International Journal of Innovative Computing, Information and Control
Volume **22**, Number **3**, June **2026**

ICIC International ©2026 ISSN 1349-4198
pp. 755–766

A KIDNEY DISEASE DATA ANALYSIS APPROACH BASED ON MODIFIED INFORMATION GRANULATION AND DATA RECONSTRUCTION

Wenyan Song* and Xiangtai Meng

School of Economics
Dongbei University of Finance and Economics
No. 217, Jianshan Street, Shahekou District, Dalian 116025, P. R. China
mxtdufe@163.com
*Corresponding author: songwydufe@dufe.edu.cn

Received June 2025; revised September 2025

Abstract. *To improve model interpretability in clinical medical datasets such as chronic kidney disease data, this study utilizes information granular technique to propose a data reconstruction and feature selection method. First, the information granulation method grounded in the principle of justifiable granularity is applied for data reconstruction. Then, a modified Davies-Bouldin index is developed as an evaluation criterion for a set of information granules. Furthermore, a novel task-oriented loss function integrating classification objective is designed, and a heuristic feature selection algorithm is developed through cross-validation procedures. Experimental validation on chronic kidney disease datasets demonstrates that the proposed information granulation-based data reconstruction method effectively identifies critical features while maintaining satisfactory classification performance.*

Keywords: Information granulation, Data reconstruction, Feature selection, Chronic kidney disease

1. **Introduction.** According to World Health Organization estimates, chronic kidney disease (CKD) affects over 850 million individuals globally and is projected to become the fifth leading cause of reduced life expectancy by 2040. In recent years, many medical and research institutions have been continuously promoting pathological research on kidney disease, while also collecting and establishing patient case databases of different scales, accumulating valuable data resources for evidence-based medicine research on chronic kidney disease, especially end-stage renal disease.

Current analytical approaches for CKD clinical data are broadly classified into two categories: statistical modeling and machine learning techniques. Based on the Australia and New Zealand dialysis and transplant registry data, a multilevel logistic (Logit) regression model was constructed to identify the influencing factors of late referral for renal replacement therapy [1]. Cox regression and LASSO models were systematically employed to pinpoint metabolites with prognostic value for diabetic kidney disease progression [2]. And another study also used these two models to construct equation for predicting the risk of chronic kidney disease in lithium created patients [3]. Meanwhile, machine learning techniques represented by support vector machine and random forest, as well as some deep learning methods, were reviewed for their application in kidney disease investigations, especially in gene phenotype analysis [4]. By applying gradient tree boosting for variable selection, a simplified scoring scale model was adopted to predict long-term outcomes

DOI: 10.24507/ijicic.22.03.755

for immunoglobulin A nephropathy patients in a retrospective cohort study [5]. And a model combining convolutional neural network and support vector machine was designed to deal with the classification of CKD [6]. Statistical models remain prevalent in clinical research due to their parsimonious structure, strong interpretability, and practical utility. However, the utility of such models is often constrained by limited generalizability, stemming from restricted study timeframes and highly selective patient cohorts that result in suboptimal sample sizes and diminished external validity. In contrast, machine learning methods achieve enhanced predictive performance through sophisticated data transformations, though their complex information processing mechanisms often compromise model interpretability.

In recent years, information granulation theory and techniques have attracted significant attention for their ability to balance model interpretability with classification or regression performance in complex data processing. These methods are widely employed in data analysis, particularly in sample classification, by extracting and refining representative multidimensional features from datasets. For instance, an enhanced granular ball computing method was discussed in [7], demonstrating robust classification performance across 24 benchmark datasets. A dominance neighborhood granularity-based classification method capable of handling both Euclidean and non-Euclidean features was proposed in [8]. Further applications of these techniques, through hybrid modeling approaches, have yielded successful results in medical datasets such as appendicitis, breast cancer, heart disease, and Parkinson's disease [9-11]. However, as data dimensionality increases, critical challenges arise in designing effective multidimensional information granules while preserving feature semantics, reducing redundant information interference, and ensuring computational efficiency. Addressing these challenges remains a pivotal research gap requiring systematic exploration.

In this study, an information granulation-based solution for multidimensional data classification in kidney disease diagnosis is proposed. The approach begins by the design of an information granulation method, followed by the development of a feature selection mechanism that incorporates enhanced granulation evaluation criteria. The framework effectively identifies typical sample representations and selects feature subsets with maximal classification relevance. By optimizing information granule construction and performing data reconstruction, the method achieves significant accuracy improvement. Crucially, the proposed method maintains clinical interpretability by preserving original medical meanings of features and evaluating their deviation from typical representations, thus ensuring both the selected features and final models being clinically meaningful.

The primary contributions of this paper are as follows.

1) A granular data reconstruction method is designed, which employs information granulation techniques to generate clinically meaningful feature representations while preserving semantic alignment with original clinical attributes.

2) An integrated evaluation metric combining information granule quality assessment with discriminative feature analysis is developed to optimize feature subset selection through quantifiable granularity measurements.

The remainder of the paper is organized as follows. Section 2 presents the principle of information granulation and the corresponding data reconstruction method. Section 3 proposes granulation criteria indicators for a set of information granules and uses them to select features. Section 4 validates the proposed analytical framework through comprehensive experiments on two kidney disease datasets. Finally, Section 5 summarizes the main conclusions and the future research.

2. **Information Granulation and Data Reconstruction.** This section mainly introduces the basic principles of information granulation and the process of reconstructing datasets based on generated information granules.

For a kidney disease data set $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\} \subseteq R^{n\times d}$, with n samples and d features, assume that the patient characteristics comprise mixed data types, including nominal (unordered) and ordinal (ordered) discrete variables, integer-valued measurements and continuous real-valued variables. In order to ensure the original demographic and physiological meanings of the features, discrete variables are not subjected to granulation processing, and integer and real features are subjected to granulation and data reconstruction. To simplify the notation, $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is still used to represent the dataset that requires granulation and processing. For the k-th feature, all samples are represented as $X^{(k)} = \left\{x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right\}$, $k \in \{1, 2, \ldots, d\}$.

Firstly, the fuzzy C-means (FCM) algorithm [12] is used to cluster the group $X^{(k)}$ into $c^{(k)}$ categories. The FCM algorithm aims to minimize the following objective function:

$$J_{FCM}^{(k)} = \sum_{i=1}^{n}\sum_{j=1}^{c^{(k)}} u_{ij}^{m(k)} \left\| x_i^{(k)} - v_j^{(k)} \right\|_2^2, \tag{1}$$

where $u_{ij}^{m(k)}$ represents the membership degree of the k-th feature of the i-th sample to the j-th cluster of the feature, satisfying $0 \le u_{ij}^{m(k)} \le 1$, $\sum_{j=1}^{c^{(k)}} u_{ij}^{m(k)} = 1$, $i = 1, 2, \ldots, n$. $v_j^{(k)}$ is the cluster prototype of the j-th cluster, $j = 1, 2, \ldots, c^{(k)}$. $u_{ij}^{m(k)}$ and $v_j^{(k)}$ are determined as follows, respectively:

$$u_{ij}^{m(k)} = 1 \Bigg/ \sum_{l=1}^{c^{(k)}} \left(\frac{\left\| x_i^{(k)} - v_j^{(k)} \right\|_2}{\left\| x_i^{(k)} - v_l^{(k)} \right\|_2} \right)^{\frac{2}{m-1}}, \quad v_j^{(k)} = \frac{\sum_{i=1}^{n} u_{ij}^{m(k)} x_i^{(k)}}{\sum_{i=1}^{n} u_{ij}^{m(k)}}.$$

Next, a set of interval-valued information granules is generated using all the prototypes of the clustering cluster $\left\{v_j^{(k)}\right\}$ and the membership degrees of the data points for each cluster. Consider a set of information granules $\left\{\left(l_j^{(k)}, v_j^{(k)}, r_j^{(k)}\right)\right\}$, where $l_j^{(k)}$ and $r_j^{(k)}$ are the left and right endpoints of the j-th interval, respectively, $l_j^{(k)} < v_j^{(k)} < r_j^{(k)}$, and they may not be symmetrical about $v_j^{(k)}$. According to the principle of justifiable granularity proposed in [13], the interval endpoints are determined by optimizing the single information granule quality evaluation indicators coverage $Q_{jl}^{(k)}$ and specificity $Q_{jr}^{(k)}$:

$$Q_{jl}^{(k)}\left(X^{(k)}\right) = \underset{l_j^{(k)}}{\operatorname{argmax}} \left(\operatorname{Cov}\left(X^{(k)}, l_j^{(k)}\right) \cdot \operatorname{Sp}^{\beta}\left(X^{(k)}, l_j^{(k)}\right)\right), \tag{2}$$

$$Q_{jr}^{(k)}\left(X^{(k)}\right) = \underset{r_j^{(k)}}{\operatorname{argmax}} \left(\operatorname{Cov}\left(X^{(k)}, r_j^{(k)}\right) \cdot \operatorname{Sp}^{\beta}\left(X^{(k)}, r_j^{(k)}\right)\right), \tag{3}$$

where
$\operatorname{Cov}\left(X^{(k)}, l_j^{(k)}\right) = \sum_{x_i^{(k)} \in \left[l_j^{(k)}, v_j^{(k)}\right] \cap D_j^{(k,\alpha)}} \frac{l_j^{(k)} - x_i^{(k)}}{l_j^{(k)} - v_j^{(k)}}$, $\operatorname{Sp}\left(X^{(k)}, l_j^{(k)}\right) = 1 - 0.5 \cdot \frac{l_j^{(k)} - v_j^{(k)}}{\min\left\{D_j^{(k,0)}\right\} - v_j^{(k)}}$,
$\operatorname{Cov}\left(X^{(k)}, r_j^{(k)}\right) = \sum_{x_i^{(k)} \in \left[v_j^{(k)}, r_j^{(k)}\right] \cap D_j^{(k,\alpha)}} \frac{r_j^{(k)} - x_i^{(k)}}{r_j^{(k)} - v_j^{(k)}}$, $\operatorname{Sp}\left(X^{(k)}, r_j^{(k)}\right) = 1 - 0.5 \cdot \frac{r_j^{(k)} - v_j^{(k)}}{\max\left\{D_j^{(k,0)}\right\} - v_j^{(k)}}$,
and $D_j^{(k,\alpha)} = \left\{x_i^{(k)} \middle| u_{ij}^{(k)} > \alpha\right\}$ is a subset of $X^{(k)}$. α represents the level of membership, where $\alpha \in (0, 1)$. β represents the level of emphasis placed on feature degree. $D_j^{(k,\alpha)}$ can

also be determined by the principle of maximum membership degree. $x_i^{(k)}$ can be clearly divided into $c^{(k)}$ categories according to $\underset{j}{\operatorname{argmax}}\, u_{ij}^{(k)}$ correspondingly. According to the reconstruction idea in [14,15], $x_i^{(k)}$ can be restructured as

$$\hat{x}_i^{(k)} = \sum_{j:u_{ij}^{(k)}>\alpha} u_{ij}^{(k)} v_j^{(k)} \Bigg/ \sum_{j:u_{ij}^{(k)}>\alpha} u_{ij}^{(k)}. \tag{4}$$

It is worth noting that this data reconstruction method not only preserves feature semantics, but also mitigates interference from noise and detection operation variances. This is achieved through quantitative comparison between specific numerical values and their corresponding information granules.

3. **Feature Selection.** This section presents a method of feature selection by optimizing the following modified evaluation function:

$$J = J_{Model} + \lambda J_{Features}, \tag{5}$$

where λ is a hyperparameter. $J_{Model} = 1-\text{F1}$, in which the F1-score is a statistical measure that reflects the effectiveness of classification models. And $J_{Features}$ is an indicator that reflects the quality of information granules, which is composed of the modified Davies-Bouldin index [16,17].

The process of computing $J_{Features}$ is as follows. First, the method proposed in Section 2 is used for information granulation and data reconstruction of features with continuous variable values. The dataset is randomly grouped for cross validation preparation. Given a dataset of candidate features $X^{(k)}$, the training subset is clustered within each cross-validation fold. Using the resulting cluster centers and membership degrees, the dispersion degree and the inter class distance are computed by

$$S_j^{(k)} = \sum_{i:u_{ij}^{(k)}>\alpha} u_{ij}^{m(k)} \left\| x_i^{(k)} - v_j^{(k)} \right\|_2^2 \Bigg/ \sum_{i:u_{ij}^{(k)}>\alpha} u_{ij}^{m(k)}, \tag{6}$$

$$M_{jl}^{(k)} = \sum_{i=1}^{n} \frac{1}{2} \left(\left| l_j^{(k)} - l_l^{(k)} \right| + \left| r_j^{(k)} - r_l^{(k)} \right| \right), \tag{7}$$

where $S_j^{(k)}$ quantifies the intra-cluster cohesion and $M_{jl}^{(k)}$ reflects the differences between different categories. Alternatively, $M_{jl}^{(k)}$ can be directly computed by $M_{jl}^{(k)} = \left\| v_j^{(k)} - v_l^{(k)} \right\|_2^2$. Further, the similarity between classes is calculated as $R_{jl}^{(k)} = \left(w_j^{(k)} S_j^{(k)} + w_l^{(k)} S_l^{(k)} \right) \Big/ M_{jl}^{(k)}$, where $w_j^{(k)} = \sum_{i:u_{ij}^{(k)}>\alpha} u_{ij}^{(k)} \Big/ \sum_{i:u_{ij}^{(k)}>0} u_{ij}^{(k)}$ is the weight that reflects the coverage of the j-th cluster. The smaller R_{jl} is, the greater the discrimination between the two clusters is.

Next, the modified Davies-Bouldin index of the k-th feature is computed as

$$\text{MDBI}^{(k)} = \frac{1}{c^{(k)}} \sum_{j=1}^{c^{(k)}} \max_{l \neq j} R_{jl}^{(k)}. \tag{8}$$

Since optimal clustering requires maximal intra-cluster similarity and inter-cluster dissimilarity, smaller values of the indicator MDBI represent better clustering performance. Furthermore, the design of this indicator adheres to the fundamental criteria for information granulation when constructing information granules. For candidate feature combinations,

the corresponding $\mathrm{MDBI}^{(k)}$ values of each feature are summed, and the second term in the loss function is defined as

$$J_{Features} = \sum_{k} \mathrm{MDBI}^{(k)}. \tag{9}$$

In summary, a forward heuristic feature selection algorithm is proposed, which integrates wrapper and embedded methods [18], while maintaining model interpretability. Given d candidate features, the dataset is randomly partitioned into several groups. Model performance is evaluated by sequentially adding features to the current feature combination across these datasets. Cross-validation is then employed to assess whether each newly added feature should be retained. The implementation process of the feature selection algorithm is shown in Table 1, which mainly includes five main steps:

Step 1. Initialization Phase. Initialize the feature subset as either an empty set, or a predefined feature set containing domain-specific essential features.

Step 2. Candidate Feature Selection. At each iteration, identify unexamined features from the complete feature space, and construct a candidate set for assessment.

Step 3. Loss Computation. For each candidate feature, partition the dataset into training and validation sets via cross-validation, apply the granularity transformation method (Section 2 and Section 3) to continuous variables, and compute the loss function value using Equation (5).

Table 1. Feature selection algorithm process

Inputs: data matrix X $(n \times d)$, class label vector Y $(n \times 1)$, feature number set F, feature number set *KeepIn* that must be included (*KeepIn* can be $\emptyset$), the final number of selected features card(*Inmodel*) (card(*KeepIn*) $\leq$ card(*Inmodel*) $\leq d$);
Outputs: number set of selected features *Inmodel*;
Initialization: *Inmodel*(0) = *KeepIn*, $F(0) = F \backslash$ *Inmodel*(0), iterative counter $t = 0$, maximum number of iterations *maxiter* (*maxiter* $\leq$ card(*Inmodel*)), loss tolerance *tol*.

```
while Flag == true
    t = t + 1;
    Generate candidate feature set:
        Generate a candidate feature set F(t) based on the current Inmodel(t−1)
        (including a single feature based on F(t − 1));
      for each candidate feature k ∈ F(t):
      Generate temporary feature number subset:
        Tmp_Inmodel = Inmodel(t − 1) ∪ {k};
    Split the dataset into training and validation sets;
    Calculate the loss: L_k = J(X(:, Tmp_Inmodel), Y);
    Determine the selected feature:
        k* = argmin_k L_k;
    Determine whether to update:
        if L_k* < current optimal loss:
            Inmodel(t) = Inmodel(t) ∪ (k*);
            Update current optimal loss L_k = L_k*, F(t) = F(t){k*};
        else
            if t > maxiter or L_k < tol
            Flag == false;
end while
return: Inmodel
```

Step 4. Feature Inclusion. Select features satisfying lower mean loss value across cross-validation folds.

Step 5. Termination Condition. The iterative process terminates upon satisfying any of the following conditions: (a) the loss function value demonstrates convergence, (b) the computed loss falls below a predefined threshold, or (c) the maximum allowable iteration count is attained.

4. **Examples.** This section focuses on the Chronic Kidney Disease dataset from the UCI dataset (UCI-CKD, http://archive.ics.uci.edu) and an end-stage renal disease clinical cohort study dataset. The proposed methodology demonstrates robust capability in processing and selecting clinically relevant features from complex medical datasets. Quantitative evaluation of feature selection performance is conducted using the UCI-CKD dataset, while the end-stage renal disease dataset is employed to validate classification accuracy improvements achieved through data reconstruction techniques. In renal disease diagnosis, a representative clinical decision-making task, logistic regression combined with data granulation and feature reconstruction is implemented. This model is selected for its dual advantages of interpretability and computational efficiency, which are critical for clinical applications.

The evaluation metrics employed in this section include the F1-score and the area under the receiver operating characteristic (ROC) curve, both standard measures for assessing binary classification models. The F1-score is defined as the harmonic mean of precision and recall, providing a balanced performance assessment that accounts for both false positives and false negatives. The calculations are

$$Precision(P) = \frac{TP}{TP + FP},\ Recall(R, Sensitivity) = \frac{TP}{TP + FN},\ F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{10}$$

where TP (True Positives) denotes correctly classified target-class instances, FP (False Positives) represents non-target instances misclassified as target class, and FN (False Negatives) indicates target-class instances misclassified as non-target. Precision measures a model's resistance to false positives, while Recall evaluates its capability to detect all positive instances. The area under the ROC curve (AUC) evaluates the discriminative ability between positive and negative classes across all classification thresholds. It is computed from the ROC curve, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). Higher F1-score and AUC values indicate superior classification performance.

Example 4.1. *CKD data set from UCI. There are 400 samples in the UCI-CKD dataset, covering 24 attributes and 1 class label (CKD or non-CKD). Specifically, the 24 features are age (x_1), blood pressure (x_2), specific gravity (x_3), albumin (x_4), sugar (x_5), red blood cells (x_6), pus cell (x_7), pus cell clumps (x_8), bacteria (x_9), blood glucose random (x_{10}), blood urea (x_{11}), serum creatinine (x_{12}), sodium (x_{13}), potassium (x_{14}), hemoglobin (x_{15}), packed cell volume (x_{16}), white blood cell count (x_{17}), red blood cell count (x_{18}), hypertension (x_{19}), diabetes mellitus (x_{20}), coronary artery disease (x_{21}), appetite (x_{22}), pedal edema (x_{23}), and anemia (x_{24}). There are 11 numeric and 13 nominal data types in these 24 attributes. This dataset analyzes the performance of patient attributes to diagnose whether they have CKD, which is a binary classification problem. The class labels given by doctors in the dataset are 250 CKD patients and 150 non-CKD patients. Following the KDIGO clinical practice guidelines 2024 (Kidney Disease: Improving Global Outcomes, KDIGO [19]), albumin (x_4), blood urea (x_{11}), serum creatinine (x_{12}), and hemoglobin (x_{15}) are core mandatory features for CKD diagnosis and evaluation.*

The classification analysis initially utilizes all 24 clinical attributes after standard data normalization. Multiple interpretable classifiers are employed to evaluate feature importance through 10 iterative experiments with randomized 70% training and 30% testing partitions. As demonstrated in Table 2, while all methods attain acceptable accuracy levels, significant limitations are observed regarding serum creatinine identification across linear discriminant analysis (LDA), support vector machines (SVM), decision trees and random forest algorithms. These technical constraints may derive from three interrelated causes. First, restricted sample sizes reduce statistical power. Second, incomplete attribute value distributions hinder pattern recognition. Third, potential deviations exist between laboratory measurements and actual physiological conditions. To address these issues, comparative analysis with information granules is proposed for data transformation, as these structures effectively capture characteristic feature representations. Subsequent diagnostic models should therefore prioritize the refined features generated through this process to enhance clinical validity.

Table 2. Classification results of the base classifiers

Model	F1	Selected top five important features
Linear discriminant analysis	0.9507	specific gravity (x_3), hemoglobin (x_{15}), albumin (x_4), packed cell volume (x_{16}), diabetes mellitus (x_{20})
Support vector machine	0.9717	packed cell volume (x_{16}), albumin (x_4), pedal edema (x_{23}), blood glucose random (x_{10}), appetite (x_{22})
Decision tree	0.9583	hemoglobin (x_{15}), specific gravity (x_3), red blood cell count (x_{18}), sodium (x_{13}), albumin (x_4)
Random forest	0.9728	packed cell volume (x_{16}), hemoglobin (x_{15}), specific gravity (x_3), albumin (x_4), serum creatinine (x_{12})

The subsequent analysis implements the granulation framework to transform all numerical variables into information granules, enhancing clinical interpretability while maintaining diagnostic fidelity. Considering that this is a diagnostic recognition problem, the number of clusters $c^{(k)}$ is set to 2 and α is set to 0.5. The selected features are shown in Table 3.

It reveals that across various initial candidate feature sets, the selected important features almost all contain hemoglobin (x_{15}), albumin (x_4) and serum creatinine (x_{12}), which is highly consistent with existing kidney disease research guidelines and clinical diagnosis and treatment experience. In most cases, blood glucose random (x_{10}) is also selected. [20] employed GridSearchCV to analyze this dataset, utilizing 8 models and 4 variable selection methods for feature identification. In contrast, our approach prioritizes the detection of the clinically significant feature x_{15} through a more parsimonious model, with subsequent selection of x_4 and x_{12}. In addition, the last three rows of Table 3 also show that progressively expanding the feature set size enhances classification performance. These results confirm that our method successfully balances predictive accuracy with clinically meaningful feature prioritization.

Table 4 presents the Logit model and parameters established using variables x_4, x_{10}, x_{12} and x_{15}. The second column displays the regression coefficients for the constant term and each variable, while the third and fourth columns report the standard deviations and t-statistics, respectively. The fifth column presents the p-values corresponding to the t-statistics. All p-values are statistically significant ($p < 0.05$), confirming the relevance of these variables. Furthermore, the signs of the regression coefficients align with findings from prior medical research.

TABLE 3. Selected features and classification effect

	Cost function J		Cost function J_{Model}	
Initial candidate features	Selected features and selection order	F1	Selected features and selection order	F1
$\emptyset$	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18}$	0.9691	$x_{15} \to x_4 \to x_{10} \to x_{19} \to x_{12}$ (x_{19} is not significant)	0.9754
$\{x_1\}$	$x_1 \to x_4 \to x_{15} \to x_{10} \to x_{12}$ (x_1 is not significant)	0.9669	$x_1 \to x_{15} \to x_4 \to x_{10} \to x_{19}$ (x_1 and x_{19} are not significant)	0.9669
$\{x_3\}$	$x_3 \to x_{15} \to x_4 \to x_{10} \to x_{12}$ (x_3 is not significant)	0.9585	$x_3 \to x_{15} \to x_{10} \to x_{18} \to x_{20}$ (x_3 and x_{20} are not significant; exclude x_{19} in advance)	0.9605
$\{x_{11}\}$	$x_{11} \to x_{15} \to x_4 \to x_{10} \to x_{12}$	0.9691	$x_{11} \to x_{15} \to x_4 \to x_{10} \to x_{12}$ (exclude x_{19} and x_{20} in advance)	0.9691
$\emptyset$	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{11}$	0.9691	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{20}$ (x_{20} is not significant; exclude x_{19} in advance)	0.9754
$\emptyset$	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{11} \to x_3$	0.9733	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{11} \to x_{22}$ (x_{22} is not significant; exclude x_{19} and x_{20} in advance)	0.9733
$\emptyset$	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{11} \to x_3 \to x_{14}$	0.9754	$x_{15} \to x_4 \to x_{10} \to x_{12} \to x_{18} \to x_{11} \to x_3 \to x_{23}$ (x_{23} is not significant; exclude x_{19} and x_{20} in advance)	0.9754

TABLE 4. Regression coefficients of Logit model

	Estimate	SE	t-statistics	p-value
const	−9.7738	2.0237	−4.8297	1.3674e-06
x_4	−6.1930	1.6787	−3.6891	2.2505e-04
x_{10}	−4.8665	1.1877	−4.0973	4.1800e-05
x_{12}	−7.0808	2.8940	−2.4468	0.0144
x_{15}	6.7298	1.3445	5.0056	5.5687e-07

Example 4.2. *End-stage renal disease dataset. The dataset is divided into two subsets (Group-A and Group-B) according to differences in patient enrollment timing. Group-A contains 1281 samples, while Group-B contains 504 samples. This dataset includes baseline characteristics and pre-hemodialysis laboratory values (measured at 3 months prior to treatment initiation) for all patients. Nine clinically significant variables (serum creatinine, age, albumin, hemoglobin, blood urea nitrogen, phosphorus, heart failure, diabetes mellitus and gender) have been identified from prior studies, which are directly used to analyze the survival status of patients after receiving hemodialysis treatment. To address class imbalance in datasets Group-A and Group-B, random under sampling is employed during group cross-validation, and performances are evaluated across multiple classification models.*

Table 5 displays classification performance for 5-year and 3-year survival endpoints, reporting mean AUC values (with variances in parentheses) from 20 under sampling trials. The comparative analysis reveals that classification models achieve superior performance

when processing granulation-reconstructed data. Empirical results from Logit, Probit, decision tree, and random forest models consistently demonstrate that granular data transformation enables more accurate diagnostic classification. Notably, even some machine learning methods (e.g., decision trees and random forests) show limited classification efficacy when applied to the original dataset. The proposed information granulation and data reconstruction techniques demonstrate the capability to extract representative and discriminative features from limited sample sets, thereby offering an effective solution for addressing diagnostic challenges in medical datasets.

TABLE 5. Comparison of classification results among different methods

Data	Logit model		Probit model		Decision tree		Random forest	
	Unprocessed	Granular reconstruction	Unprocessed	Granular reconstruction	Unprocessed	Granular reconstruction	Unprocessed	Granular reconstruction
Group-A (5 years)	0.5952 (0.0002)	0.6662 (0.0001)	0.5937 (0.0004)	0.6560 (0.0003)	0.5293 (0.0003)	0.9439 (0.0001)	0.5303 (0.0005)	0.9261 (0.0002)
Group-B (3 years)	0.7501 (0.0005)	0.8703 (0.0002)	0.7460 (0.0002)	0.8560 (0.0003)	0.5668 (0.0018)	0.9370 (0.0014)	0.5680 (0.0011)	0.9065 (0.0015)

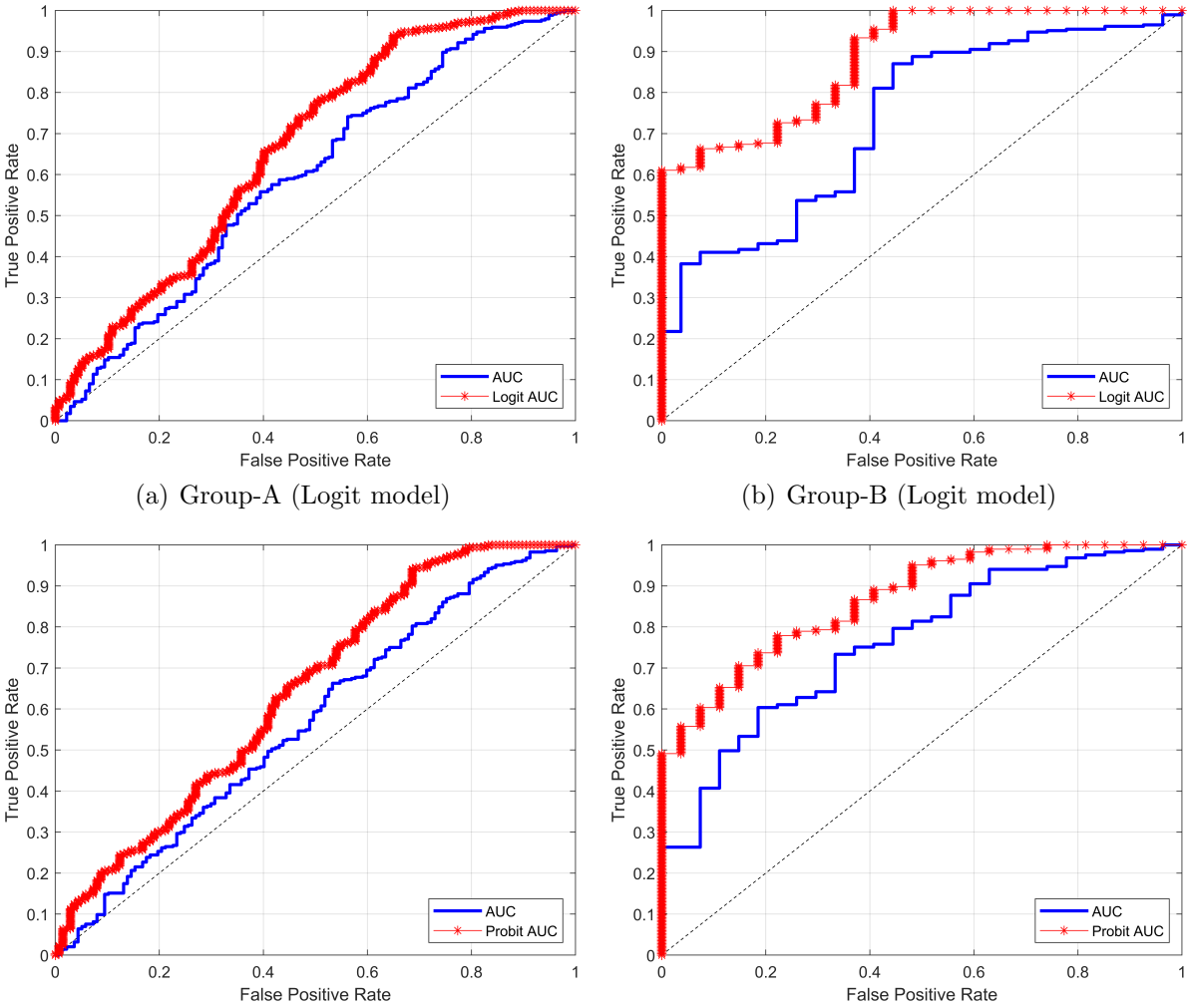
In addition, based on Logit model, Table 6 reveals significant heterogeneity (all $p < 0.05$ by likelihood ratio test) across subgroups stratified by heart failure, diabetes status, and sex. These findings substantiate that the proposed feature selection algorithm not only maintains high consistency with internationally recognized kidney disease diagnostic guidelines, but also identifies the clinically significant biomarkers that physicians predominantly rely on during diagnostic decision-making.

TABLE 6. Heterogeneity analysis likelihood ratio test p-value

Data	Heart failure	Diabetes mellitus	Gender
Group-A (5 years)	0.0158	0.0272	0.0405
Group-B (3 years)	0.0098	0.0313	0.0039

Figure 1 presents the ROC curves of the Logit model and the Probit model under different data processing methods. The solid lines represent the modeling results using the original (non-granulated) data, while the lines with asterisks denote the results obtained from the granulation-reconstructed data. For datasets Group-A and Group-B, the Logit model demonstrates robust classification performance without reconstruction, achieving an AUC value above 0.75 for Group B. However, data reconstruction significantly enhances classification accuracy and yields improved discriminative power for quality-of-life assessment.

The investigation of information granulation and data reconstruction's influence on feature selection and representation is conducted through an extension of existing granular computing frameworks [13,21,22], incorporating two computational formulations: triangular fuzzy sets and Gaussian fuzzy sets. Following data reconstruction with these granular representations, comparative evaluations of classification model performance are conducted across both datasets. In the initial processing of raw data, three clustering algorithms are employed for cluster center determination, including fuzzy C-means, K-means, and K-medoids. The identified cluster centers are then utilized as the core parameters for the triangular fuzzy sets (defining endpoints) and Gaussian fuzzy sets (determining variance). Subsequently, the membership degrees of sample data relative to these fuzzy set-based information granules are computed to enable data reconstruction. Table 7 presents the mean AUC values (with variances in parentheses) averaged over 10 classification runs. The consistently high classification accuracy observed across varying granular representations in Table 7 demonstrates the robustness of the proposed feature selection methodology.



(a) Group-A (Logit model) (b) Group-B (Logit model)

(c) Group-A (Probit model) (d) Group-B (Probit model)

FIGURE 1. Comparison of ROC curves

TABLE 7. Comparison of results based on different information granular methods

Data	Interval			Triangular fuzzy set			Gaussian fuzzy set		
	FCM	K-means	K-medoids	FCM	K-means	K-medoids	FCM	K-means	K-medoids
Group-A	0.6520 (0.0007)	0.7648 (0.0002)	**0.8264** (0.0002)	0.6982 (0.0006)	0.7382 (0.0007)	**0.7454** (0.0000)	**0.6835** (0.0001)	0.6656 (0.0001)	0.6453 (0.0001)
Group-B	0.8643 (0.0003)	**0.9385** (0.0001)	0.9160 (0.0002)	0.8884 (0.0005)	0.9282 (0.0001)	**0.9500** (0.0002)	**0.8247** (0.0001)	0.8037 (0.0001)	0.7962 (0.0001)

5. **Conclusions.** This study proposes an optimized framework for information granulation and data reconstruction in clinical datasets, with specific applicability to renal disease research. A novel indicator is designed to evaluate the collective role of information granules in feature expression, which is subsequently combined with classification accuracy to construct a cost function. Additionally, a progressive feature selection algorithm is introduced, demonstrating robust performance even under limited sample sizes while successfully identifying clinically interpretable features aligned with diagnostic expertise. Future research directions will focus on extending the application of the proposed method to high-dimensional datasets, with particular emphasis on developing adaptive

information granulation techniques and dynamic feature representation frameworks for evolving data environments.

Acknowledgment. This work is partially supported by the Humanities and Social Science project for Ministry of Education of China (21YJA630079). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] C. Foote, P. A. Clayton, D. W. Johnson et al., Impact of estimated GFR reporting on late referral rates and practice patterns for end-stage kidney disease patients: A multilevel logistic regression analysis using the Australia and New Zealand dialysis and transplant registry (ANZDATA), *American Journal of Kidney Diseases*, vol.64, no.3, pp.359-366, 2014.

[2] J. Zhang, F. T. Fuhrer, H. Ye et al., High-throughput metabolomics and diabetic kidney disease progression: Evidence from the chronic renal insufficiency (CRIC) study, *American Journal of Nephrology*, vol.53, pp.215-225, 2022.

[3] J. K. N. Chan, M. Solmi, C. U. Correll et al., Predicting 10-year risk of chronic kidney disease in lithium-treated patients with bipolar disorder: A risk model development and internal cross-validation study, *European Neuropsychopharmacology*, vol.95, pp.24-30, 2025.

[4] R. S. G. Sealfon, L. H. Mariani, M. Kretzler et al., Machine learning, the kidney, and genotype-phenotype analysis, *Kidney International*, vol.97, pp.1141-1149, 2020.

[5] T. Chen, X. Li, Y. Li et al., Prediction and risk stratification of kidney outcomes in IgA nephropathy, *American Journal of Kidney Diseases*, vol.74, no.3, pp.300-309, 2019.

[6] K. Ramu, S. Patthi, Y. N. Prajapati et al., Hybrid CNN-SVM model for enhanced early detection of chronic kidney disease, *Biomedical Signal Processing and Control*, vol.100, 107084, 2025.

[7] Q. Xie, Q. H. Zhang, S. Y. Xia et al., GBG++: A fast and stable granular ball generation method for classification, *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol.8, no.2, pp.2022-2036, 2024.

[8] B. Yu, X. He and W. P. Ding, A classification method based on dominance neighborhood granularity, *Knowledge-Based Systems*, vol.316, 113276, 2025.

[9] B. Yu, X. He and J. H. Dai, CLIG: A classification method based on bidirectional layer information granularity, *Information Sciences*, vol.650, 119662, 2023.

[10] D. Wang, M. Li and W. Song, A kind of information granular fuzzy broad learning system based on the Takagi-Sugeno model, *International Journal of Innovative Computing, Information and Control*, vol.19, no.1, pp.279-287, 2023.

[11] Y. Cui, H. E, W. Pedrycz et al., From fuzzy rule-based models to granular models, *IEEE Transactions on Fuzzy Systems*, vol.33, no.2, pp.644-656, 2025.

[12] J. C. Bezdek, R. Ehrlich and W. Full, FCM: The fuzzy C-means clustering algorithm, *Computers & Geosciences*, vol.10, pp.191-203, 1984.

[13] W. Pedrycz and W. Homenda, Building the fundamentals of granular computing: A principle of justifiable granularity, *Applied Soft Computing*, vol.13, pp.4209-4218, 2013.

[14] X. Zhu, W. Pedrycz and Z. Li, Granular representation of data: A design of families of ε-information granules, *IEEE Transactions on Fuzzy Systems*, vol.26, no.4, pp.2107-2119, 2018.

[15] W. Pedrycz, Granular data compression and representation, *IEEE Transactions on Fuzzy Systems*, vol.31, no.5, pp.1497-1505, 2023.

[16] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no.2, pp.224-227, 1979.

[17] F. Ros, R. Riad and S. Guillaume, PDBI: A partitioning Davies-Bouldin index for clustering evaluation, *Neurocomputing*, vol.528, pp.178-199, 2023.

[18] M. C. Barbier, B. I. Grisci and M. Dorn, Analysis and comparison of feature selection methods towards performance and stability, *Expert Systems with Applications*, vol.249, 123667, 2024.

[19] P. E. Stevens, S. B. Ahmed, J. J. Carrero et al., KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease, *Kidney International*, vol.105, no.4, pp.S117-S314, 2024.

[20] P. A. Moreno-Sanchez, Features importance to improve interpretability of chronic kidney disease early diagnosis, *IEEE International Conference on Big Data*, pp.3786-3792, 2020.

[21] W. Pedrycz and X. Wang, Designing fuzzy sets with the use of the parametric principle of justifiable granularity, *IEEE Transactions on Fuzzy Systems*, vol.24, no.2, pp.489-496, 2015.
[22] D. G. Wang, H. Liu, W. Pedrycz et al., Design Gaussian information granule based on the principle of justifiable granularity: A multi-dimensional perspective, *Expert Systems with Applications*, vol.197, 116763, 2022.

Author Biography

Wenyan Song received the B.Sc. degree in Mathematics from Beijing Normal University, China, in 2002; the Ph.D. degree in Applied Mathematics from Beijing Normal University, China, in 2007.

Dr. Song is currently a full-time associate professor at the School of Economics, Dongbei University of Finance and Economics, China. Her research interests include fuzzy system modelling, neural network and management optimization.

Xiangtai Meng received his B.Sc. degree in Financial Engineering from Linyi University, China, in 2019. He is currently pursuing his Master's degree at the School of Economics, Dongbei University of Finance and Economics, China. His research interests include population aging and machine learning.