

A comparative analysis of clustering methods for identifying patient subgroups in Chronic Kidney Disease using feature engineering
Kandasamy, N., Basukoski, A. and Chausalet, T.

This is a copy of the author's accepted version of a paper subsequently published in the proceedings of the 38th IEEE International Symposium on Computer-Based Medical Systems (IEEE CBMS 2025). Madrid, Spain, 18 - 20 Jun 2025.

The final published version will be available online at:

<https://doi.org/10.1109/CBMS65348.2025.00023>

© 2025 IEEE. This manuscript version is made available under the CC-BY 4.0 license

<https://creativecommons.org/licenses/by/4.0/>

For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

The WestminsterResearch online digital archive at the University of Westminster aims to make the research output of the University available to a wider audience. Copyright and Moral Rights remain with the authors and/or copyright owners.

A comparative analysis of clustering methods for identifying patient subgroups in Chronic Kidney Disease using feature engineering.

Nichalini Kandasamy
School of Computer Science and
Engineering
University of Westminster
London, United Kingdom
kandasn@westminster.ac.uk

Dr. Artie Basukoski
School of Computer Science and
Engineering
University of Westminster
London, United Kingdom
A.basukoski@westminster.ac.uk

Prof. Thierry Chausalet
School of Computer Science and
Engineering
University of Westminster
London, United Kingdom
chausst@westminster.ac.uk

Abstract— Chronic Kidney Disease (CKD) is a chronic disease that progressively deteriorates health conditions of the affected patients. It also poses substantial challenges for clinicians in accurately predicting disease progression trajectories, detecting cohorts of patients and anticipating patient management strategies. This study investigates the potential benefits of using unsupervised clustering techniques as a feature engineering method to enhance the identification of subgroups in CKD patients. Specifically, we evaluate the performance of four frequently used clustering algorithms, each with different characteristics and speciality - K-Means, HDBSCAN, Hierarchical Clustering, and Gaussian Mixture Models - on a publicly available clinical dataset (MIMIC IV) formed of detailed real-world records related to CKD patients, including relevant laboratory test results, demographic details, and sequences of ICD-10 codes. The analysis indicates that complex clustering algorithms, like GMM and HDBSCAN, do not have a higher accuracy in detecting clusters on the given clinical dataset than simpler algorithms such as K-Means and Hierarchical clustering.

Keywords — *KMeans, HDBSCAN, Hierarchical Clustering, GMM, MIMIC IV, Chronic Kidney Disease, ICD-10 Codes*

I. INTRODUCTION

Chronic Kidney Disease, also known as CKD, is a health condition characterized by a gradual and progressive deterioration of renal function, often accompanied by a range of comorbidities and adverse outcomes that develop throughout the deterioration of the patient's health. It is estimated that 700 million patients are affected by Chronic Kidney Disease worldwide [1]. However, the complexity of the disease and the progression pattern being unique to each patient, poses significant challenges for clinicians in accurately predicting disease trajectories and developing preventive treatment at an earlier stage of the disease progression. This complexity and variability in CKD progression indicates the imperative need for novel analytical approaches that can effectively capture the underlying patterns and support clinicians in delivering treatments that are tailored for each patient based on the risk and level of progression of the disease.

One suggestive path for improving the prediction of CKD progression is the application of unsupervised clustering techniques as a feature engineering method prior to the predictive task. Clustering methods have the potential to identify distinct subgroups within the CKD population, each with unique disease progression profiles and associated risk factors. By identifying these clustering-derived subgroups and using the cluster labels into predictive models, researchers may be able to improve the granularity and accuracy of predicting severe disease trajectories, enabling the development of personalized risk assessment frameworks and targeted treatment strategies tailored to the specific needs and characteristics of individual patients.

In this study, we conduct a comparative evaluation of four widely used clustering algorithms - K-Means, HDBSCAN, Hierarchical Clustering, and Gaussian Mixture Models - to explore their potential as feature engineering tools for identifying subgroups of patients with similar characteristics. Leveraging a comprehensive dataset comprising of detailed clinical records, including laboratory test results, demographic information, and sequences of ICD-10 codes, we investigate the performance of these clustering methods in uncovering meaningful subgroups within the CKD population. The presented analysis forms a component of an overall hybrid framework that is being developed to predict disease trajectories in patients with Chronic Kidney Disease, using Process Mining and Machine Learning. The output of this study will be sourced as a feature in the downstream sequential predictive activities.

II. EXISTING LITERATURE REVIEW

Extensive research has highlighted the significant heterogeneity in the presentation and progression of Chronic Kidney Disease, presenting the need for more sophisticated predictive approaches to capture the complexity and support medical practitioners in delivering personalized, risk-based preventive treatments. As mentioned earlier, one feasible method for optimizing the prediction of CKD progression is

the prior application of unsupervised clustering techniques as a feature engineering method.

Clustering methods have the potential to identify distinct subgroups within the CKD population, each with unique disease progression aspects and associated risk factors. D.T. Anh Luong et al. presented a modified version of the K-Means algorithm to cluster Chronic Kidney Disease progressions and identify disease subtypes with the objective to anticipate future diagnosis [2]. The algorithm has been adapted to cluster on time series data as clinical time series are scattered and disparate amongst patients. Thus, similarity metrics used in traditional K-means will not perform well. With $K = 10$ as optimal number of K , K determined by evaluating the BIC and AIC values for various K values, the generated clusters have been profiled on 50 patients with the eGFR as key marker. The profiling showed that the algorithm has well differentiated patients, with the first 3 clusters showing early signs of CKD (high eGFR), followed by clusters 4 to 9 with signs of CKD stage 3-4 (eGFR gradually deteriorating) and cluster 10 with CKD stage 5 (low eGFR).

Another study of similar degree has been conducted by Mohammed A. Al-Mamun et al., where a hybrid clustering approach was proposed to evaluate kidney disease progression from an acute kidney injury (AKI) [3]. The study used EHR as data to work with and focused on patients with ICD-9 code 585 and ICD-10 code N18 only. Also, the dataset was separated in three distinct cohorts as: hospital acquired AKI (HA-AKI), community acquired AKI (CA-AKI) and No-AKI. The objective of the study is to tackle the aspects overlooked by traditional clustering methods where different phenotypes are not well identified as part of disease trajectory. As a result, Mohammed A. Al-Mamun et al. suggests a hybrid framework with the joint use of variable clustering and network-based clustering. The variable clustering would uncover correlated clinical and demographic variables while network-based clustering would allow visual inspection and comorbidities grouping for each cohort. The variable clustering method has well clustered the clinical and demographic components while the network-based clustering method identified comorbidities unique to each cohort, e.g. HA-AKI cohort had hypertension, pneumonia and dehydration identified. In contrast the no-AKI cohort had IHD and breast cancer screening as key comorbidities.

Existing studies have demonstrated the need to augment or modify traditional clustering algorithms to achieve successful subgrouping of patients. Therefore, the following study provides a comparative analysis of traditional clustering algorithms' performances in identifying groups of patients with similar characteristics.

III. DATA

The dataset used in this study consists of electronic health records from a cohort of patients diagnosed with Chronic Kidney Disease. The dataset was extracted from a publicly available database (MIMIC IV database) that includes comprehensive clinical data, including demographics, laboratory test results, diagnosis codes, and

hospitalization details for a large population of patients treated at the Beth Israel Deaconess Medical Centre in Boston, MA [4]. The data is presented in an anonymised format where the patient's identification nor other personal information can be retrieved. The records within the dataset span from 2008 to 2022. It has been anonymised by shifting to dates into the future between 2100 and 2200. The shifting has been applied independently for each patient and is temporally not comparable. The EHR data was extracted and pre-processed to create a dataset of unique CKD patients per row, which can be used to model and compare the various models as part of this study. The detailed data pre-processing steps are depicted in Figure 1.

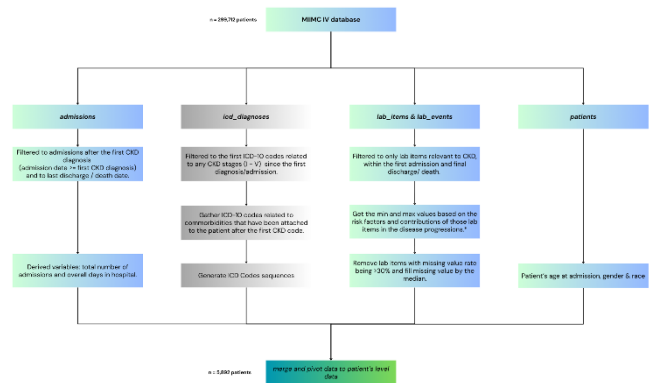


Fig. 1. Detailed data pre-processing steps

The MIMIC-IV database contains information of a 299,712 unique patients with each patient presenting differing characteristics. As the study solely focuses on patients presenting signs of CKD, the data has been meticulously limited to the cohort of interest by using ICD-10 codes respectively linked to any stages of CKD: *N181* (CKD, stage 1), *N182* (CKD, stage 2), *N183* (CKD, stage 3), *N184* (CKD, stage 4) and *N185* (CKD, stage 5). The ICD-10 coding system has been preferred to the ICD-9 system due to the granularity and precision of the former coding system. The data filtering allowed the creation of a cohort of 5,892 unique patients presenting signs of CKD. Based on the extracted cohort and to provide a representation of the disease progression, the comorbidities with relevance to CKD have been filtered, where only comorbidities detected after the first diagnosis of CKD have been considered. Comorbidities related to the reproductive systems or external organs (e.g. the skin) have been discarded. From the available variables in the dataset, 2 features were derived from: *Total number of hours in care* and *Total number of admissions*. The *Total number of hours in care* was calculated from the first date of admission since diagnosis of CKD till the last discharge date, with the assumptions that patients might undergo treatments outside of the hospital (e.g. medication at home). On the other hand, the *Total number of admissions* is the total number of admissions since the first diagnosis of CKD inclusive. These features provide temporal significance in the disease progression.

To summarise, the key variables extracted from the datasets available in MIMIC-IV are:

1. Laboratory Details - all severe values to provide a severe progression path.

2. Patients Race.
3. Patients Age at admission.
4. ICD 10 diagnosis code related to CKD
5. Total Number of hours in care.
6. Total number of admissions.
7. ICD 10 diagnosis code attached to the comorbidities.

It is important to note that the CKD patients present in the dataset do not necessarily have their entire medical records in the database. This study solely focuses on the admissions within a single medical centre. The laboratory details collected from the database range from the first laboratory test up to the last laboratory test within the overall admission period from the first diagnosis of CKD. Missing biomarker values have been imputed using the median to preserve the variance and original data structure. The biomarkers extracted from the database and key indicators of kidney functionalities (minimum or maximum concentration) are presented in Table 1 [5].

Key Biomarkers	Severity Indicator
Albumin	Low
Bicarbonate	Low
Calcium	Low
Creatinine (Blood)	High
Creatinine (Urine)	High
Hemoglobin	Low
Phosphate	High
Potassium	High
Urea Nitrogen	High

TABLE 1. KEY CHRONIC KIDNEY DISEASE BIOMARKERS

By applying the various clustering algorithms to this extensive dataset, we aim to uncover distinct subgroups of CKD patients that exhibit unique patterns in terms of their clinical characteristics, disease progression, and comorbidity profiles. An important task that had to be taken to have the EHR data into an appropriate format for clustering purposes is to have it as a patient level dataset. The sequential ICD 10 codes were then embedded using Word2Vec. The sequences of ICD 10 codes follow the chronological order of the diagnosis, from the first assigned code to the latest. As the succeeding step of the framework revolves around the sequential prediction of diagnosis code, the latest 3 codes have not been included in the embedding process, to prevent data leakage in the predictive modelling. Word2Vec is a popular method for converting words into numerical vector representations, which preserves the semantic relationships between the different ICD codes [6]. The algorithm is structured based on a Neural Network architecture, where each ICD code is treated as a word, and the model learns the vector representations of these codes based on their co-occurrence patterns in the sequence of diagnoses for each patient. Amongst existing embedding methods, such as GloVe (Global Vectors) or BERT (Bidirectional Encoder Representations from Transformers) which are better performing in capturing contextual information, we decided to use Word2Vec to

ensure that institutions with low computational power can reproduce this analysis and yet have a meaningful output.

IV. IMPLEMENTATION

The implementation of the various clustering models was performed using Python and libraries such as scikit learn for data science aspects (modelling & hyperparameters tuning), and Pandas and NumPy for data manipulation. Each of the models had their hyperparameters tuned to obtain the optimal number of clusters and parameters, where required.

The optimal k values of 4 for the k-means model were derived using the elbow method with a minimized WCSS and shown in Figure 2. The model's parameters were set with basic factors using 'kmeans ++' as algorithm and $k = 4$ with the maximum iteration set to 1000.

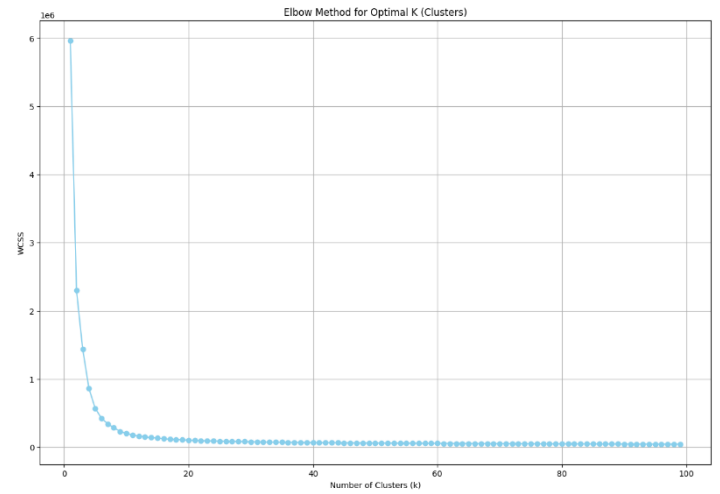


Fig. 2. Optimal k using Elbow method

To determine the appropriate parameters for the hierarchical clustering model, a dendrogram (Figure 3) was fitted on the dataset with common parameters: the Ward's method as the linkage and Euclidean as metric. Ward linkage was selected because it is more appropriate and efficient for segmenting patients with similar characteristics. Based on the produced dendrogram, the suggested optimal number of components is $k = 4$.

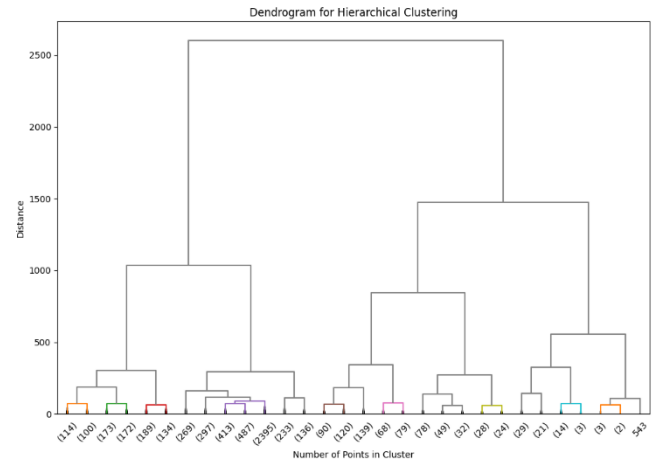


Fig. 3. Hierarchical Clustering Dendrogram

As with the HDBSCAN, the optimal number of clusters do not need to be specified in advance. Instead, the minimum size and minimum number of samples must be configured. To get the optimal values, a range between 2 and 101, has been used to loop through and evaluate against the silhouette score. Figure 4 shows the evaluation of the cluster size against the silhouette score. The optimal values which produced the highest silhouette score, were determined as *minimum number of size = 4* and *minimum number of samples = 1*. Using these parameters, the HDBSCAN model formed 37 clusters of various sizes.

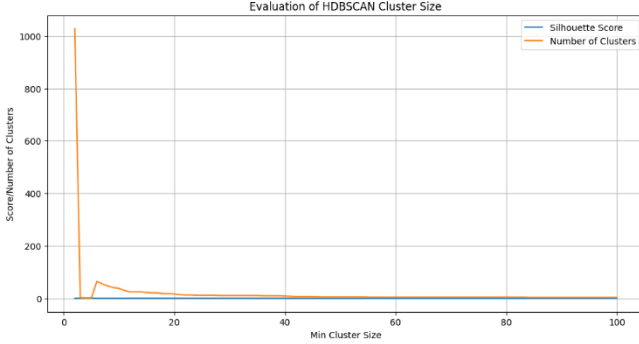


Fig. 4. HDBSCAN – Silhouette score per various minimum size and cluster

Based on the output of the various models, GMM's hyperparameters were tuned within specific ranges. To effectively decide on the optimal values, 3 elements were measured against the silhouette score: *covariance type* (full, tied, diagonal or spherical), *number of components* (between 2 and 36) and the *initiation method* (*k-means++* or *random*). Whilst the silhouette score was the preferred metric in assessing the quality of the formed clusters and to tune the hyperparameters of the previous models, the BIC was also considered as a metric to assess the hyperparameters of the GMM model. The Bayesian Information Criterion, also known as BIC, is a model selection criterion specific to the GMM. The BIC helps in deciding the optimal number of components that best suit the data without causing overfitting [7]. Figures 5 and 6 show the variation in the BIC and Silhouette score by number of components and covariance type. The final optimal values set for the GMM model are *covariance type = tied*, *initiation = 'k-means++'* and *number of components = 5*.

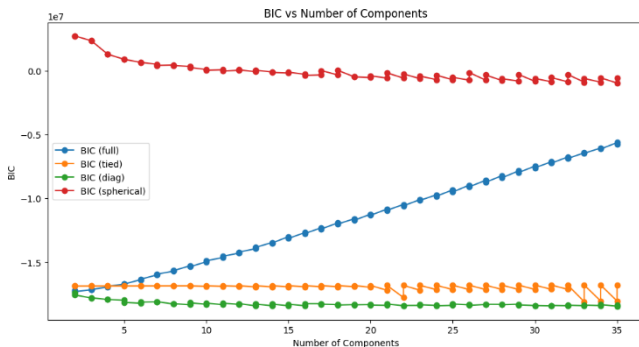


Fig. 5. BIC evaluation per covariance

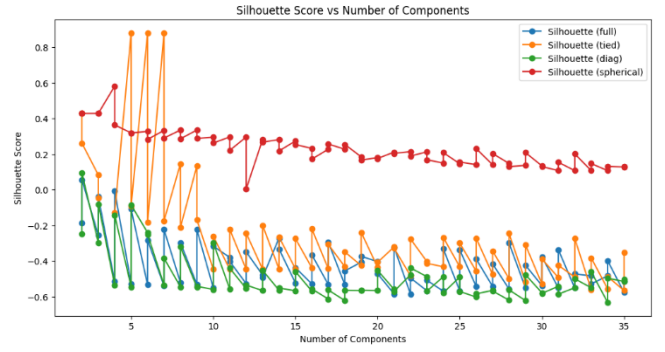


Fig. 6. Silhouette Score evaluation per covariance

V. RESULTS

The 4 models were evaluated using various metrics and visual inspection techniques: the silhouette score, Calinski-Harabasz Index and Davies-Bouldin Index. The silhouette score measures the quality of a cluster by considering the mean intra-cluster distance and the mean nearest-cluster distance for each sample. The score ranges between -1 and 1, where the higher the silhouette score, the better the distinction between clusters [8]. Similarly, the Calinski-Harabasz Index measures the quality of the formed clusters by assessing the ratio of between-cluster distribution to within-cluster distribution. While the index score does not have any predefined ranges, a higher value indicates a distinct and well-separated clustering [9]. On the other hand, the Davies-Bouldin Index measures the average similarity between each formed cluster, by calculating the ratio of within-cluster distance to between-cluster distance. The index score ranges between 0 and ∞ , with the lower values indicating better clustering. Furthermore, a score closer to 0 indicates minimal overlap between clusters [10]. The clusters will also be visually inspected using Principal Component Analysis (PCA) components to ensure they capture meaningful patterns in the data.

The silhouette score was chosen as one of the primary metrics for assessing the quality of the clusters, as it is well-suited for unsupervised problems. PCA was also used as a technique to reduce the dimensionality of the data and visually inspect the structure of the clusters. Table 2 presents the final silhouette scores, Davies-Bouldin Index scores and Calinski-Harabasz Index scores for each clustering method investigated in this paper.

Metrics	Clustering methods			
	<i>K-Means</i>	<i>Hierarchical Clustering</i>	<i>HDBSCAN</i>	<i>GMM</i>
Silhouette Score	0.672	0.672	-0.025	0.0170
Calinski-Harabasz Index	11612.960	10345.498	118.423	8.051
Davies-Bouldin Index	0.521	0.497	1.977	2.410

TABLE 2. METRICS SCORES PER CLUSTERING METHODS

From the presented scores, it is evident that simple clustering techniques such as K-Means clustering, and Hierarchical

clustering outperform more complex methods like GMM and HDBSCAN for this dataset. It is apparent from the metrics that density-based algorithms (HDBSCAN) and probabilistic models (GMM) are not well suited for the given dataset and problem domain. Based on the Calinski-Harabasz Index score and the Davies-Bouldin Index score, the worse performing model is the GMM followed by the HDBSCAN, with the scores indicating a poor separation of the clusters and high similarity in the cluster's formations. This might also suggest a possible re-tuning of the GMM's hyperparameters by focusing on further optimization of the BIC values and adjusting to a different covariance type. The PCA analysis, presented in Figure 7, reinforces the results presented by the scores of each model. Upon visual inspection of the PCAs, HDBSCAN and GMM present a poor separation of the clusters in comparison to K-Means and Hierarchical clustering. GMM presents an overlap of data points between clusters 1 and 4, reflecting the low Calinski-Harabasz score.

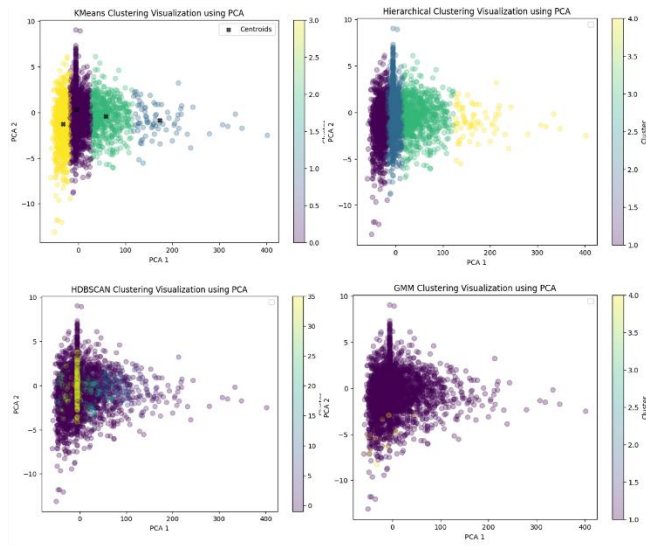


Fig. 7. PCA - Visualisation of the clusters per model

Table 3 shows the cluster formation, and the number of data points assigned to each cluster. HDBSCAN's output has not been included for comparative purposes due to the high number of components that have been derived.

Cluster Assignment			
Cluster	K-Means	Hierarchical Clustering	Gaussian Mixture Model
1	4421	882	5853
2	80	4230	0
3	521	707	0
4	870	73	39

TABLE 3. DATA POINTS ASSIGNMENTS PER CLUSTER

The data assignment for cluster 3 are of similar magnitude for the K-Means and Hierarchical models. Nevertheless, a sparse difference is noticeable amongst the other clusters for each

model. The differences can possibly result from the cluster shape assumptions or the sensitivity to initializations. Although, the optimal components for GMM were set to 5 based on the silhouette score tuning, only 2 effective clusters have been created. This might have resulted due to the overlapping of components in the feature space and by assigning a hard clustering instead of soft clustering. A sole focus on the BIC for component selection would be more appropriate.

An interesting variable to focus on is the creatinine range of each cluster for the K-Means and Hierarchical clustering models. The creatinine value is a critical element in determining the CKD stage associated with the patient's condition. The analysis can be comparable to the ones conducted by D.T. Anh Luong et al. as part of their study [2]. Figures 8 and 9 present the creatinine level variation of each cluster per method.

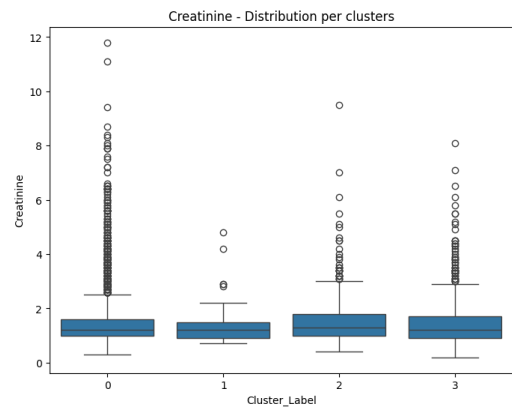


Fig. 8. Creatinine level per clusters – K-Means

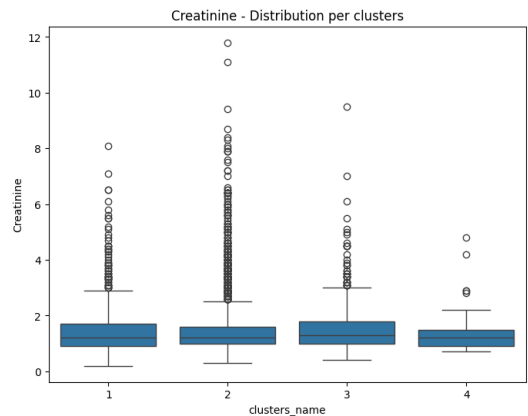


Fig. 9. Creatinine level per clusters – Hierarchical

It is visible that certain clusters (1 and 4) have a lower creatinine range compared to the other ones. Clusters 1 of the Hierarchical model, and cluster 0 of the K-Means model imply grouping patients with lower creatinine value, therefore patients at an early stage of the disease (Stage 1). On the other hand, clusters 4 of the Hierarchical model, and cluster 1 of the K-means model have higher low creatinine value, which indicates a possibility of clustering patients at a more advanced stage of the disease (Stage 4 – 5). It also

comes to attention that cluster 0 of the K-mean model and cluster 2 of the Hierarchical model contain a high number of outliers, explaining the high assignments of patients within those clusters. With a similar interest, urea nitrogen is another indicator of CKD severity. Figures 10 and 11 depict the urea nitrogen level variation per cluster for each clustering method, which present similar patterns to the creatinine ranges per cluster.

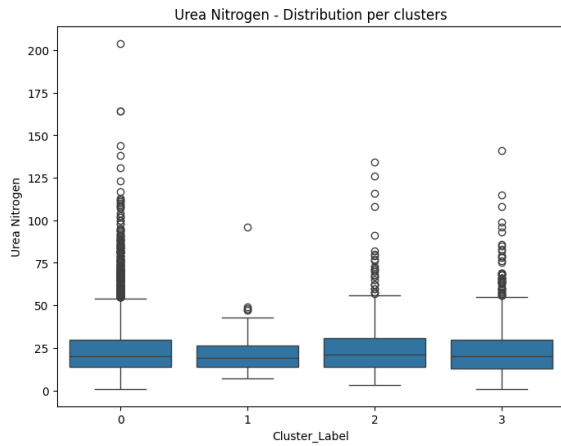


Fig. 10. Urea nitrogen level per clusters – K-Means

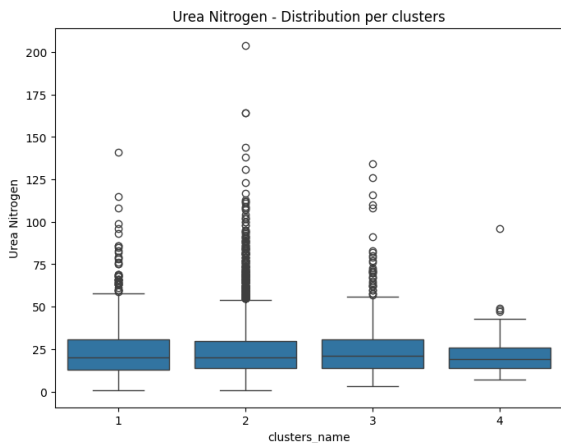


Fig. 11. Urea nitrogen level per clusters – Hierarchical

VI. CONCLUSION

This study provides a comparative analysis of the various clustering techniques on EHR data, with emphasis on the superiority of traditional clustering algorithms (K-Means & Hierarchical Clustering) over complex algorithms (GMM & HDBSCAN).

A potential expansion of the study will focus on the data pre-processing framework where other embedding methods can be investigated to scrutinize its impact on the performance of the clustering algorithms. As an example, Bio_Clinical BERT, which is a modified version of the

BERT algorithm trained on EHR and ICD code data, will be of interest [11]. As previously mentioned, this comparative study is part of extensive research surrounding the simultaneous application of machine learning and process mining to predict severe disease progressions for Chronic Kidney Disease patients as well as any other chronic illnesses [12]. The output from the well-performing clustering model will be utilized as a feature and fed into the succeeding process model and predictive model.

REFERENCES

- [1] Francis, A., Harhay, M.N., Ong, A.C.M. *et al.* (2024), Chronic kidney disease and the global public health agenda: an international consensus. *Nat Rev Nephrol* **20**, 473–485, doi: <https://doi.org/10.1038/s41581-024-00820-6>
- [2] D. T. Anh Luong and V. Chandola (2017), "A K-Means Approach to Clustering Disease Progressions," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, USA, pp. 268-274, doi: 10.1109/ICHI.2017.18.
- [3] Mohammad.A. Al-Mamun, Ki Jin Jeun, Todd Brothers, Ernest Asare, Khaled Shawwa, Imtiaz Ahmed (2024). Evaluating the kidney disease progression using a comprehensive patient profiling algorithm: A hybrid clustering approach, medRxiv 2024.09.08.24313275, doi : <https://doi.org/10.1101/2024.09.08.24313275>.
- [4] Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., & Naumann, T. (2020, April). MIMIC-Extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM conference on health, inference, and learning* (pp. 222-235).
- [5] Lopez-Giacoman, S., & Madero, M. (2015). Biomarkers in chronic kidney disease, from kidney function to kidney damage. *World journal of nephrology*, 4(1), 57.
- [6] Steiger, E., & Kroll, L. E. (2023). Patient embeddings from diagnosis codes for health care prediction tasks: Pat2Vec machine learning framework. *JMIR AI*, 2, e40755.
- [7] H. Wan, H. Wang, B. Scotney and J. Liu, "A Novel Gaussian Mixture Model for Classification," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 3298-3303, doi: 10.1109/SMC.2019.8914215.
- [8] Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)* (pp. 747-748). IEEE.
- [9] Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta. *Journal of Applied Informatics and Computing*, 7(1), 95-103.
- [10] Jumadi Dehotman Sitompul, B., Salim Sitompul, O., & Sihombing, P. (2019, June). Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. In *Journal of Physics: Conference Series* (Vol. 1235, No. 1, p. 012015). IOP Publishing.
- [11] Ling, Y. (2023). Bio+ Clinical BERT, BERT Base, and CNN performance comparison for predicting drug-review satisfaction. *arXiv preprint arXiv:2308.03782*.
- [12] Kandasamy, N., Chaussalet, T., & Basukoski, A. (2023). A Conceptual Framework to Predict Disease Progressions in Patients with Chronic Kidney Disease, Using Machine Learning and Process Mining. In *Healthcare Transformation with Informatics and Artificial Intelligence* (pp. 190-193). IOS Press.