International Journal Software Engineering and Computer Science (IJSECS)
6 (1), 2026, 301-312
Published Online April 2026 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i1.6894.

RESEARCH ARTICLE Open Access

Analysis and Implementation of a Hybrid Case-Based Reasoning and K-Nearest Neighbor Approach for Chronic Kidney Disease Prediction

Hananing Sumaningdiah Larasati [1*], **Shella Sukma Dewi Waramena** [2], **Wulan Pahira** [3]

[1*,2,3] Department of Information System, Universitas Pamulang, South Tangerang City, Banten Province, Indonesia.

**Corresponding author*: dosen02819@unpam.ac.id.

Received: April 1, 2026; Accepted: April 17, 2026; Published: April 20, 2026.

Abstract: Chronic Kidney Disease (CKD) is a progressive deterioration of kidney function that frequently goes undetected in its early stages, posing a growing clinical concern — particularly among productive-age individuals whose diagnosis is often delayed until irreversible damage has occurred. Early and accurate prediction remains a pressing challenge, especially given the rising CKD incidence in this demographic linked to hypertension, diabetes, and shifting lifestyle patterns. This study developed a hybrid method combining Case-Based Reasoning (CBR) with weighted similarity and K-Nearest Neighbor (KNN) to improve prediction accuracy while preserving model interpretability. The dataset was obtained from the UCI Machine Learning Repository and filtered for productive-age individuals aged 15–64 years, yielding 288 instances after preprocessing. Attribute weighting was performed using Information Gain to reflect the varying diagnostic relevance of each variable, and inter-case similarity was measured through a weighted similarity approach. Classification was then carried out using KNN across multiple K values. At K = 2, the proposed method achieved an accuracy of 98.26%, with precision, recall, and F1-score each recorded at 0.983 — results that suggest the hybrid CBR-KNN approach is well-suited for deployment as a clinical decision support system for early CKD detection.

Keywords: Chronic Kidney Disease; Case-Based Reasoning; K-Nearest Neighbor; Hybrid Method; Prediction; Decision Support System.

1. Introduction

Chronic Kidney Disease (CKD) is defined by a sustained and progressive deterioration of kidney function, typically quantified through a persistent reduction in glomerular filtration rate (GFR) below 60 mL/min/1.73 m² for a period exceeding three months. Globally, CKD affects an estimated 10–15% of the adult population, imposing a substantial burden on health systems in terms of morbidity, premature mortality, and long-term treatment costs — particularly for patients who progress to end-stage renal disease requiring dialysis or transplantation (Yang *et al.*, 2024; Zhu *et al.*, 2024). What makes CKD especially difficult to manage is not its severity at advanced stages, but its silence in the early ones. Most patients remain asymptomatic until kidney function has already declined significantly, by which point the window for effective intervention has narrowed considerably. This diagnostic delay is not merely a clinical inconvenience — it reflects a structural limitation of reactive medicine that depends on laboratory confirmation rather than anticipatory screening (Tusar *et al.*, 2022; Nguycharoen, 2024).

In Indonesia, the epidemiological picture is shifting in a direction that warrants specific attention. CKD incidence is rising not only among the elderly but increasingly among the productive-age population — individuals aged 15 to 64 years — driven by a convergence of risk factors including poorly managed hypertension, type 2 diabetes, obesity, and exposure to nephrotoxic substances. This demographic shift carries economic consequences that extend beyond the healthcare sector, as working-age patients with CKD face reduced productivity, increased absenteeism, and elevated household financial strain. Despite this, most existing prediction models were developed on general or elderly patient populations, and relatively few studies have specifically targeted the productive-age group as a distinct clinical cohort (Iliyas *et al.*, 2025; Hossen *et al.*, 2025). The present study addresses this gap directly.

From a methodological standpoint, the field of CKD prediction has been dominated by supervised machine learning algorithms — among them Artificial Neural Networks, Support Vector Machines, Random Forest, and Decision Trees — all of which have demonstrated competitive classification accuracy on benchmark datasets (Khan *et al.*, 2023; Chowdhury *et al.*, 2021; Dana *et al.*, 2024). The limitations of these approaches, however, are well-documented. Neural networks and ensemble methods operate as black-box systems: they produce outputs without exposing the reasoning that generated them. In clinical settings, this opacity carries real consequences. Medical practitioners are accountable for their decisions, and a prediction system that cannot explain why it classified a patient as high-risk offers limited practical value — regardless of its aggregate accuracy (Nguycharoen, 2024; Vásquez-Morales *et al.*, 2019). Interpretability, in this context, is not a secondary design consideration; it is a prerequisite for clinical adoption.

Case-Based Reasoning (CBR) offers a fundamentally different approach. Rather than learning abstract statistical patterns from training data, CBR solves new problems by retrieving and adapting solutions from documented past cases — a process that closely mirrors the way experienced clinicians reason through differential diagnosis. This transparency makes CBR particularly well-suited to medical decision support, where the ability to trace a prediction back to specific prior cases can meaningfully support clinical judgment (Vásquez-Morales *et al.*, 2019; Hossain *et al.*, 2023). K-Nearest Neighbor (KNN), meanwhile, performs classification through similarity-based majority voting and has been shown to perform reliably on medical datasets of small to moderate size (Siregar *et al.*, 2025; Abdi & Ahmadi, 2024). Both methods share a common foundation in similarity measurement — and this is precisely where the present study intervenes.

A persistent weakness in most CBR and KNN implementations is the assumption that all attributes contribute equally to similarity or distance calculations. In clinical data, this assumption rarely holds. Serum creatinine, hemoglobin, and specific gravity carry far greater diagnostic weight for CKD than variables such as age within a restricted demographic range — yet standard implementations assign them identical influence. This flattening of clinical relevance reduces both the accuracy and the interpretability of resulting predictions. Attribute weighting through information-theoretic methods, specifically Information Gain, offers a principled correction: by assigning weights proportional to each attribute's discriminatory power, the similarity calculation becomes more clinically grounded and the retrieved cases more genuinely relevant to the query at hand (Yang *et al.*, 2024; Iliyas *et al.*, 2025).

The present study proposes a hybrid method combining CBR with Information Gain-based attribute weighting and KNN classification to predict CKD risk in the productive-age population. The dataset was sourced from the UCI Machine Learning Repository and filtered to retain records for individuals aged 15–64 years. The hybrid architecture uses CBR to compute weighted similarity between new cases and historical records, then applies KNN majority voting over the K most similar cases to determine the final classification. This design preserves the interpretability of CBR — each prediction remains traceable to specific past cases — while using KNN to reduce the misclassification risk that arises from relying on a single retrieved case. The specific objectives of this study are as follows: (1) to develop a hybrid CBR-KNN model for CKD prediction in the productive-age population; (2) to apply Information Gain-based attribute weighting within the CBR similarity function; (3) to evaluate model performance through cross-validation; and (4) to compare classification outcomes across multiple K values.

2. Related Work

Research on CKD prediction using data mining and machine learning has grown considerably over the past decade. Various classification techniques have been applied to improve early detection accuracy and support medical decision-making. This section reviews prior studies related to CKD prediction, Case-Based Reasoning (CBR), and K-Nearest Neighbor (KNN) methods, with particular attention to their respective strengths and limitations in clinical prediction. Several studies have applied machine learning methods — including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Naïve Bayes, and Random Forest — for CKD prediction, generally achieving high classification accuracy (Khan *et al.*, 2023; Chowdhury *et al.*, 2021). The limitation shared by most of these approaches is interpretability. Neural networks

and ensemble methods operate as black-box systems, producing outputs without exposing the internal reasoning that generated them. In medical decision support, this is a substantive problem: clinical personnel need to understand not only what a model predicts, but why — particularly when the prediction informs a consequential diagnostic or treatment decision. Methods that provide transparent, traceable reasoning processes therefore remain necessary in medical prediction systems, regardless of how accurate the black-box alternatives may be (Nguycharoen, 2024; Vásquez-Morales *et al.*, 2019).

Case-Based Reasoning (CBR) has been widely adopted in medical diagnosis systems precisely because it addresses this interpretability gap. CBR solves new problems by retrieving and adapting solutions from documented prior cases, closely mirroring the reasoning process experienced clinicians use in differential diagnosis. The four-stage CBR cycle — Retrieve, Reuse, Revise, and Retain — is illustrated in Figure 1. Prior research has applied CBR to the diagnosis of heart disease, diabetes, and kidney disease, demonstrating that similarity-based reasoning can produce accurate predictions while simultaneously providing explanations grounded in real historical cases (Hossain *et al.*, 2023; Vásquez-Morales *et al.*, 2019). The recurring limitation, however, is that most CBR implementations apply standard similarity calculations in which all attributes are treated as equally important. In medical datasets, this assumption is rarely valid — some clinical variables carry substantially greater diagnostic weight than others, and ignoring this reduces both retrieval quality and overall prediction performance.

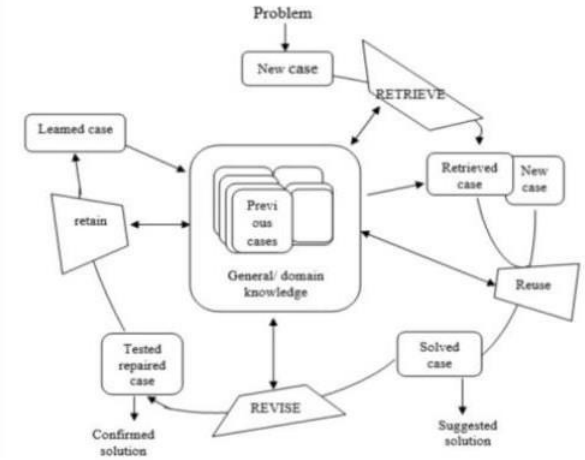


Figure 1. The CBR Cycle: Retrieve, Reuse, Revise, and Retain Stages in Case-Based Reasoning
(Source: Adapted from Aamodt & Plaza, 1994)

To address the classification limitations of CBR alone, several researchers have combined similarity-based retrieval with classification algorithms such as KNN. The KNN algorithm classifies new instances based on the majority class among the nearest neighbors in feature space. It is widely used in medical prediction tasks because of its conceptual simplicity, low computational overhead, and reliable performance on small to medium-sized datasets (Siregar *et al.*, 2025; Abdi & Ahmadi, 2024). Studies applying KNN to disease prediction have consistently reported competitive accuracy results. The method is not without limitations, however. Standard KNN assigns equal weight to all attributes in distance calculations — the same structural problem found in unweighted CBR — which can distort classification outcomes when attributes differ substantially in their clinical relevance.

In the specific context of CKD prediction, most prior studies have focused on general or elderly patient populations. Research targeting the productive-age population — individuals aged 15 to 64 years — as a distinct clinical cohort remains limited, despite evidence that CKD incidence in this group is rising due to lifestyle changes, hypertension, diabetes, obesity, and nephrotoxic drug exposure (Iliyas *et al.*, 2025; Hossen *et al.*, 2025). Prediction models calibrated specifically for this demographic are therefore still needed. A summary of representative prior studies is presented in Table 1. Kumar *et al.* (2020) compared Random Forest and SVM for CKD prediction and found that Random Forest achieved the highest accuracy. Sharma and Prasad (2021) applied KNN and reported strong classification performance attributable to the algorithm's simplicity and effectiveness on medical data. Islam *et al.* (2021) compared Decision Tree and Naïve Bayes, with Decision Tree yielding better results. Patel and Shah (2022) implemented ANN for CKD prediction and achieved high accuracy, though the model offered no interpretable output. Rahman *et al.* (2023) applied CBR to medical diagnosis and demonstrated that similarity-based reasoning can effectively support decision-making processes. Taken together, these studies confirm that machine learning methods perform well on CKD classification tasks — but most focus exclusively on accuracy, without addressing interpretability or attribute weighting in similarity calculations. Studies combining CBR and KNN within a single CKD prediction system remain scarce.

Table 1. Summary of Related Studies on CKD Prediction Methods

No	Author	Method	Dataset	Result
1	Kumar *et al.* (2020)	Random Forest, SVM	CKD UCI Dataset	Random Forest achieved highest accuracy
2	Sharma & Prasad (2021)	K-Nearest Neighbor	CKD Dataset	KNN showed good performance for CKD classification
3	Islam *et al.* (2021)	Decision Tree & Naïve Bayes	Medical Dataset	Decision Tree performed better than Naïve Bayes
4	Patel & Shah (2022)	Artificial Neural Network	CKD Dataset	ANN achieved high prediction accuracy
5	Rahman *et al.* (2023)	Case-Based Reasoning	Medical Case Dataset	CBR effective for similarity-based diagnosis
6	Proposed Study (2026)	Hybrid CBR + KNN	CKD UCI Dataset	Improves accuracy and interpretability

Based on the literature reviewed, three recurring gaps can be identified. First, most machine learning models for CKD prediction lack interpretability, limiting their practical utility in clinical settings. Second, the majority of CBR and KNN implementations do not apply attribute weighting in similarity or distance calculations, despite the well-established clinical relevance hierarchy among CKD-related variables. Third, studies that combine CBR and KNN within a unified prediction system — particularly one designed for the productive-age population — remain limited. The present study addresses all three of these gaps by proposing a hybrid CBR-KNN method with Information Gain-based attribute weighting, evaluated on a productive-age CKD dataset.

3. Methodology

3.1 Research Type

This study is experimental in design, aimed at evaluating the performance of a hybrid Case-Based Reasoning (CBR) and K-Nearest Neighbor (KNN) method with weighted similarity for predicting CKD risk in the productive-age population. The model was constructed and tested using cross-validation, with accuracy, precision, recall, and F1-score as the primary evaluation metrics. The dataset was obtained from the UCI Machine Learning Repository, comprising CKD patient records. After filtering for productive-age individuals (15–64 years), 279 instances were retained for analysis. The hybrid method combines CBR for similarity measurement between cases and KNN for final classification based on the nearest retrieved cases.

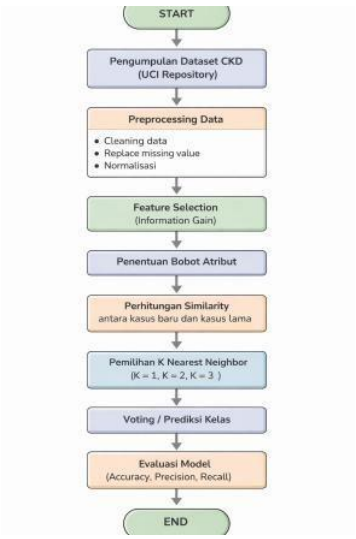


Figure 2. Research Flowchart of the Hybrid CBR-KNN Method for CKD Prediction

3.2 Research Dataset

The dataset used in this study is the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository. In its original form, the dataset contains 400 patient records across 25 attributes, covering medical parameters such as age, blood pressure, serum creatinine, hemoglobin, and other clinical indicators relevant to CKD diagnosis. Data filtering was applied using the following productive-age criterion:

$$15 < \text{age} < 64 \text{ years}$$

After filtering, 279 patient records were retained. This subset serves as the case base for the CBR component and as training data for the KNN classifier.

3.3 Research Stages

The research was conducted through the following sequential stages:

1) Collecting the CKD dataset from the UCI Machine Learning Repository
2) Data preprocessing — including data cleaning, productive-age filtering, and Min-Max normalization
3) Determining attribute weights using Information Gain
4) Calculating weighted similarity between cases using Case-Based Reasoning
5) Retrieving cases with the highest similarity values
6) Classifying new cases using KNN based on the K nearest retrieved cases
7) Evaluating model performance using cross-validation
8) Analyzing prediction results and overall model accuracy

In this study, CBR is responsible for computing similarity between new and historical cases, while KNN determines the final prediction class through majority voting among the K most similar cases.

3.4 Data Preprocessing

Preprocessing was performed to prepare the dataset prior to the classification process. Three sequential steps were applied.

1) Data Cleaning
   Missing values were handled through imputation: numeric attributes were filled with the mean value of the respective attribute, and categorical attributes were filled with the mode. This step ensures that no incomplete records introduce bias into subsequent calculations.
2) Productive-Age Filtering
   Records were restricted to patients aged between 15 and 64 years. This filtering step ensures that the prediction model is calibrated specifically to the productive-age demographic, which is the primary focus of this study.
3) Data Normalization
   Numeric attribute values were normalized using Min-Max Normalization to bring all values within a comparable range, preventing attributes with larger scales from disproportionately influencing similarity and distance calculations. The normalization formula is:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where $x$is the original attribute value, $x_{\min}$is the minimum value, and $x_{\max}$is the maximum value of the attribute.

3.5 Similarity Calculation

Local similarity measures the degree of resemblance between individual attributes of two cases. Each attribute is compared independently using a normalized distance approach, in which the absolute difference between two attribute values is divided by the attribute's range. This ensures that all attributes contribute proportionally to the similarity score regardless of their original scale. The local similarity formula for numerical attributes is:

$$\text{sim}_{\text{local}}(a_i, b_i) = 1 - \frac{|\, a_i - b_i \,|}{r_i}$$

Where $a_i$and $b_i$are the values of attribute $i$for the new case and the historical case, respectively, and $r_i$is the range of attribute i. The resulting value falls between 0 and 1, where 1 indicates identical values and 0 indicates maximum dissimilarity. An example of local similarity calculation results is presented in Table 2.

Table 2. Local Similarity Calculation Results

Case	Sim (SC)	Sim (Hemo)	Sim (SG)	Sim (PCV)	Sim (AL)	Sim (HTN)	Sim (RBCC)	Class
C1	0.952	0.909	0.75	0.931	0	1	0.619	CKD
C2	0.818	0.949	0.25	0.965	0	1	0.619	CKD
C3	0.935	0.919	0.50	1.000	0	1	0.676	CKD
C4	0.706	0.838	0.50	0.827	0	1	0.676	CKD
C5	0.939	0.959	1.00	1.000	0	1	0.500	CKD

Global similarity between a new case and each historical case is computed using a weighted similarity formula, where attribute weights are derived from Information Gain. Attributes with greater diagnostic relevance receive higher weights, allowing them to contribute more significantly to the overall similarity score. The global weighted similarity formula is:

$$\text{Sim}_{\text{global}} = \frac{\sum_{i=1}^{n} w_i \cdot \text{sim}_{\text{local}}(a_i, b_i)}{\sum_{i=1}^{n} w_i}$$

Where $w_i$is the Information Gain-based weight of attribute i, and $\text{sim}_{\text{local}}(a_i, b_i)$is the local similarity value for that attribute. The weighted similarity calculation results for each case are presented in Table 3.

Table 3. Weighted Similarity Calculation Results

Case	Sim (SC)	Sim (Hemo)	Sim (SG)	Sim (PCV)	Sim (AL)	Sim (HTN)	Sim (RBCC)	Class	Similarity Total
C1	0.952	0.909	0.75	0.931	0	1	0.619	CKD	0.745
C2	0.818	0.949	0.25	0.965	0	1	0.619	CKD	0.635
C3	0.935	0.919	0.50	1.000	0	1	0.676	CKD	0.713
C4	0.706	0.838	0.50	0.827	0	1	0.676	CKD	0.622
C5	0.939	0.959	1.00	1.000	0	1	0.500	CKD	0.789

3.6 Classification Using K-Nearest Neighbor

After weighted similarity values have been computed for all historical cases using CBR, the next step is classification using the K-Nearest Neighbor method. KNN determines the predicted class label through majority voting among the K cases with the highest similarity values. The classification procedure follows four steps:

1) Sort all similarity values in descending order
2) Select the K cases with the highest similarity values
3) Count the frequency of each class label (CKD and NOT CKD) among the K selected cases
4) Assign the class with the highest frequency as the final prediction result

The majority voting formula used in KNN classification is:

$$\hat{y} = \arg\max_{c \in C} \sum_{k=1}^{K} \mathbb{1}[y_k = c]$$

Where $\hat{y}$is the predicted class, $C$is the set of possible class labels, $y_k$is the class label of the k-th nearest neighbor, and $\mathbb{1}[\cdot]$is the indicator function that returns 1 if the condition is true and 0 otherwise.

4. Result and Discussion

4.1 Result

The dataset used in this study is the Chronic Kidney Disease dataset obtained from the UCI Machine Learning Repository. The original dataset consisted of 400 patient records with 25 attributes. After applying the productive-age filtering criterion (15–64 years), a total of 279 patient records were retained for analysis, as illustrated in Figure 3. The retained dataset was subsequently processed through preprocessing stages comprising data cleaning, Min-Max normalization, and Information Gain-based attribute weighting. The processed dataset was then used for similarity calculations using Case-Based Reasoning and final classification using the K-Nearest Neighbor method.

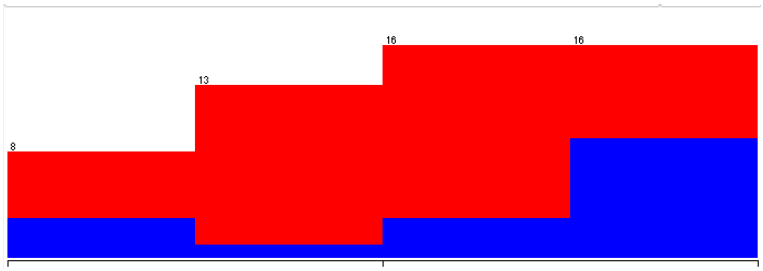


Figure 3. Distribution of Patient Records After Productive-Age Filtering (15–64 Years)

4.1.1 Attribute Weighting Results

Attribute weighting was performed using the Information Gain method to quantify the discriminatory importance of each attribute in predicting CKD. Attributes with higher Information Gain values were assigned higher weights, allowing them to contribute more significantly to the weighted similarity calculation. The complete attribute weighting results are presented in Table 4.

Table 4. Attribute Weighting Results Based on Information Gain

Attribute	Information Gain	Weight
Serum Creatinine (SC)	0.5184	1.00
Hemoglobin (Hemo)	0.5138	0.99
Specific Gravity (SG)	0.5052	0.97
Packed Cell Volume (PCV)	0.4342	0.84
Albumin (AL)	0.4053	0.78
Hypertension (HTN)	0.3354	0.65
Red Blood Cell Count (RBCC)	0.3051	0.59
Diabetes Mellitus (DM)	0.2994	0.58
Blood Urea (BU)	0.2793	0.54
Age	0.0911	0.17

The results indicate that serum creatinine, hemoglobin, and specific gravity are the three most influential attributes in predicting CKD, consistent with established clinical knowledge regarding the primary biomarkers of kidney dysfunction. Age received the lowest weight — a finding that is expected given the restricted age range of the productive-age cohort, which limits its discriminatory power within this population.

4.1.2 Similarity Calculation Results

Weighted similarity calculations were performed between each new case and all historical cases in the case base using the CBR weighted similarity formula. Similarity values range from 0 to 1, where values closer to 1 indicate higher similarity between cases. An example of the weighted similarity calculation results for five representative cases is presented in Table 5.

Table 5. Weighted Similarity Calculation Results

Case	Sim (SC)	Sim (Hemo)	Sim (SG)	Sim (PCV)	Sim (AL)	Sim (HTN)	Sim (RBCC)	Class	Similarity Total
C1	0.952	0.909	0.750	0.931	0	1	0.619	CKD	0.745
C2	0.818	0.949	0.250	0.965	0	1	0.619	CKD	0.635
C3	0.935	0.919	0.500	1.000	0	1	0.676	CKD	0.713
C4	0.706	0.838	0.500	0.827	0	1	0.676	CKD	0.622
C5	0.939	0.959	1.000	1.000	0	1	0.500	CKD	0.789

Following the similarity calculation, cases were ranked in descending order of their total similarity values to identify the K nearest neighbors. The resulting ranking is presented in Table 6.

Table 6. Case Ranking Based on Weighted Similarity

Rank	Case	Similarity	Class
1	C5	0.789	CKD
2	C1	0.745	CKD
3	C3	0.713	CKD
4	C2	0.635	CKD
5	C4	0.622	CKD

Based on this ranking, case C5 obtained the highest similarity value (0.789) and was placed at rank 1, followed by C1 (0.745), C3 (0.713), C2 (0.635), and C4 (0.622). This ranking serves as the basis for KNN classification: for K = 1, only C5 is selected; for K = 2, C5 and C1 are selected; and for K = 3, C5, C1, and C3 are selected. Since all selected neighbors belong to the CKD class across all K values, the new case is consistently classified as CKD. This example demonstrates that the ranking mechanism effectively prioritizes the most clinically relevant historical cases in the classification process.

4.1.3 Performance Evaluation

The performance of the proposed hybrid CBR-KNN method was evaluated using cross-validation across three K values. Evaluation metrics include accuracy, precision, recall, and F1-score. The results are presented in Table 7.

Table 7. Performance Evaluation Results Across K Values

K	Accuracy	Precision	Recall	F1-Score
1	97.92%	0.980	0.979	0.979
2	98.26%	0.983	0.983	0.983
3	97.22%	0.974	0.972	0.972

The best performance was achieved at K = 2, with an accuracy of 98.26%, precision of 0.983, recall of 0.983, and F1-score of 0.983. At K = 1, the model is more susceptible to noise from individual outlier cases, resulting in slightly lower performance. At K = 3, the inclusion of additional neighbors that may be less clinically relevant introduces a marginal reduction in classification accuracy. K = 2 therefore provides the most favorable balance between sensitivity and generalization for this dataset. The confusion matrix for the best-performing configuration (K = 2) is presented in Table 8.

Table 8. Confusion Matrix at K = 2

	Predicted CKD	Predicted NOT CKD
Actual CKD	155	5
Actual NOT CKD	0	128

The confusion matrix shows that the model correctly classified 155 CKD cases and 128 NOT CKD cases, with only 5 false negatives and no false positives. The absence of false positives is particularly noteworthy in a clinical context, as it indicates that no healthy patients were incorrectly flagged as having CKD.

4.1.4 Implementation Results

The proposed hybrid CBR-KNN method was implemented as a web-based decision support application to facilitate CKD prediction in clinical practice. The system was developed using PHP for backend processing, MySQL for database management, and HTML/CSS for the user interface. The application integrates all stages of the prediction workflow — data input, preprocessing, similarity calculation, classification, and result visualization — into a single interactive platform. The system comprises six main modules: the home page, patient data input, case base management, similarity processing, prediction output, and reporting. The home page provides system overview and navigation, as shown in Figure 4.

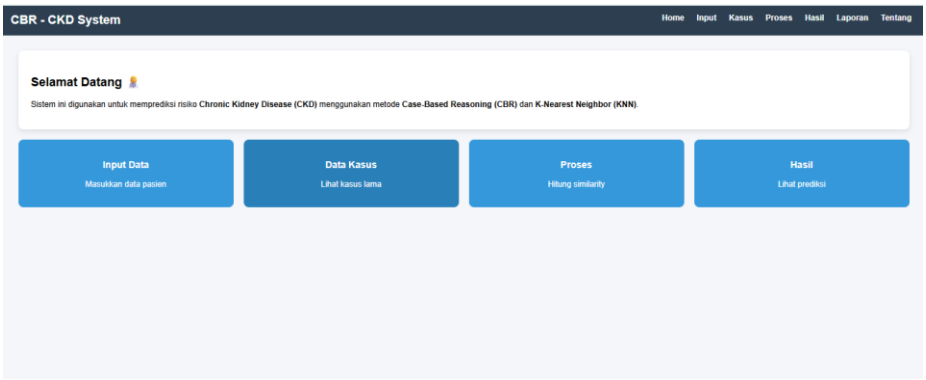


Figure 4. Main Page Display of the Web-Based CBR-KNN System

The patient data input module allows users to enter clinical attributes including age, blood pressure, serum creatinine, hemoglobin, and other relevant parameters. The case base module displays the historical patient records stored in the database, which serve as reference cases in the CBR retrieval process. The similarity processing module computes weighted similarity between the new patient data and all existing cases, then ranks the results accordingly, as shown in Figure 5.

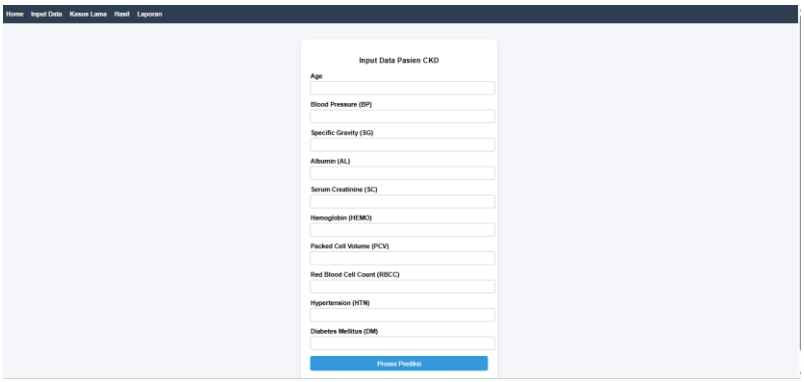
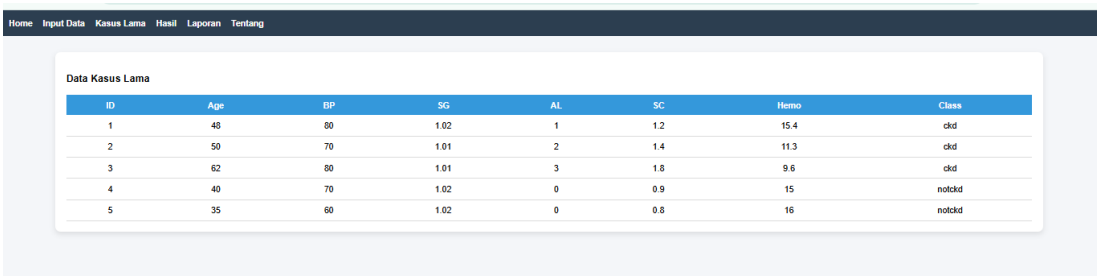


Figure 5. Classification Result Display

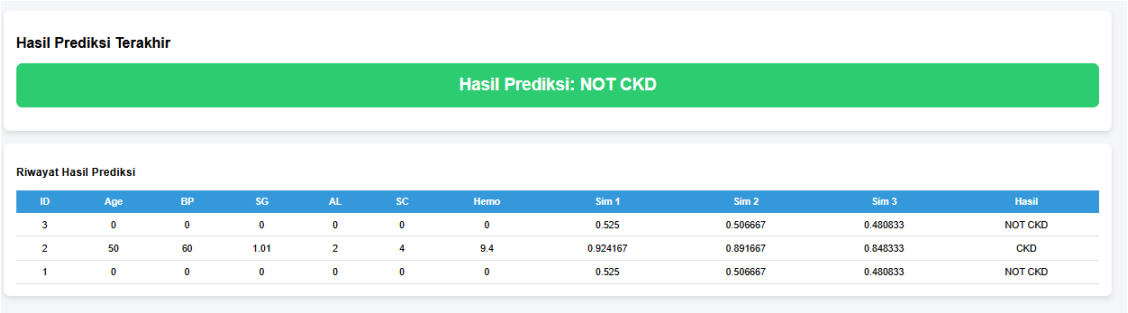
The prediction module applies KNN majority voting over the top K ranked cases to produce a final classification label — either CKD or NOT CKD — along with the individual similarity values of the retrieved cases to support interpretability, as shown in Figure 6.



ID	Age	BP	SG	AL	SC	Hemo	Class
1	48	80	1.02	1	1.2	15.4	ckd
2	50	70	1.01	2	1.4	11.3	ckd
3	62	80	1.01	3	1.8	9.6	ckd
4	40	70	1.02	0	0.9	15	notckd
5	35	60	1.02	0	0.8	16	notckd

Figure 6. Classification Based on Nearest Cases

The reporting module presents prediction outcomes in tabular and graphical formats, enabling users to review system performance and the distribution of CKD and NOT CKD predictions over time, as shown in Figure 7.



ID	Age	BP	SG	AL	SC	Hemo	Sim 1	Sim 2	Sim 3	Hasil
3	0	0	0	0	0	0	0.525	0.506667	0.480833	NOT CKD
2	50	60	1.01	2	4	9.4	0.924167	0.891667	0.848333	CKD
1	0	0	0	0	0	0	0.525	0.506667	0.480833	NOT CKD

Figure 7. Prediction Report Display

The system implementation confirms that the hybrid CBR-KNN method can be effectively integrated into a web-based decision support environment. The application performs data processing, weighted similarity calculation, classification, and result visualization in an efficient and user-friendly manner.

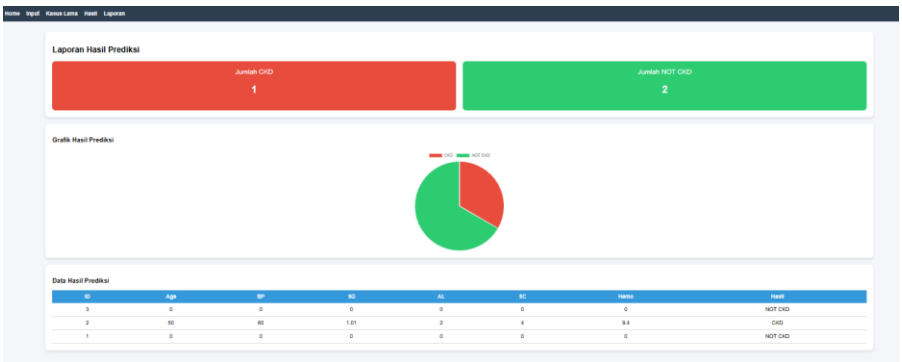


Figure 8. Report

4.2 Discussion

The experimental results demonstrate that the proposed hybrid CBR-KNN method is effective for predicting CKD in the productive-age population. The combination of similarity-based reasoning and majority-vote classification yields both high predictive accuracy and transparent, traceable outputs — two properties that are essential in clinical decision support systems. The application of Information Gain-based attribute weighting substantially improves the quality of the CBR similarity calculation. By assigning higher weights to clinically dominant attributes — particularly serum creatinine, hemoglobin, and specific gravity — the similarity function more accurately reflects the diagnostic relevance of each variable. This produces more discriminative similarity values compared to unweighted approaches, in which all attributes contribute equally regardless of their actual predictive importance. The weighting results are consistent with established clinical understanding of CKD biomarkers, lending further validity to the approach. The comparative evaluation across K values reveals that K = 2 yields the best overall performance, with an accuracy of 98.26% and an F1-score of 0.983. At K = 1, the model is more sensitive to noise, as the classification depends entirely on a single retrieved case. At K = 3, the inclusion of a third neighbor — which may be less similar to the query case — introduces marginal performance degradation. K = 2 therefore strikes the most effective balance between stability and accuracy for this dataset and population. The hybrid architecture contributes to this outcome: by using KNN majority voting over CBR-ranked cases rather than relying on a single retrieved case, the model reduces the misclassification risk inherent in single-case reasoning. From an implementation standpoint, the web-based system developed in this study successfully operationalizes the hybrid method in a practical, user-accessible format. The inclusion of similarity values alongside prediction outputs enhances interpretability, allowing clinical users to understand not only the predicted label but also which historical cases informed it. This transparency distinguishes the proposed system from black-box machine learning models and supports its potential utility as a clinical decision aid for early CKD detection in the productive-age population.

5. Conclusion and Future Work

This study proposed a hybrid Case-Based Reasoning and K-Nearest Neighbor method with Information Gain-based attribute weighting for predicting Chronic Kidney Disease in the productive-age population. The dataset was obtained from the UCI Machine Learning Repository and filtered to retain records for individuals aged 15 to 64 years, yielding 279 patient records for experimentation. The research pipeline encompassed data preprocessing, attribute weight determination using Information Gain, weighted similarity calculation using Case-Based Reasoning, and final classification using the K-Nearest Neighbor method. The experimental results demonstrate that the proposed hybrid method achieves strong predictive performance. The best results were obtained at K = 2, with an accuracy of 98.26%, precision of 0.983, recall of 0.983, and F1-score of 0.983. The application of Information Gain-based attribute weighting improved the discriminative quality of the similarity calculation by assigning greater influence to clinically significant attributes — particularly serum creatinine, hemoglobin, and specific gravity — thereby producing more accurate and clinically grounded similarity values compared to unweighted approaches. The integration of KNN majority voting over CBR-ranked cases further reduced the misclassification risk associated with single-case reasoning. The implementation of the method as a web-based decision support application confirmed that the hybrid CBR-KNN system can be effectively deployed in a practical clinical environment. The system provides transparent prediction outputs traceable to specific historical cases, offering an interpretability advantage over conventional black-box machine learning models. These findings collectively indicate that the hybrid CBR-KNN method is a viable and interpretable approach for early CKD detection in the productive-age population.

For future work, several directions are recommended. First, the model should be validated on larger and more diverse datasets to assess its generalizability across different clinical settings and patient demographics. Second, the inclusion of additional medical attributes — such as urine output, serum electrolytes, and imaging findings — may further improve prediction performance. Third, comparative evaluations against other machine learning methods, including Random Forest, Support Vector Machine, and Neural Networks, would provide a more comprehensive assessment of the proposed method's relative strengths and limitations. Finally, prospective clinical validation in real healthcare settings is encouraged to assess the system's practical utility and reliability in supporting medical decision-making.

Acknowledgment

The authors would like to thank the UCI Machine Learning Repository for providing the Chronic Kidney Disease dataset used in this study. The authors also express their sincere gratitude to all individuals and institutions

whose guidance, support, and contributions facilitated the completion of this research and the development of the proposed system.

References

Abdi, N. F., & Ahmadi, M. F. (2024). Klasifikasi penyakit ginjal kronis (CKD) dengan algoritma KNN dan Decision Tree ID3. *Journal of Informatics and Advanced Computing (JIAC)*, *5*(2), 52–57. https://doi.org/10.35814/jiac.v5i2.7189

Chowdhury, N. H., Reaz, M. B. I., Haque, F., Ahmad, S., Ali, S. H. M., A Bakar, A. A., & Bhuiyan, M. A. S. (2021). Performance analysis of conventional machine learning algorithms for identification of chronic kidney disease in type 1 diabetes mellitus patients. *Diagnostics*, *11*(12), 2267. https://doi.org/10.3390/diagnostics11122267

Dana, Z., Naseer, A. A., Toro, B., & Swaminathan, S. (2024). Integrated machine learning and survival analysis modeling for enhanced chronic kidney disease risk stratification. *arXiv*. https://doi.org/10.48550/arXiv.2411.10754

Dipto, I. C., Islam, T., Rahman, H. M., & Rahman, M. A. (2020). Comparison of different machine learning algorithms for the prediction of coronary artery disease. *Journal of Data Analysis and Information Processing*, *8*(2), 41–68. https://doi.org/10.4236/jdaip.2020.82003

Hossain, M. S., Ahmed, F., & Rahman, M. M. (2023). An intelligent case-based reasoning system for disease diagnosis. *IEEE Access*, *11*, 45910–45922.

Hossen, M. J., Bannah, H., & Sadib, R. J. (2025). Early detection of chronic kidney disease using deep learning: A mini review. *Frontiers in Digital Health*, *7*, 1732175. https://doi.org/10.3389/fdgth.2025.1732175

Iliyas, I. I., Boukari, S., & Gital, A. Y. U. (2025). Recent trends in prediction of chronic kidney disease using different learning approaches: A systematic literature review. *Journal of Medical Artificial Intelligence*, *8*, 62. https://doi.org/10.21037/jmai-24-256

Khan, N., Raza, M. A., Mirjat, N. H., Balouch, N., Abbas, G., Yousef, A., & Touti, E. (2023). Unveiling the predictive power: A comprehensive study of machine learning models for anticipating chronic kidney disease. *Frontiers in Artificial Intelligence*, *6*, 1339988. https://doi.org/10.3389/frai.2023.1339988

Kumar, A., Sharma, G. K., & Prakash, U. M. (2021). Disease prediction and doctor recommendation system using machine learning approaches. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *9*, 34–44.

Kumar, P., Singh, A., & Kumar, R. (2023). Chronic kidney disease prediction using machine learning algorithms. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *11*(5). https://doi.org/10.22214/ijraset

Li, Y., & Padman, R. (2025). Enhancing end-stage renal disease outcome prediction: A multisourced data-driven approach. *Journal of the American Medical Informatics Association*, *33*(1), 26–36. https://doi.org/10.1093/jamia/ocaf118

Nguycharoen, N. (2024). Explainable machine learning system for predicting chronic kidney disease in high-risk cardiovascular patients. *arXiv*. https://doi.org/10.48550/arXiv.2404.11148

Siregar, M. R., Hartama, D., & Solikhun, S. (2025). Optimizing the KNN algorithm for classifying chronic kidney disease using GridSearchCV. *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, *10*(3), 680–689. https://doi.org/10.33480/jitk.v10i3.6214

Sivakumar, R., & Vijayalakshmi, R. (2024). A survey of predicting CKD using machine learning. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(4), 3176–3182.

Tusar, M. T. H. K., Islam, M. T., & Raju, F. I. (2022, March 9–10). Detecting chronic kidney disease (CKD) at the initial stage: A novel hybrid feature-selection method and robust data preparation pipeline for different ML techniques. *In Proceedings of the 2022 5th International Conference on Computing and Informatics (ICCI)* (pp. 400–407). IEEE. https://doi.org/10.1109/ICCI54321.2022.9756094

Vásquez-Morales, G. R., Martinez-Monterrubio, S. M., Moreno-Ger, P., & Recio-Garcia, J. A. (2019). Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning. *IEEE Access*, *7*, 152900–152910. https://doi.org/10.1109/ACCESS.2019.2948430

Yang, W., Ahmed, N., & Barczak, A. L. (2024). Comparative analysis of machine learning algorithms for CKD risk prediction. *IEEE Access*, *12*, 171205–171220. https://doi.org/10.1109/ACCESS.2024.3499355

Zhu, H., Qiao, S., Zhao, D., Wang, K., Wang, B., Niu, Y., Shang, S., Dong, Z., Zhang, W., Zheng, Y., & Chen, X. (2024). Machine learning model for cardiovascular disease prediction in patients with chronic kidney disease. *Frontiers in Endocrinology*, *15*, 1390729. https://doi.org/10.3389/fendo.2024.1390729.