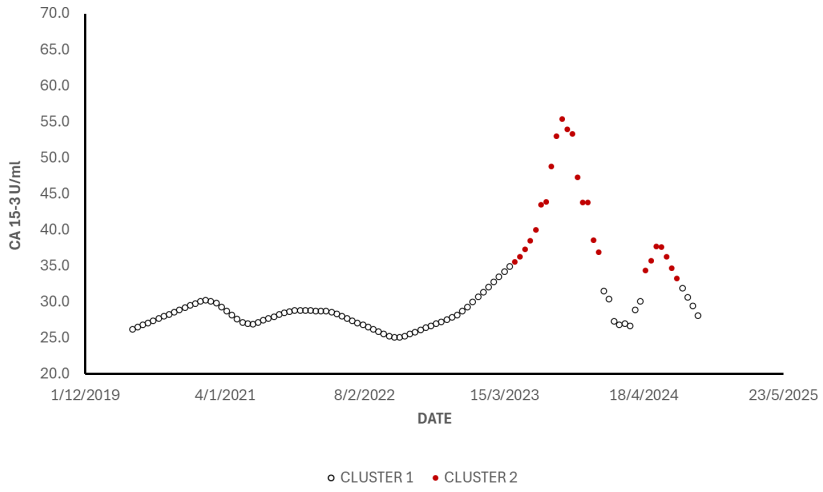

Figure 7. Clustering analysis of CA 15-3 processed data, sampled at 15-day intervals, using clustering algorithm with Code Number 9. Open circles represent CA 15-3 values in cluster 1, while filled circles denote values in cluster 2.

Discussion

Cluster analysis of the WBC time-series using nine clustering algorithms yielded promising results. Most algorithms, specifically those numbered 1, 2, 3, 4, 5, 6, and 9, successfully identified the two distinct clusters expected within the WBC time series. The first cluster spans October 15, 2020, to August 31, 2023, and the second extends from September 30, 2023, to the most recent measurement on September 24, 2024. Notably, the last WBC data point in Cluster 1 was recorded 15 days before the patient began treatment with palbociclib, whereas the first WBC data point in Cluster 2 was recorded 15 days after treatment commenced. Additionally, the WBC values in Cluster 1 were more than twice as high as those in Cluster 2 were. Because a significant reduction in WBC levels is a well-known side effect of palbociclib, and the boundary between the two clusters aligns with the treatment start date, these clustering results underscore the clinical relevance of algorithms 1, 2, 3, 4, 5, 6, and 9 for analyzing the specific WBC time-series data. By contrast, clustering algorithms with code numbers 7 and 8 failed to identify the two expected clusters in the time-series data. Instead, Algorithm 7 detects 10 clusters, while Algorithm 8 identifies nine clusters.

Clustering algorithms 4, 5, 6, 7, and 8 failed to identify the expected two clusters in the CA 15-3 time-series data (see Table 2), which represent the two distinct phases: the pre-recurrence stage and the metastatic stage. Specifically, the clustering algorithms Jarvis-Patrick (code number 6) and Mean Shift (code number 7) produced a disproportionately high number of clusters—25 and 15, respectively—compared to the anticipated two clusters. While the Jarvis-Patrick and Mean Shift algorithms are indeed commonly applied to large datasets due to their clustering efficiency and ability to handle complex structures, it is worth noting that their performance may vary depending on the dataset's characteristics. For smaller datasets, these algorithms might not always provide meaningful or interpretable clusters. However, it is worth noting that although the Jarvis-Patrick algorithm identified a disproportionately high number of clusters in the CA 15-3 time-series data, its performance was successful for the WBC time-series cluster analysis. On the other hand, the Mean-Shift algorithm also failed in the cluster analysis of the WBC time-series, identifying ten clusters instead of the expected two.

Although clustering algorithm 9 (Figure 6) successfully identified two clusters in the CA 15-3 time-series, the clustering lacked clinical relevance. Instead of differentiating between the pre-recurrence and metastatic stages, it merely segregated CA 15-3 values based on their levels, with lower values assigned to Cluster 1 and higher values to Cluster 2. However, this approach was problematic. For instance, from November 1, 2024, to March 21, 2024, a period within the metastatic stage, there is a subset of four CA 15-3 measurements ranging between 26 and 29 U/mL, with an average of 27 U/mL. These values show similar variation and mean as those in the pre-recurrence stage (27.6 ± 1.6 U/mL), yet they do not indicate a return to the pre-recurrence stage. However, in certain cases, this algorithm may offer some utility in tracking CA 15-3 values to monitor illness progression, provided its limitations are clearly understood.

Clustering algorithms 1, 2, and 3 successfully identified the two expected clusters in the CA 15-3 time-series data. Data points in Cluster 1 correspond to the period before the recurrence of the disease (metastatic breast cancer), whereas those in Cluster 2 represent the advanced (metastatic) stage of the disease. The Silhouette Score for this clustering analysis was 0.594, indicating reasonably good clustering performance. A difference was observed in the boundary between the two clusters identified in this study compared with a previous study [7]. In this analysis, time-series measurements were processed using a combination of linear interpolation and moving average smoothing, converting the irregularly measured data into regularly spaced 15-day intervals, while the previous study used the original irregularly measured CA 15-3 data. In the present study, the first CA 15-3 value associated with Cluster 2 was observed at the end of December 2022, whereas in a previous study [7], it was noted on November 9, 2022. These dates correspond to 7.5 months and 9 months, respectively, before recurrence was confirmed via PET/CT. This aligns well with the findings of De Cock et al. [3], who reported a gradual increase in CA 15-3 levels 6–12 months prior to the detection of the first metastases.

In summary, three algorithms—K-means with Euclidean distance, K-means with Manhattan distance, and K-medoids—demonstrated clinical relevance for analyzing both WBC and CA 15-3 time-series data. However, it is important to note that this study was limited to a single patient, and these findings cannot be generalized without further research involving additional patients. Nevertheless, this study represents a step forward toward the ultimate goal of investigating the potential to detect the onset of the metastatic stage through cluster analysis of

CA 15-3 time-series data.

Conclusions

This study utilized time-series measurements of White Blood Cells (WBC) and plasma Cancer Antigen 15-3 (CA 15-3) from a patient with metastatic breast cancer to evaluate the clinical relevance of nine clustering algorithms. Among these, K-means with Euclidean distance, K-means with Manhattan distance, and K-medoids proved clinically relevant for analyzing both WBC and CA 15-3 time-series data. However, the study was limited to a single patient. To validate and generalize these findings, further research involving additional patients and time-series datasets is necessary.

List of Abbreviations: CA 15-3: Cancer Antigen 15-3; WBC: White Blood Cells; DBSCAN: Density-Based Spatial Clustering of Applications with Noise; PR: Progesterone Receptor; ER: Estrogen Receptor; HER2: Human Epidermal Growth Factor Receptor 2; GMM: Gaussian Mixture Model; WCSS: Within-Cluster Sum of Squares; PET/CT: Positron Emission Tomography/Computed Tomography.

Author Contributions: Not applicable, as there is only one author.

Funding: This research received no external funding.

Ethics Statement: Not applicable.

Data Availability Statement: The author declares that the data supporting the findings of this study are available within the paper as figures. The computations for clustering the WBC and CA 15-3 time-series data using the nine clustering algorithms were carried out on the Wolfram Cloud platform using the Wolfram Language.

Acknowledgments: The author would like to thank the medical staff of the following departments for the high-quality care provided during his treatment: 1) the C' Department of Clinical Oncology and Chemotherapy at Theagenio Cancer Hospital of Thessaloniki, 2) the 3rd Department of Orthopedics at Papageorgiou Hospital, and Dr. Anna Makridou, Chief of the Medical Physics Department at Theagenio Cancer Hospital of Thessaloniki.

Conflict of Interest: The author declares no conflict of interest.

References

1. Mariam A, Javidi H, Zabor EC, Zhao R, Radivoyevitch T, Rotroff DM. Unsupervised clustering of longitudinal clinical measurements in electronic health records. PLOS Digit Health. 2024;3(10):e0000628. doi: 10.1371/journal.pdig.0000628.
2. Gang Y, Adachi I, Ohkura H, Yamamoto H, Mizuguchi Y, Abe K. [CA 15-3 is present as a novel tumor marker in the sera of patients with breast cancer and other malignancies]. Gan To Kagaku Ryoho. 1985;12(12):2379-86. Japanese.
3. De Cock L, Heylen J, Wildiers A, Punie K, Smeets A, Weltens C, Neven P, Billen J, Laenen A, Wildiers H. Detection of secondary metastatic breast cancer by measurement of plasma CA 15.3. ESMO Open. 2021;6(4):100203. doi: 10.1016/j.esmoop.2021.100203.
4. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, Zackrisson S, Senkus E; ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2019;30(8):1194-1220. doi: 10.1093/annonc/mdz173.
5. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr; American Society of Clinical Oncology. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol. 2007;25(33):5287-312. doi: 10.1200/JCO.2007.14.2364.
6. Khatcheressian JL, Hurley P, Bantug E, Esserman LJ, Grunfeld E, Halberg F, Hantel A, Henry NL, Muss HB, Smith TJ, Vogel VG, Wolff AC, Somerfield MR, Davidson NE; American Society of Clinical Oncology. Breast cancer follow-up and management after primary treatment: American Society of Clinical Oncology clinical

practice guideline update. J Clin Oncol. 2013;31(7):961-5. doi: 10.1200/JCO.2012.45.9859.
7. Clouvas, A. (2024) "Exploring the Significance of Cluster Analysis on Time-Series Measurement of Plasma Cancer Antigen 15-3 in a Patient with Metastatic Breast Cancer", Applied Medical Informatics, 46(3). Available at: https://ami.info.umfcluj.ro/index.php/AMI/article/view/1064 (Accessed: 28 October 2024).
8. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-10. doi: 10.1111/j.1365-2559. 1991.tb00229.
9. Wolfram Research, Inc., Wolfram|Alpha Notebook Edition, Champaign, IL. 2024.
10. MacQueen J.B. Some methods for classification and analysis of multivariate observations Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1967 University of California Press 281-297.Available from: http://projecteuclid.org/euclid.bsmsp/1200512992 (c) (Accessed: 28 October 2024).
11. Kaufman, Leonard; Rousseeuw, Peter J. (1990-03-08), "Partitioning Around Medoids (Program PAM)", Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 68–125, doi:10.1002/9780470316801.ch2.
12. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise (PDF). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).
13. Wolfram Language & System Documentation Center "NeighborhoodContraction" (Machine Learning Method) https://reference.wolfram.com/language/ref/method/NeighborhoodContraction.html (Accessed: 28 October 2024).
14. Jarvis R.A, Patrick E.A. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. IEEE Transactions on Computers, 22(11):1025-1034. doi: 10.1109/T-C.1973.223640
15. Wolfram Language & System Documentation Center "MeanShift" (Machine Learning Method) https://reference.wolfram.com/language/ref/method/MeanShift.html (Accessed: 28 October 2024).
16. Wolfram Language & System Documentation Center "Spectral" (Machine Learning Method) https://reference.wolfram.com/language/ref/method/Spectral.html (Accessed: 28 October 2024).
17. Wolfram Language & System Documentation Center "GaussianMixture" (Machine Learning Method) https://reference.wolfram.com/language/ref/method/GaussianMixture.html(Accessed: 28 October 2024).
18. Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means clustering. Int J Adv Res Comput Sci Manag Stud. 2013;1:2321-7782.
19. Rousseeuw PJ. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7.