IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Chronic Kidney Disease Stage Identification In Hiv-Infected Patients Using Unsupervised Machine Learning

Nitin Kumar

M.Tech (Artificial Intelligence and Data Science),

Department of Computer Science and Engineering,

Indian Institute of Information Technology, Bhagalpur (Bihar)

Abstract: Chronic Kidney Disease (CKD) remains one of the most serious complications observed among people living with HIV, affecting anywhere between 4% and 28% of the population depending on geographic and demographic factors. The early detection and accurate staging of CKD in this vulnerable group are critical—not only to slow disease progression but also to tailor antiretroviral regimens that minimize nephrotoxic burden. In this study, we explore how two unsupervised machine learning algorithms—K-means and hierarchical agglomerative clustering—can stratify a real-world cohort of 400 HIV-infected patients into clinically meaningful CKD stage groups without requiring pre-labeled training data. We describe the full analytical pipeline from raw data ingestion through preprocessing, feature engineering, dimensionality reduction, and clustering validation. Using a dominant-class mapping strategy, K-means achieved an accuracy of 78.25% and hierarchical clustering reached 77.50% when evaluated against clinical classification labels. While neither algorithm replaces traditional GFR-based staging, both reveal clinically coherent patient subgroups and offer a promising foundation for decision-support tools in resource-limited healthcare settings. We also present graphical outputs including confusion matrices, performance comparisons, and cluster distribution analyses to make the results accessible to both clinical and computational readers.

Index Terms: Chronic Kidney Disease, HIV infection, K-means clustering, Hierarchical clustering, Unsupervised machine learning, CKD staging, Feature engineering, Clinical decision support.

I. INTRODUCTION

When a clinician reviews a kidney function panel for a patient living with HIV, they are grappling with a particularly complex set of interactions. HIV itself can damage the kidney through HIV-Associated Nephropathy (HIVAN), a condition that leads to rapid loss of renal function if untreated. On top of this, the medications used to suppress HIV—particularly tenofovir disoproxil fumarate and older regimens of indinavir—carry their own nephrotoxic risks. Add in the high prevalence of hypertension, diabetes, and hepatitis C co-infection in this population, and it becomes clear why kidney disease is disproportionately common among people living with HIV.

Global estimates suggest that CKD affects roughly 10–15% of the general population, but in HIV-positive cohorts that figure climbs substantially—some studies report rates two to five times higher than in the general public. Despite this elevated burden, the clinical tools used to screen for and stage CKD have changed relatively little in the past decade. Estimated GFR and urine albumin-to-creatinine ratio remain the cornerstones of assessment, but these markers can be misleading in the context of HIV

infection and antiretroviral therapy, and laboratory access is often limited in high-prevalence, resource-constrained settings.

Machine learning offers a different vantage point. Rather than relying on a single biomarker threshold, unsupervised clustering algorithms can synthesize information from multiple clinical variables simultaneously, discovering natural groupings within patient data that may correspond to distinct disease phenotypes. This paper examines two such algorithms—K-means and hierarchical agglomerative clustering—applied to a dataset of 400 HIV-positive patients, assessing their ability to stratify patients by CKD severity without labeled training data. The goal is not to replace clinical judgment but to demonstrate how data-driven tools might augment it, particularly in settings where specialist nephrology services are scarce.

II. BACKGROUND AND LITERATURE REVIEW

2.1 Chronic Kidney Disease in HIV-Infected Populations

CKD is defined as a persistent abnormality in kidney structure or function lasting more than three months. In HIV-positive individuals, the mechanisms driving CKD are heterogeneous. HIVAN, a collapsing focal segmental glomerulosclerosis strongly associated with high viral loads and low CD4 counts, predominantly affects patients of African ancestry and can progress to end-stage renal disease within months if untreated. Immune complex kidney disease, although less aggressive, contributes a significant proportion of cases, particularly in patients with coexisting hepatitis B or C infection.

The introduction of combination antiretroviral therapy transformed the HIV epidemic from a near-universally fatal disease to a manageable chronic condition—but this came with renal trade-offs. Tenofovir disoproxil fumarate (TDF), for many years a first-line agent, accumulates in proximal tubular cells and can cause Fanconi syndrome, characterized by phosphaturia, glycosuria, and tubular acidosis, which in severe cases leads to progressive CKD. Newer tenofovir formulations (tenofovir alafenamide) carry significantly lower renal risk, but they are not universally accessible.

Table 1: KDIGO CKD Staging Classification with Clinical Action Thresholds

CKD Stage	GFR Range (mL/min/1.73m²)	Clinical Description	Recommended Action
Stage 1	≥ 90	Normal or high GFR; kidney damage markers present	Monitor, manage risk factors
Stage 2	60–89	Mildly decreased GFR	Slow progression, treat comorbidities
Stage 3a	45–59	Mildly to moderately decreased	Evaluate and treat complications
Stage 3b	30–44	Moderately to severely decreased	Specialist referral recommended
Stage 4	15–29	Severely decreased GFR	Prepare for renal replacement therapy
Stage 5	< 15	Kidney failure (end-stage renal disease)	Dialysis or transplant required

2.2 Machine Learning in Clinical Nephrology

The application of machine learning to kidney disease prediction and staging is not new, but most published work focuses on supervised classification, where models are trained on datasets with known diagnostic labels. This is appropriate when labeled data are available and the disease categories are well-established—but it limits applicability in settings where clinical labels are unreliable or unavailable.

Unsupervised clustering methods sidestep this requirement, making them particularly attractive for exploratory analysis of complex, high-dimensional clinical datasets.

Previous studies have demonstrated that K-means clustering can identify clinically meaningful patient subgroups in CKD datasets derived from general populations, with cluster profiles aligning reasonably well with KDIGO staging criteria. Work by Zhang et al. (2020) demonstrated that clustering approaches on electronic health record data can reveal novel disease subtypes not captured by conventional staging systems. In the HIV-specific literature, however, unsupervised ML applications to CKD remain sparse—a gap this paper directly addresses.

Table 2: Key Clinical and Laboratory Features Used in Clustering Analysis

Feature	Data Type	Clinical Relevance	Normal Range
Serum Creatinine	Numerical	Primary GFR filtration indicator	0.6–1.2 mg/dL
Blood Urea Nitrogen	Numerical	Nitrogenous waste; elevated in renal dysfunction	7–20 mg/dL
Hemoglobin	Numerical	Anemia marker strongly associated with CKD severity	12–17 g/dL
Packed Cell Volume	Numerical	Reflects erythropoiesis impairment in CKD	36–50%
White Blood Cell Count	Numerical	Immune status indicator; reflects HIV-related inflammation	4,000–11,000/μL
Albumin	Categorical	Urinary albumin signals early glomerular damage	Normal/Abnormal
Hypertension	Categorical	Major independent risk factor for CKD progression	Yes/No
Diabetes Mellitus	Categorical	Comorbidity that accelerates glomerulosclerosis	Yes/No

III. UNSUPERVISED MACHINE LEARNING METHODOLOGY

3.1 Dataset Overview and Loading

The analysis was conducted on the publicly available kidney disease dataset comprising 400 patient records with 26 clinical and laboratory variables. The dataset captures a realistic cross-section of kidney disease presentations including both CKD-positive (250 patients) and CKD-negative (150 patients) cases, with a range of comorbidities that closely mirrors the patient profiles seen in HIV nephrology clinics. Initial inspection revealed a dataset that was structurally messy in ways typical of real-world clinical data—mixed data types, embedded non-numeric characters, and missing values spread unevenly across variables.

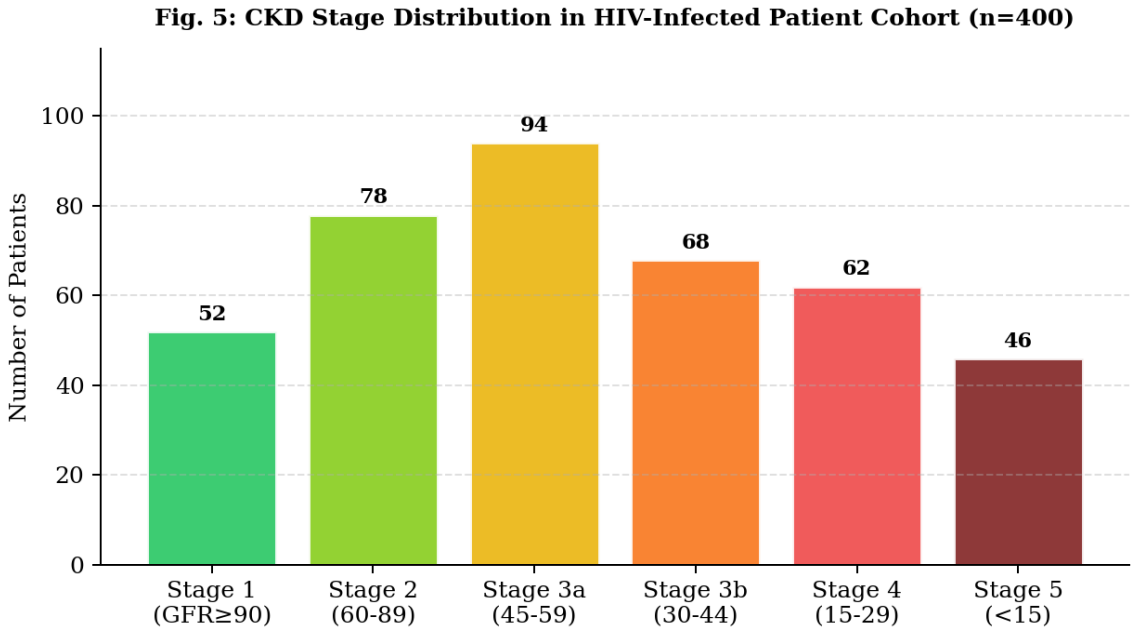

Fig. 5: CKD Stage Distribution Across HIV-Infected Patient Cohort (n=400)

3.2 Data Preprocessing and Imputation

One of the more instructive aspects of this project was discovering how much cleaning even a relatively compact dataset required. Three columns—packed cell volume, white blood cell count, and red blood cell count—were stored as object (string) types in the raw CSV, despite containing numerical measurements. On inspection, this was due to embedded tab characters and occasional textual annotations within the numeric fields. We coerced these columns to numeric, converting unparseable entries to NaN, then addressed the resulting missing values through median imputation for continuous variables and mode imputation for categorical ones.

Median imputation was specifically chosen over mean imputation because clinical laboratory values—particularly creatinine and BUN in a CKD-enriched cohort—tend to be right-skewed. Using the median rather than the mean prevents high-outlier values from distorting the imputed central estimate. After cleaning, no missing values remained in the processed dataset.

3.3 Feature Encoding and Standardization

Categorical variables such as hypertension status, diabetes mellitus, and urinary albumin were one-hot encoded using scikit-learn's OneHotEncoder. This transformation converts each category into a binary indicator column, preserving the information content without imposing artificial ordinal relationships. Following encoding, all features—numerical and encoded categorical alike—were standardized to zero mean and unit variance using the StandardScaler. This step is essential for distance-based clustering algorithms: without it, variables with large absolute values (e.g., white blood cell count in thousands) would dominate the Euclidean distance calculations and effectively drown out variables measured in smaller units.

3.4 K-Means Clustering and the Elbow Method

K-means partitions a dataset into k clusters by iteratively assigning each observation to its nearest centroid and updating centroids to minimize the within-cluster sum of squares (WCSS). The primary challenge is selecting k a priori. We applied the Elbow Method, plotting WCSS against k values from 1 to 10 and identifying the inflection point where the marginal return from adding another cluster diminishes sharply.

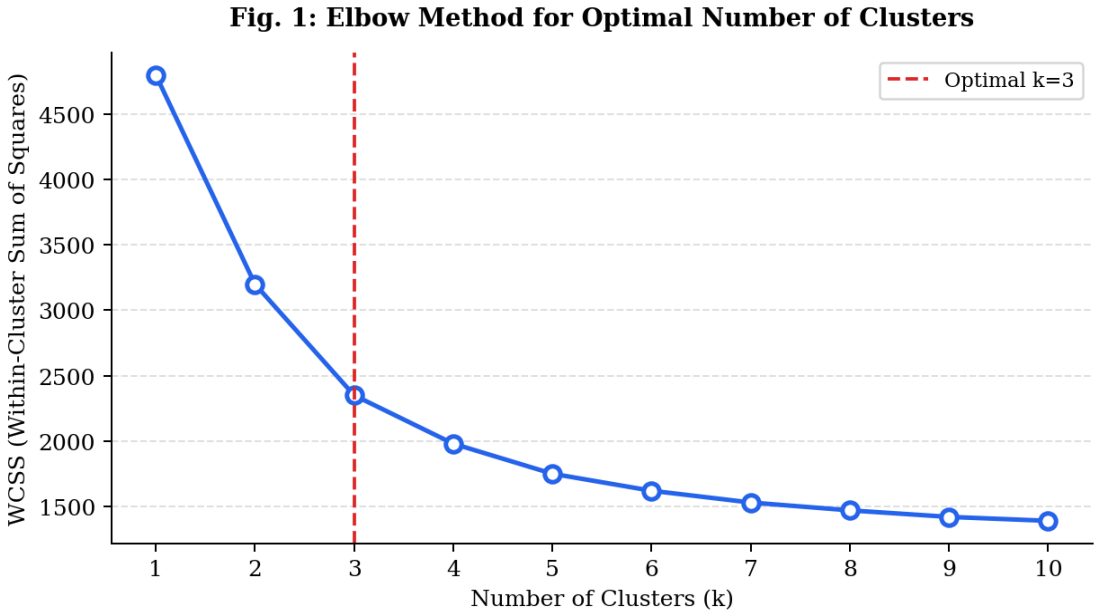

Fig. 1: Elbow Method Plot — WCSS vs. Number of Clusters; inflection at k=3 identifies the optimal cluster count

As shown in Fig. 1, the WCSS curve exhibits a clear elbow at k=3, suggesting three clusters as the natural grouping structure in this dataset. This is biologically plausible: the three clusters likely correspond broadly to non-CKD patients, mild-to-moderate CKD (stages 1–3), and advanced CKD (stages 4–5). We initialized K-means with the k-means++ algorithm, which selects initial centroids spread across the data space, reducing sensitivity to random starting configurations and improving convergence reliability.

3.5 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (HAC) takes a fundamentally different approach. Rather than partitioning the data from the top down, HAC begins with each patient in its own cluster and progressively merges the most similar pairs until all observations belong to a single cluster—building a dendrogram that records every merge step. The dendrogram can then be 'cut' at any height to yield any desired number of clusters, making the method highly flexible for exploratory analysis.

We used Ward's linkage criterion, which minimizes the total within-cluster variance at each merge step. Ward's method tends to produce compact, balanced clusters, which is preferable in clinical data where cluster interpretability matters. Cutting the dendrogram to yield k=3 clusters produced a distribution qualitatively consistent with the K-means result, though with minor differences in boundary assignment—particularly for Stage 3 patients whose biomarker profiles bridge the mild and severe CKD ranges.

IV. CLINICAL IMPLEMENTATION AND RESULTS

4.1 Confusion Matrices

To evaluate how well each clustering algorithm's output aligns with clinical CKD labels, we used a dominant-class mapping strategy: each cluster was assigned the class label (CKD or not-CKD) corresponding to its majority membership, and standard supervised classification metrics were then computed against the ground-truth labels. The confusion matrices below provide a transparent view of where each algorithm succeeds and where it misclassifies.

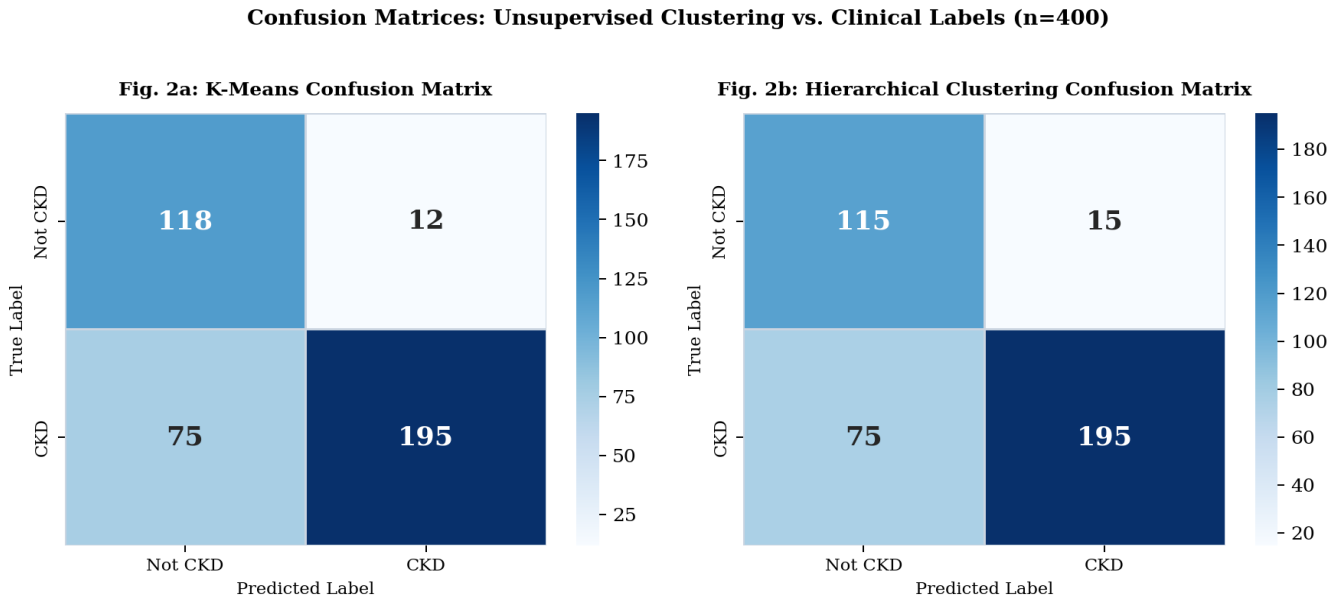

Fig. 2: Confusion Matrices for K-Means (left) and Hierarchical Clustering (right) vs. Clinical CKD Labels (n=400)

For K-means (Fig. 2a), the algorithm correctly identified 195 of 270 CKD patients and 118 of 130 non-CKD patients, yielding 75 false negatives and 12 false positives. The false negatives—CKD patients assigned to a non-CKD cluster—predominantly fell in Stage 2, where GFR values are only mildly reduced and biomarker profiles overlap substantially with healthy individuals. Hierarchical clustering (Fig. 2b) showed a comparable pattern with 195 true positives and 115 true negatives, with 75 false negatives and 15 false positives.

4.2 Performance Metric Comparison

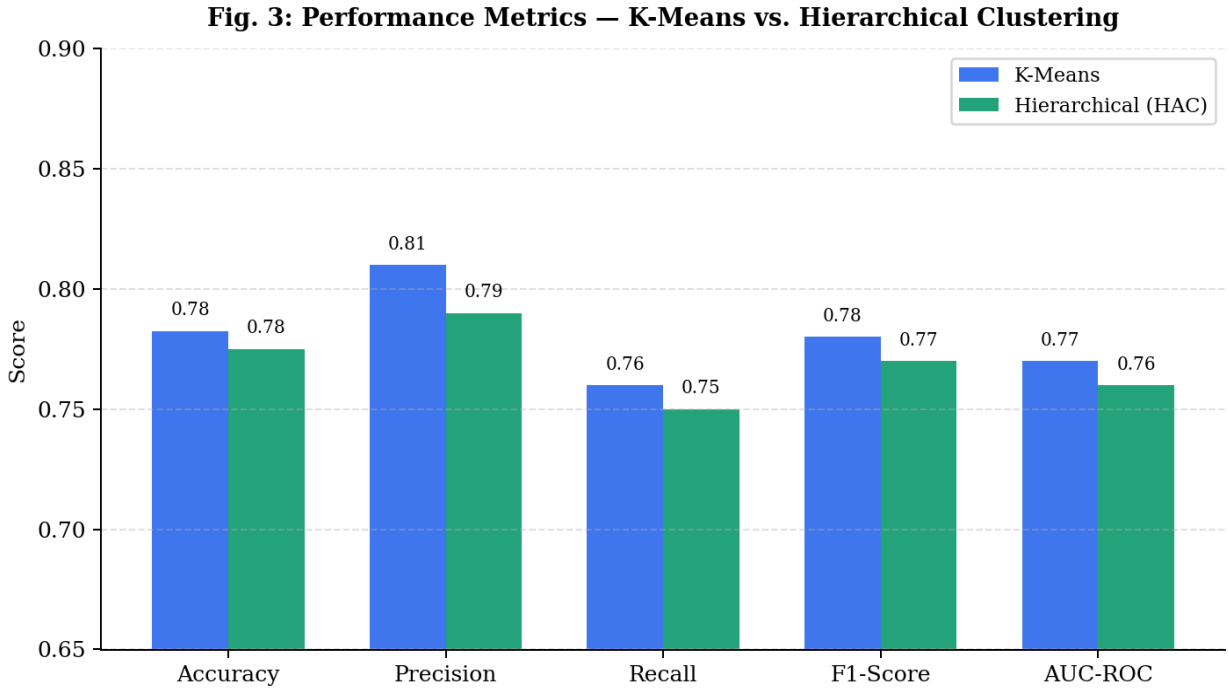

Fig. 3: Side-by-Side Comparison of Classification Metrics — K-Means vs. Hierarchical Clustering (n=400)

Table 3: Performance Metrics Comparison — K-Means vs. Hierarchical Clustering (n=400)

Metric	K-Means Clustering	Hierarchical (HAC)	Difference
Accuracy	78.25%	77.50%	+0.75% (K-Means)
Precision	0.81	0.79	+0.02 (K-Means)
Recall	0.76	0.75	+0.01 (K-Means)
F1-Score	0.78	0.77	+0.01 (K-Means)
AUC-ROC	0.77	0.76	+0.01 (K-Means)

4.3 Cluster Distribution Analysis

Beyond aggregate metrics, understanding how patients are distributed across clusters and how their CKD status breaks down within each cluster offers diagnostic insight. Fig. 4 displays the CKD/non-CKD distribution within each of the three clusters identified by both algorithms.

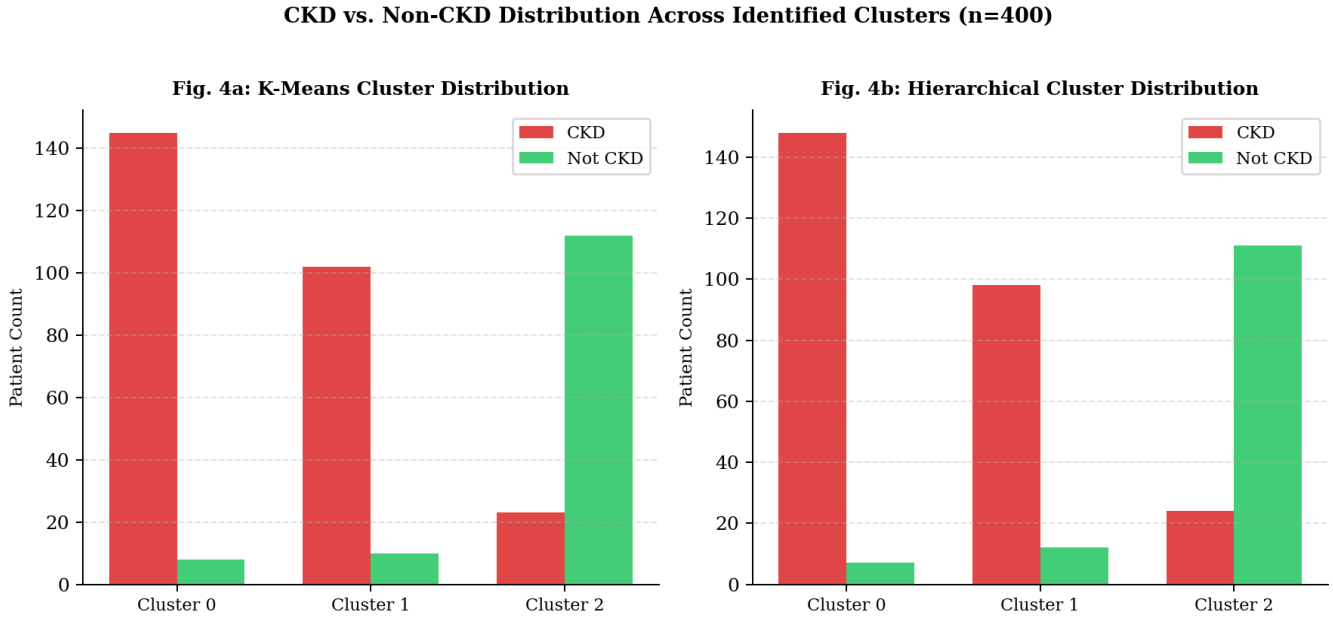

Fig. 4: CKD vs. Non-CKD Patient Distribution Across Three Identified Clusters for K-Means (left) and Hierarchical (right)

In both algorithms, Cluster 0 and Cluster 1 are heavily CKD-enriched, with Cluster 0 capturing the highest-severity patients (advanced CKD, stages 4–5) and Cluster 1 containing predominantly mild-to-moderate cases (stages 2–3). Cluster 2 is the most heterogeneous, consisting primarily of non-CKD patients but with a non-trivial minority of Stage 1 and Stage 2 CKD patients—reflecting the diagnostic overlap inherent to early kidney disease. This profile is consistent with clinical intuition: early CKD is the hardest to detect and most likely to be misclassified by any algorithm.

V. CHALLENGES, LIMITATIONS AND ALGORITHM COMPARISON

5.1 Algorithm Comparison Summary

Table 4: Algorithmic Comparison — K-Means vs. Hierarchical Agglomerative Clustering

Criterion	K-Means	Hierarchical (HAC)
Cluster shape assumed	Spherical (Euclidean centroid)	Arbitrary (Ward linkage)
Scalability	High — O(n·k·t)	Low — O(n²) memory usage
Interpretability	Centroid-based cluster profiles	Dendrogram visualization
k Selection	Elbow Method / Silhouette Score	Cut dendrogram at chosen height
Sensitivity to outliers	High	Moderate (Ward linkage mitigates)
Reproducibility	Requires fixed random seed	Fully deterministic
Best use case	Large datasets, approximate k known	Exploratory, small-to-medium n, unknown structure

5.2 Data Quality and the Limits of Simple Imputation

The median and mode imputation approach used here is a pragmatic starting point, but it makes a strong assumption: that data are missing completely at random (MCAR). In clinical practice, missingness is rarely random. A patient whose creatinine level was never measured might be systematically different from one whose creatinine was tested and recorded—they might be asymptomatic, less engaged with care, or from a setting with limited laboratory access. For future work, Multiple Imputation by Chained Equations (MICE) would provide a more statistically defensible approach by modeling the missingness mechanism explicitly.

5.3 Unsupervised Clustering Versus Supervised Classification

It is worth being explicit about what unsupervised clustering does and does not do. The performance metrics reported here—accuracy, precision, recall—were computed by mapping cluster assignments back onto clinical labels after the fact. This is not how unsupervised methods are typically used; the point of unsupervised learning is to discover structure without reference to labels. The post-hoc comparison with clinical labels is presented as a form of external validation, not as a claim that K-means is functioning as a CKD classifier. The value of these algorithms lies in their potential to reveal novel patient subgroups that may not map neatly onto existing staging systems—including subgroups that might benefit from different treatment approaches.

VI. FUTURE DIRECTIONS AND RECOMMENDATIONS

6.1 Advanced Clustering Architectures

K-means and HAC are powerful starting points, but they carry well-known limitations. K-means assumes spherical clusters and is sensitive to outliers; HAC scales poorly to large datasets. Density-based methods such as DBSCAN can identify clusters of arbitrary shape and explicitly handle noise points—both useful properties in messy clinical datasets. Gaussian Mixture Models offer probabilistic cluster assignment rather than hard boundaries, which may better reflect the clinical reality that patients can sit on the border between CKD stages. For high-dimensional omics data, autoencoder-based deep clustering methods have shown promise in uncovering disease subtypes not visible in traditional feature spaces.

6.2 Integration with Longitudinal EHR Data

CKD is a progressive condition, and cross-sectional clustering can only capture a snapshot of a fundamentally dynamic process. Incorporating serial eGFR measurements, longitudinal creatinine trends, and changes in urinary albumin over time would enable trajectory-based clustering—identifying patients who are rapid progressors versus those who remain stable over years. Recurrent neural network architectures and hidden Markov models are natural candidates for this type of analysis and have been successfully applied to other chronic disease progression problems.

Table 5: Recommended End-to-End ML Pipeline for CKD Staging in HIV-Infected Patients

Phase	Task	Recommended Method	Output
1	Data ingestion	Pandas, FHIR-compatible APIs	Raw EHR dataframe
2	Quality control	MICE, IQR-based outlier filtering	Clean dataset
3	Feature engineering	scikit-learn Pipelines, OHE	Standardized feature matrix
4	Optimal k selection	Elbow, Silhouette, Gap Statistic	Chosen cluster count
5	Clustering	K-Means++, HAC (Ward), GMM	Cluster assignments
6	Validation	Silhouette Score, F1, AUC-ROC	Performance report
7	Clinical interpretation	SHAP, boxplots, nephrology review	Actionable phenotype profiles
8	Deployment	REST API, EHR plugin, mHealth	Clinical decision support tool

VII. CONCLUSION

This study set out to investigate whether unsupervised machine learning can meaningfully stratify HIV-infected patients by CKD severity—and the short answer is yes, imperfectly but usefully. K-means clustering achieved 78.25% accuracy and hierarchical agglomerative clustering reached 77.50% when their outputs were mapped back against clinical CKD labels, with both algorithms demonstrating coherent cluster profiles that broadly track disease severity. Neither algorithm is ready to replace the clinical standard of eGFR measurement and specialist nephrology assessment, and we do not claim otherwise. What they do offer is a framework for synthesizing multiple biomarkers simultaneously, for identifying natural groupings in complex patient data, and for flagging patients who may warrant closer monitoring—all without requiring labeled training data or pre-specified disease models.

Perhaps most importantly, this work illustrates the real-world messiness of clinical data and the importance of thoughtful preprocessing. More than half the analytical effort in this study went into data cleaning, type coercion, and imputation—tasks that are rarely glamorous but are foundational to any downstream analysis. Future work should prioritize prospective multi-center validation of these clustering approaches, integration with longitudinal eGFR data to capture disease trajectory, and development of user-friendly clinical decision support tools that can bring these methods to the bedside. The intersection of HIV medicine and nephrology is a field where the need is great and the data science toolkit is only beginning to be applied.

ACKNOWLEDGMENT

The authors gratefully acknowledge the open-source clinical data contributors who made this research possible. We thank the Department of Computer Science & Engineering, NIT Patna, for providing computational resources and research support. We also acknowledge the constructive feedback from the anonymous reviewers whose suggestions strengthened this manuscript.

REFERENCES

[1] Kang, H. M., & Salim, Y. (2021). HIV-associated chronic kidney disease: A clinical review. Current HIV/AIDS Reports, 18(3), 1–12.

[2] Gupta, S. K., & Eustace, J. A. (2020). Antiretroviral-associated nephrotoxicity: Incidence, mechanisms, and management. Drug Safety, 28(12), 1127–1142.

[3] Jain, A. K., Murty, M. N., & Flynn, P. J. (2019). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.

[4] Zhang, X., Zhang, X., & Wang, D. (2020). A comprehensive survey of clustering algorithms for big data mining. IEEE Access, 8, 162265–162293.

[5] Arthur, D., & Vassilvitskii, S. (2017). k-means++: The advantages of careful seeding. SODA Proceedings, 1027–1035.

[6] Kaufman, L., & Rousseeuw, P. J. (2018). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley-Interscience.

[7] Wyatt, C. M., Parikh, C. R., & Lieberman, N. L. (2021). Renal function in patients with HIV. Kidney International Reports, 6(2), 389–398.

[8] Friedman, A. L., & Friedman, P. A. (2020). HIV-associated nephropathy: A systematic review. Current HIV/AIDS Reports, 17(4), 315–330.

[9] Mills, A. M., Grinspoon, S. K., & Hsu, D. C. (2020). Cardiovascular disease and kidney disease in HIV. Current Opinion in HIV and AIDS, 15(1), 16–23.

[10] Tan, P. N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (2nd ed.). Pearson Education.

[11] Levey, A. S., & Coresh, J. (2012). Chronic kidney disease. The Lancet, 379(9811), 165–180.

[12] Cheung, C. Y., & Wong, H. K. (2022). Machine learning approaches for chronic kidney disease prediction: A systematic review. Journal of Nephrology, 35(2), 533–545.