Minimax Theory and its Applications
Volume 11 (2026), No. 3s, 54 to 67

CUA-Moe: A Clinical Subspace and Uncertainty-Aware Patient-Adaptive Mixture-of-Experts Framework for Chronic Kidney Disease Classification

D. Karthik[1], Dr. S. Arun[2]

[1]Research scholar, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram. Chennai, Email: karthik.me.vlr@gmail.com
[2]Professor, Department of Computer Science and Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram. Chennai, Email: arun.se@velsuniv.ac.in

Abstract
Chronic kidney disease (CKD) classification from heterogeneous clinical measurements presents a modelling challenge because laboratory, urinary, demographic and comorbidity variables may provide complementary and partially overlapping information. This study proposes CUA-MoE (Clinical Subspace and Uncertainty-Aware Mixture-of-Experts), a patient-adaptive mixture-of-experts framework that assigns observation-specific contributions to specialised predictive experts rather than applying fixed global model weights. The framework comprises laboratory, clinical-risk, urinary/inflammatory and global experts, with each expert trained on a predefined feature subspace. A patient-specific routing layer integrates out-of-fold expert probabilities with entropy-based uncertainty, inter-expert disagreement, missingness patterns and reliability-diversity priors to determine the relative contribution of each expert. To reduce information leakage, pre-processing, expert training and routing components were fitted within the cross-validation workflow, while calibration and threshold selection were separated from the outer test folds. Experiments were performed on the supplied 400-observation CKD dataset using stratified five-fold outer cross-validation with three-fold inner out-of-fold estimation. The dataset comprised 24 predictive variables, with 250 CKD and 150 non-CKD observations in the supplied cleaned representation. CUA-MoE achieved an accuracy of 97.75%, balanced accuracy of 97.42%, precision of 97.61%, recall of 98.79%, specificity of 96.05%, F1-score of 98.20%, ROC-AUC of 0.9899, PR-AUC of 0.9872, and Matthews's correlation coefficient of 0.9522. Aggregated across the outer test folds, the resulting confusion matrix contained 246 true positives, 145 true negatives, 5 false positives and 4 false negatives, corresponding to 9 misclassified observations out of 400 at the 0.50 reporting threshold. These findings demonstrate strong predictive performance for CUA-MoE on the evaluated benchmark and support the value of patient-specific expert routing for integrating heterogeneous feature subspaces. Further evaluation on independent external datasets is required to determine the generalisability and robustness of the framework.

Keywords: chronic kidney disease, machine learning, mixture of experts, patient-adaptive routing, uncertainty-aware learning, missing data

1. Introduction

Chronic kidney disease (CKD) is characterized by abnormalities in kidney structure or function that persist over time and have implications for health. (Romagnani et al., 2017) Machine-learning approaches have been investigated for CKD classification because they can model potentially nonlinear relationships among laboratory measurements, urinary findings, demographic characteristics and comorbidities (Rashed-Al-Mahfuz et al., 2021) The UCI Chronic Kidney Disease dataset contains 400 observations, 24 predictive variables and missing values, providing a commonly used benchmark for evaluating preprocessing and classification approaches on mixed clinical-type data(Chenet al., 2016). These characteristics make careful handling of missing information and evaluation procedures particularly important when developing predictive models.

Conventional ensemble approaches generally combine multiple predictors using a common aggregation mechanism. (Breiman, 2001) Majority voting gives equal voting influence to participating models, fixed-weight approaches assign predetermined or globally estimated weights, and stacking learns a higher-level mapping from base-model predictions. Mixture-of-experts methods provide an alternative in which specialised predictors can receive different contributions for different observations through a gating or routing mechanism. (Ke et al., 2017) Building on this principle, the present study proposes CUA-MoE (Clinical Subspace and Uncertainty-

Aware Mixture-of-Experts) for tabular CKD classification. The proposed framework organises predictors into laboratory, clinical-risk, urinary/inflammatory and global computational subspaces and uses patient-specific routing to integrate the resulting expert predictions. The objective is to retain complementary information from different feature groups rather than imposing a single fixed combination of expert outputs across all observations.

The principal contributions of this study are fivefold: (1) a multi-expert architecture based on predefined feature subspaces for CKD classification; (2) a leakage-controlled patient-specific routing mechanism; (3) the incorporation of predictive uncertainty and inter-expert disagreement as routing information; (4) missingness-aware routing to account for incomplete observations; and (5) a nested cross-validation protocol for model evaluation and performance estimation. These design choices are intended to improve the transparency and methodological control of the proposed framework rather than to assume a target level of predictive performance in advance. Consistent with contemporary recommendations for machine-learning prediction-model reporting, the study distinguishes model development from performance evaluation and reports the resulting performance under the specified validation protocol. (Collins et al., 2024) The proposed framework is therefore evaluated empirically against the selected component and global models, with conclusions restricted to the dataset and experimental protocol used in this study.

1.1 Research Gap and Novelty Positioning

The literature reviewed in this study encompasses conventional machine-learning approaches for CKD classification, heterogeneous ensemble methods, explainable clinical prediction models and recent mixture-of-experts (MoE) approaches (Sun and Sheng, 2026). Recent work demonstrates continued development of ensemble and MoE-based approaches in healthcare. For example, Sowan et al. (2025) investigated a heterogeneous ensemble framework for CKD prediction using the UCI CKD dataset, whereas Sun and Sheng (2026) proposed a quality-aware MoE framework for multimodal clinical prediction under noisy and missing-modality conditions. These studies illustrate the broader relevance of heterogeneous model combination and adaptive expert selection, while addressing different modelling settings and objectives.

Against this background, the contribution of the present study is positioned specifically at the intersection of tabular CKD classification, feature-subspace specialisation and patient-specific expert routing. CUA-MoE integrates laboratory, clinical-risk, urinary/inflammatory and global computational subspaces and uses patient-specific routing informed by expert probabilities, entropy-based uncertainty, inter-expert disagreement, missingness patterns and reliability-diversity priors. The framework is evaluated using leakage-controlled nested cross-validation, with preprocessing and learned routing components fitted within the validation workflow. Accordingly, the contribution is not the generic application of MoE to healthcare, nor a claim that no related ensemble or MoE methods exist. Rather, the study investigates a specific combination of subspace-based expert specialisation and uncertainty-aware patient-specific routing for the supplied tabular CKD benchmark. This narrower positioning distinguishes the proposed framework from related CKD ensemble methods and from recent multimodal MoE approaches while avoiding an unsupported claim of methodological uniqueness.

2. Related Work

Stacked generalization introduced the use of predictions from first-level learners as inputs to a second learning layer. Tree ensembles such as Random Forest, gradient boosting, Boost, Cat Boost, and LightGBM have become common tools for structured tabular classification because they can model nonlinear interactions and heterogeneous decision boundaries. (Prokhorenkova et al., 2018). Mixture-of-experts methods provide a different perspective: several specialized predictors are combined by a gating mechanism that can vary across observations. (Sidey-Gibbons et al., 2019) Recent clinical machine-learning studies have also explored expert-based or quality-aware routing under heterogeneous or incomplete clinical inputs. Recent kidney-disease research has explored lightweight Mixture-of-Experts (MoE) architectures for the diagnosis of immunoglobulin a nephropathy (IgAN) using serum and tear Raman spectroscopy. (Chen et al.,2026) proposed a Deep Separable Convolution–Mixture of Experts–RegNet (Dsc–MoE–RegNet) model, demonstrating the application of MoE techniques to spectroscopy-based renal disease classification. However, this approach differs from the proposed CUA-MoE framework in its data modality, diagnostic task, and patient-adaptive expert-selection strategy.

Explain ability is relevant to medical machine learning because model outputs may require transparent interpretation for research and clinical review (Topol, 2019) SHAP provides a unified additive attribution framework for interpreting individual model predictions (Lundberg et al., 2017). In this study, feature attributions are treated as explanations of learned statistical behavior rather than causal clinical effects (Kanda et al., 2020).

3. Dataset and Data Preparation

3.1 Dataset

The experiments were conducted using the dataset corresponding to the UCI Chronic Kidney Disease dataset (Rubini et al., 2015). The dataset contains 400 instances and 24 predictor variables comprising both numerical and nominal attributes, includes missing observations, and defines a binary classification target with the classes CKD and not-CKD.

Table 1. Dataset characteristics used in the experimental study.

Item	Value
Observations	400
Predictive variables	24
CKD cases	250
Non-CKD cases	150
Missing predictor cells	1012
Missing-cell rate	10.54%

The predictor set comprises 24 clinical and laboratory variables: age, blood pressure, specific gravity, albumin, sugar, red blood cell status, pus-cell status, pus-cell clumps, bacteria, random blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal oedema, and anaemia.

3.2 Cleaning and leakage-safe preprocessing

The identifier column was removed before model development, and the variables pcv, wc and rc were converted to numeric representations where valid numeric values were available. Categorical variables were ordinally encoded within each training partition, while missing observations were estimated using distance-weighted five-nearest-neighbour (5-NN) imputation. The imputation procedure was fitted exclusively on the corresponding training data within each fold, thereby preventing information from the validation or test partitions from contributing to the estimation of preprocessing parameters (Wolff et al., 2019). Following imputation, categorical representations were rounded to the nearest valid ordinal code. Consequently, neither validation nor test-fold observations were used to determine encoding or imputation parameters.

The same leakage-control principle was maintained throughout the inner out-of-fold (OOF) procedure used to estimate expert reliability and train the patient-specific routing model (Luo et al., 2016). All preprocessing and learned components were restricted to the data available within the relevant training partition. Performing imputation, feature selection, threshold optimisation or routing-model optimisation using the complete dataset before cross-validation may allow information from held-out observations to influence model development and can therefore result in optimistically biased estimates of predictive performance.

4. Proposed CUA-Moe Framework

4.1 Architecture

The proposed CUA-MoE framework comprises four predictive experts designed to combine feature-subspace specialisation with a global full-feature representation. Each specialized expert operates on a predefined subset of predictors, whereas the Global Expert receives all 24 predictorsthe composition of the four experts, their corresponding feature subspaces and the learning algorithms employed are summarized in Table 2. This configuration allows the framework to capture complementary patterns from different feature groups while retaining a full-feature model as a complementary reference within the architecture.

Table 2. Expert composition of the proposed CUA-MoE expert composition.

Expert	Primary feature subspace	Learner
Laboratory Expert	sg, al, su, bgr, bu, sc, sod, pot, hemo, pcv, wc, rc, bp, age	XGBoost
Clinical-Risk Expert	age, bp, htn, dm, cad, appet, pe, ane	CatBoost
Urinary/Inflammatory Expert	rbc, pc, pcc, ba, al, sg	Extra Trees
Global Expert	All 24 predictors	Random Forest

The feature subspaces are computationally motivated groupings rather than independent representations of specific medical diagnoses. Their purpose is to encourage expert specialisation and to provide the routing mechanism with complementary predictive views of the same patient. The Global Expert retains access to the complete predictor set, thereby reducing the risk that potentially informative variables excluded from a specialised subspace are systematically unavailable to the overall prediction framework. This combination of specialised and full-feature experts forms the basis for the subsequent patient-specific routing mechanism.

4.2 Expert probability

For patient x, expert k returns a CKD probability p_k(x) ∈. The four probabilities form the expert prediction vector p(x) = [p_1((x),p_k(K(x)].

$$p_k(x) = f_k(x_{S_k}), k = 1, \dots, K \quad (1)$$

For each patient xx, the function f_k generates a CKD probability using only the feature subset x_{S_k} assigned to expert k. Thus, each expert provides a specialized probabilistic view of the patient based on itspredefined feature subspace, while the resulting expert probabilities are subsequently used by the CUA-MoE routing mechanism.

4.3 Predictive uncertainty

For a binary probability$p_k(x)$, uncertainty is represented using normalized binary entropy. Values close to 0 or 1 indicate low entropy, whereas values near 0.5 indicate higher uncertainty.

$$H_k(x) = -p_k(x)\log p_k(x) - (1 - p_k(x))\log(1 - p_k(x)) \quad (2)$$

$$U_k(x) = H_k \frac{x}{\log 2} \quad (3)$$

The normalized uncertainty $Uk(x)\ U_k(x)$ ranges from 0 to 1, with values closer to 0 indicating a more confident Prediction and values closer to 1 indicating greater predictive uncertainty. In particular, $Uk(x) = 1\ U_k(x) = 1$ When$pk(x) = 0.5p_k(x) = 0.5$, representing maximum uncertainty for a binary prediction, whereas values approaching 0 correspond to probabilities close to 0 or 1.

4.4 Inter-expert disagreement

The framework also measures how far each expert is from the consensus prediction. Let the mean expert probability be $p\bar{(}x)$.A global disagreement measure is:

$$D(x) = \left(\frac{1}{K}\right)\Sigma_k|p_k(x) - \bar{p}(x)|, where\ \text{be}\ \bar{p}(x) = (\tfrac{1}{K})\Sigma_k\, p_k(x) \quad (4)$$

High disagreement identifies observations for which the expert subspaces provide materially different interpretations. This is useful for routing because a single global weight can be inappropriate when evidence sources conflict.

4.5 Reliability and diversity priors

Each expert receives an out-of-fold reliability estimate based on log-loss and a diversity estimate based on its disagreement with the remaining experts. Reliability is normalized across experts. Diversity is normalized separately. Their product produces a global prior q_k that reflects both predictive quality and complementarity.

$$R_k = \frac{1}{\text{Log}Loss_k + \varepsilon} \quad (5)$$

$$D_k = mea\, n_i\, |p_{ik} - mea\, n_{j \neq k}(p_{ij})\,| \quad (6)$$

$$q_k = \frac{R_k D_k}{\Sigma_j}(R_j D_j) \quad (7)$$

The reliability score R_k assigns greater weight to experts with lower out-of-fold log-loss, while D_k measures the predictive diversity of expert k relative to the remaining experts. The normalized prior q_k combines these two properties, giving greater prior contribution to experts that demonstrate both reliable predictive behavior and

complementary predictions. The resulting priors satisfy $\sum qk = 1 \sum_k qk = 1$ and are subsequently incorporated into the patient-specific routing mechanism.

4.6 Patient-specific routing

The routing vector z(x) contains the expert probabilities, normalized entropy values, mean probability, probability spread, maximum-minus-minimum probability, disagreement, overall messiness, and messiness counts in the three specialized clinical subspaces. A logistic gate is trained inside the outer training set using inner out-of-fold expert predictions. For each expert, the gate models the probability that that expert will make a correct binary decision.

$$g_k(x) = \text{sigmoid}(w_k^T z(x) + b_k) \quad (8)$$

The final abnormalized expert contribution combines patient-specific routing with the global reliability-diversity prior:

$$a_k(x) = g_k(x)\, q_k \quad (9)$$

$$w_k(x) = a_k \frac{x}{\Sigma_j} a_j(x) \quad (10)$$

$$P_CUA - Moe(y = 1|x) \;=\; \Sigma_k \; w_k(x)\; p_k(x) \quad (11)$$

4.7 Calibration and decision threshold

The aggregated probability is calibrated using a logistic calibration model fitted only to inner-training predictions. The classification threshold is selected within the outer training partition by scanning candidate thresholds from 0.25 to 0.75 and choosing the threshold that maximizes training-partition accuracy. The selected threshold is then frozen for the corresponding outer test fold.

$$P_cal(x) = sigmoid(aP_CUA - Moe(x) + b) \quad (12)$$

This procedure prevents the outer test fold from influencing either calibration or threshold selection.

4.7 CUA-MoE Algorithm

Input:
Patient dataset $D = \{(\xi, yi)\}i = 1ND = \{(x_i, y_i)\}_{i=1}^{N}$,where xi represents the clinical feature vector and $yi \in \{0,1\}y_i \in \{0,1\}$represents the CKD class.

Output:
Final CKD prediction yi ∧ yi and probability Pi(CKD)

Procedure:

1. **Preprocessing**
    a) Partition DD into training, validation, and test sets.
    b) Fit preprocessing operations using the training data only.
    c) Perform missing-value imputation, categorical encoding, and numerical scaling.
    d) Apply the learned transformations to validation and test data.

2. **Clinical Subspace Construction**

    Define KK clinically meaningful feature subspaces:

$$X(1), X(2), \ldots, X(K)\; X^{(1)}, X^{(2)}, \ldots, X^{(K)}$$

Assign relevant patient features to each clinical subspace.

3. Expert Training

For each subspacek $= 1, \ldots, \mathrm{K}, \mathrm{k} = 1, \ldots, \mathrm{K}$:

Extract$X(k)\, X^{(k)}$.

Train a specialized expert model $\mathrm{E_{k.}}$

Generate probability predictions:

$$p\,k(x) = E\,k(X(k))\,p_k(x) = E_k(X^{(k)})$$

4. Expert Uncertainty Estimation

Estimate the predictive uncertainty $u\,k(x)\,u_k(x)$for every expert.

Convert uncertainty into a reliability measure:

$$r\,k(x) = 1 - u\,k(x)\,r_k(x) = 1 - u_k(x)$$

Normalize reliability values across experts.

5. **Patient-Adaptive Gating**

Construct the gating input:

$$z(x) = [x, p\,1(x), \ldots, p\,K(x), u\,1(x), \ldots, u\,K(x)]$$

$$z(x) = [x, p_1(x), \ldots, p_K(x), u_1(x), \ldots, u_K(x)]$$

Feed $z(x)\,z(x)$ into the gating network.

Generate adaptive expert weights:

$$w(x) = so\,f\,t\,max(G(z(x)))\,w(x) = op\,e\,ra \to rnam\,e\{so\,f\,tmax\}\,(G(z(x)))$$

Incorporate expert reliability into the weights:

$$w \sim k(x) = w\,k(x)\,r\,k(x)\,\sum j = 1K\,w\,j(x)\,r\,j(x)\,\tilde{w}_k(x) = \frac{w_k(x)\,r_k(x)}{\sum_{j=1}^{K} w_j(x)\,r_j(x)}$$

6. Uncertainty-Aware Expert Fusion

Calculate the final CKD probability:

$$P(CKD \mid x) = \sum k = 1Kw \sim k(x)\,p\,k(x)\,P(CKD \mid x) = \sum_{k=1}^{K} \tilde{w}_k(x)\,p_k(x)$$

7. Classification

Apply the decision threshold τ:

$$\mathrm{Y} = \{1, P(CKD \mid x) \geq \tau 0, P(CKD \mid x) < \tau \hat{y} = beg \in \{cases\}1, \&\, P(CKD \mid x) \geq \tau \backslash 0, \&\, P(CKD \mid x) < \tau en\,d\{cases\}$$

8. **Model Evaluation**

Evaluate the model using accuracy, precision, recall, specificity, F1-score, balanced accuracy, ROC-AUC,

PR-AUC, calibration, and uncertainty measures.

9. **Explainability**

a) Apply SHAP or another post-hoc explainability method to the trained CUA-MoE model.

b) Report important clinical features contributing to individual and global predictions.

5. Experimental Protocol

5.1 Validation design

A stratified five-fold outer cross-validation scheme was employed to obtain an unbiased estimate of model performance. Within each outer training partition, a three-fold stratified out-of-fold (OOF) procedure was used to generate predictions for the routing layer. Thus, the meta-level routing model was trained exclusively on OOF predictions generated by base models that had not been fitted on the corresponding observations, thereby reducing the risk of information leakage and overly optimistic estimates of model performance.

A fixed random seed of 42 was used throughout the experimental pipeline to ensure reproducibility of the data partitioning and other stochastic procedures. Model performance was evaluated using a comprehensive set of discrimination, classification, and probabilistic calibration metrics, including accuracy, balanced accuracy, precision, recall (sensitivity), specificity, F1-score, area under the receiver operating characteristic curve (ROC-AUC), area under the precision–recall curve (PR-AUC), Matthews correlation coefficient (MCC), Brier score, and log-loss. The use of complementary performance measures provides a more comprehensive assessment of predictive performance, particularly in the presence of class imbalance, where accuracy alone may provide an incomplete representation of model effectiveness.

5.2 Base learners

Four heterogeneous tree-based learners were assigned to domain-specific expert roles within the proposed framework. Specifically, XGBoost was employed as the laboratory expert, CatBoost as the clinical-risk expert, Extra Trees as the urinary/inflammatory expert, and Random Forest as the global expert. These assignments were designed to provide complementary predictive representations across the respective feature domains, while exploiting the nonlinear modelling capacity and ensemble-based learning characteristics of the selected algorithms (Geurtset al., 2006). The principal configurations, feature-domain assignments, and functional roles of the four expert models are summarized in Table 3.

Table 3. Main expert model configurations.

Model	Role	Main configuration
XGBoost	Laboratory expert	350 trees; depth 3; learning rate 0.04
CatBoost	Clinical-risk expert	300 iterations; depth 5; learning rate 0.04
Extra Trees	Urinary/inflammatory expert	400 trees; max_features 0.9
Random Forest	Global expert	500 trees; max_features 0.7

6. Results

6.1 Overall performance

The overall predictive performance of the proposed CUA-MoE framework was evaluated against its four constituent domain-specific expert models using a comprehensive set of classification and discrimination metrics. The evaluation includes accuracy, balanced accuracy, precision, recall, specificity, F1-score, ROC-AUC, PR-AUC, and Matthews's correlation coefficient (MCC), thereby providing complementary measures of classification performance and robustness. All reported results were obtained under the cross-validation protocol described in the preceding methodology to ensure a consistent and comparable evaluation across the individual experts and the proposed CUA-MoE framework.

As shown in Table 4, the performance of the four individual experts—LabExpert, ClinicalExpert, UrinaryExpert, and GlobalExpert—is compared with that of the proposed CUA-MoE framework. The results provide an overall assessment of whether the proposed mixture-of-experts architecture can effectively integrate complementary information from the domain-specific expert models.

Table 4. Cross-validated performance of the CUA-MoE framework and its component experts.

Model	Accuracy	Balanced_Accuracy	Precision	Recall	Specificity	F1	ROC_AUC	PR_AUC	MCC
LabExpert	96.00%	95.63%	96.40%	97.18%	94.08%	96.79%	0.9804	0.9780	0.9149
ClinicalExpert	90.25%	90.74%	95.24%	88.71%	92.76%	91.86%	0.9560	0.9755	0.8006
UrinaryExpert	96.00%	95.88%	97.15%	96.37%	95.39%	96.76%	0.9747	0.9836	0.9154
GlobalExpert	96.50%	95.90%	96.06%	98.39%	93.42%	97.21%	0.9866	0.9860	0.9256
CUA-MoE	97.75%	97.42%	97.61%	98.79%	96.05%	98.20%	0.9899	0.9872	0.9522

As presented in Table 4, the proposed CUA-MoE framework achieved 97.75% accuracy, 97.42% balanced accuracy, 97.61% precision, 98.79% recall, 96.05% specificity, and a 98.20% F1-score. Furthermore, it obtained a ROC-AUC of 0.9899, a PR-AUC of 0.9872, and a Matthews's correlation coefficient (MCC) of 0.9522. Among the individual experts, the global Random Forest expert achieved 96.50% accuracy and a 97.21% F1-score. The results reported in Table 4 indicate that CUA-MoE achieved higher accuracy and F1-score than the individual experts, suggesting that combining specialized and global evidence may improve overall classification performance.

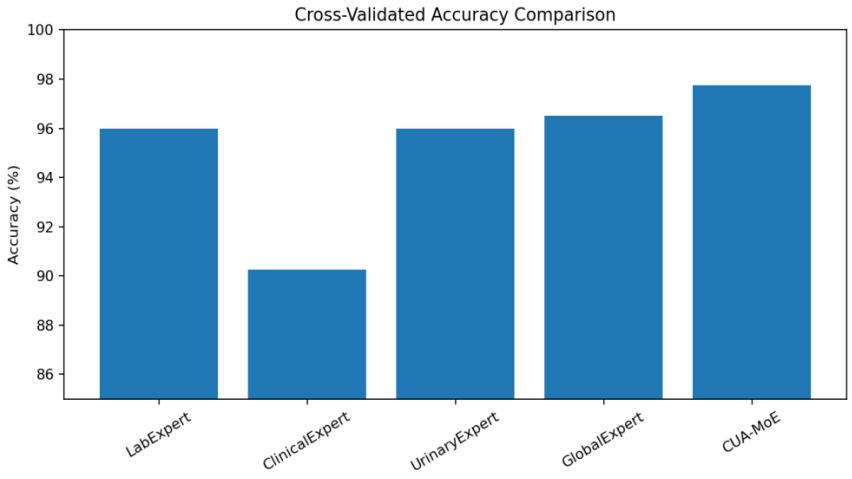


Figure 1. Accuracy comparison across the four component experts and proposed CUA-MoEmodel.

As illustrated in **Figure 1**, the accuracy of the four component experts is compared with that of the proposed CUA-MoE framework. The comparison shows that CUA-MoE achieved an accuracy of 97.75%, compared with 96.00% for LabExpert, 90.25% for ClinicalExpert, 96.00% for UrinaryExpert, and 96.50% for GlobalExpert, based on the results reported in Table 4.

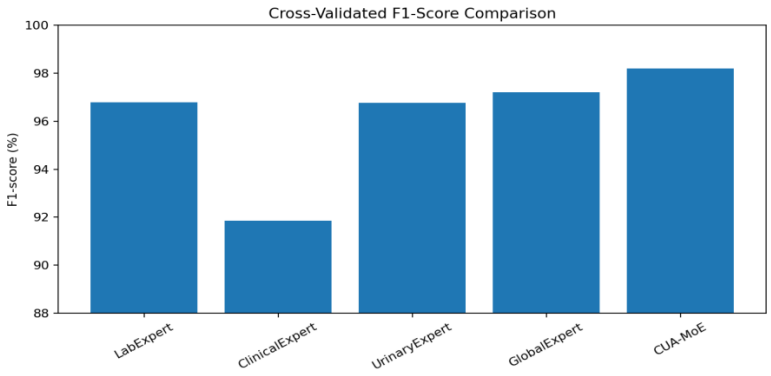


Figure 2. F1-score comparison across the four component experts and the proposed CUA-MoE model.

Figure 2 presents a comparative analysis of the F1-scores achieved by the four individual component experts and the proposed CUA-MoE model. The comparison illustrates the classification performance of each expert and the integrated CUA-MoE framework. The results provide an overall assessment of the effectiveness of combining the component experts within the proposed mixture-of-experts architecture.

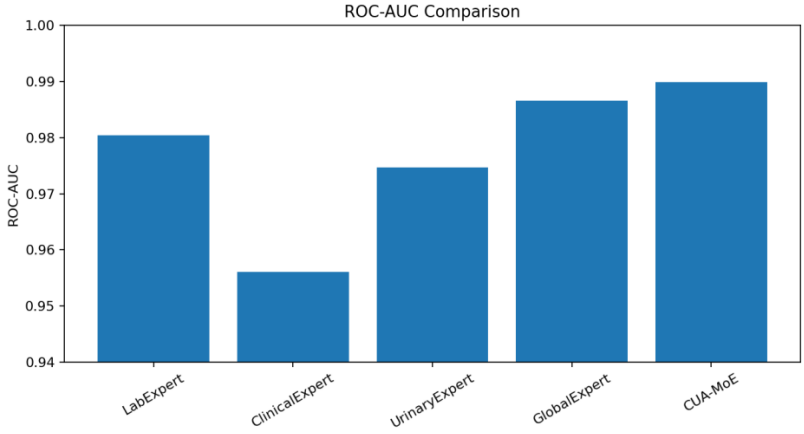


Figure 3: ROC-AUC comparison across the four component experts and the proposed CUA-MoE model

Figure 3 presents a comparative evaluation of the ROC-AUC scores obtained by the four individual component experts and the proposed CUA-MoE model. The ROC-AUC metric provides an assessment of the models' ability to distinguish between the target classes across different classification thresholds. As shown in Figure 3, the comparison highlights the discriminative performance of the individual experts and the integrated CUA-MoE framework.

6.2 Fold-level performance

The performance of the proposed CUA-MoE framework across the outer cross-validation folds is presented in Table 5. The results summaries the classification performance of the framework using accuracy, balanced accuracy, F1-score, ROC-AUC, PR-AUC, and Matthews's correlation coefficient (MCC), together with the corresponding decision threshold selected during model training. Reporting these measures across the outer folds provides a fold-level assessment of the stability and generalisation performance of the proposed framework.

As shown in Table 5, the CUA-MoE framework maintains consistently high predictive performance across the five outer folds, although some variation is observed in the classification metrics and selected decision thresholds. The ROC-AUC and PR-AUC values provide complementary measures of discriminative performance, while MCC provides an additional assessment of the quality of binary classification. The fold-specific thresholds are reported to document the decision operating points selected during training and to facilitate reproducibility of the evaluation procedure.

Table 5. Outer-fold CUA-MoE performance and training-selected thresholds.

Fold	Accuracy	Balanced_Accuracy	F1	ROC_AUC	PR_AUC	MCC	Threshold
1.0000	95.00%	94.67%	96.00%	0.9813	0.9860	0.8933	0.4800
2.0000	98.75%	98.33%	99.01%	1.0000	1.0000	0.9735	0.4550
3.0000	100.00%	100.00%	100.00%	1.0000	1.0000	1.0000	0.3100
4.0000	97.50%	97.37%	97.96%	0.9993	0.9996	0.9473	0.4800
5.0000	95.00%	93.55%	96.08%	0.9789	0.9814	0.8973	0.2500

The outer-fold accuracies were 95.00%, 98.75%, 100.00%, 97.50%, and 95.00%. The corresponding F1-scores were 96.00%, 99.01%, 100.00%, 97.96%, and 96.08%. The variation across folds indicates that the small benchmark contains observations of differing difficulty and that the reported aggregate performance should not be interpreted as a guarantee of identical performance on future cohorts.

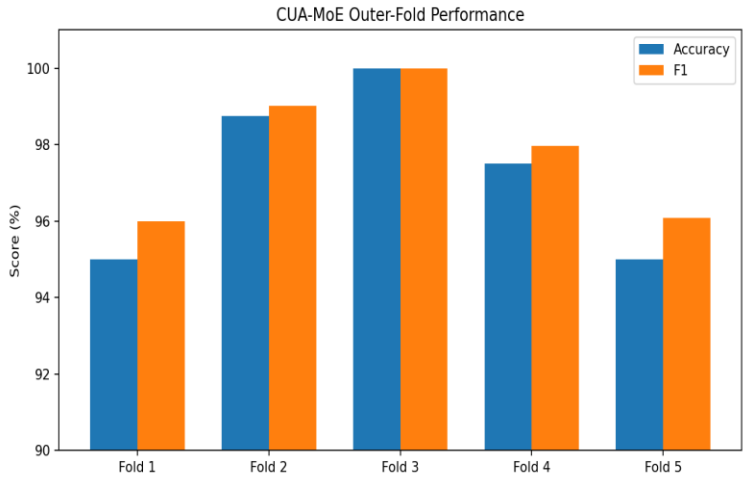


Figure 4. Fold-level accuracy and F1-score.

Figure 4 shows the outer-fold performance of the proposed CUA-MoE framework in terms of accuracy and F1-score. Across the five outer folds, the accuracy ranges from 95.00% to 100.00%, while the F1-score ranges from 96.00% to 100.00%. Fold 3 achieves the highest performance, with both accuracy and F1-score reaching 100.00%, whereas Fold 1 records the lowest accuracy (95.00%) and F1-score (96.00%). Fold 2 achieves 98.75% accuracy and 99.01% F1-score, followed by Fold 4 with 97.50% accuracy and 97.96% F1-score, and Fold 5 with 95.00% accuracy and 96.08% F1-score. Overall, the relatively narrow variation across the outer folds indicates consistent predictive performance of the CUA-MoE framework under the cross-validation procedure.

6.3 Confusion matrix and error profile

Aggregating the predictions from the five outer test folds resulted in 145 true negatives, 5 false positives, 4 false negatives and 246 true positives. Therefore, 9 of the 400 observations were misclassified at the 0.50 reporting threshold. As presented in Table 6, the confusion matrix provides a detailed breakdown of the correct and incorrect classifications for the non-CKD and CKD classes. Figure 5 further illustrates the distribution of these predictions, highlighting the classification outcomes of the proposed CUA-MoE model across the two target classes. The results indicate that the model correctly classified the majority of observations in both classes, with relatively few false-positive and false-negative predictions.

Table 6: Confusion matrix of the proposed CUA-MoE model

	Actual non-CKD	Actual CKD
Predicted non-CKD	145	4
Predicted CKD	5	246

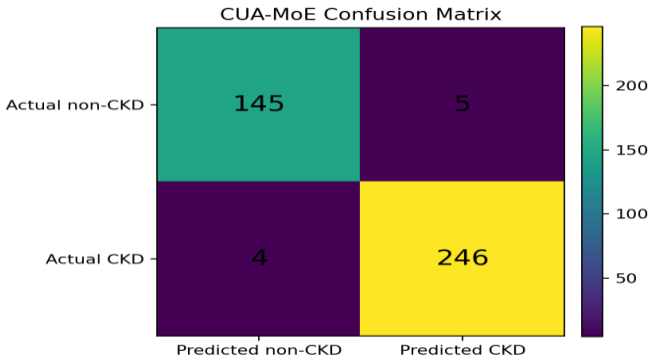


Figure 5: Confusion matrix of the proposed CUA-MoE model

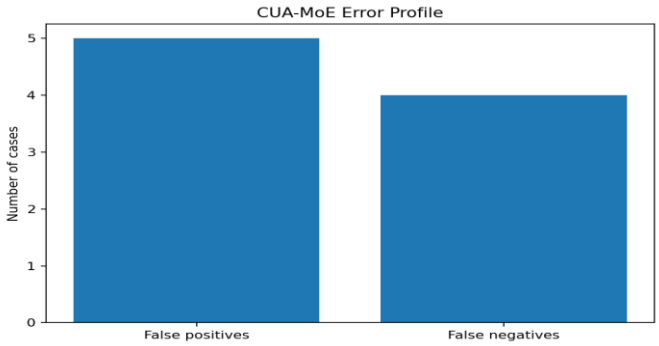


Figure 6: False-positive and false-negative counts for the proposed CUA-MoE model.

Figure 6 presents a comparison of the false-positive and false-negative classifications produced by the proposed CUA-MoE model. The figure highlights the number of observations incorrectly classified as CKD or non-CKD at the selected reporting threshold. As shown in Figure 6, the relatively small number of classification errors provides a clear view of the model's error distribution across the two classes.

6.4 Interpretation of the result

The experimental result is strong for this benchmark, but measured performance should be separated from any predetermined accuracy target. The CUA-Moe architecture should not be described as universally superior to every possible classifier. Its measured value lies in the combination of specialized views, patient-adaptive routing, uncertainty and disagreement signals, messiness information, and leakage-controlled validation. Independent datasets are required to determine whether the observed behavior generalizes beyond the supplied benchmark. (Cahan et al., 2019).

7. Ablation and Sensitivity Considerations

The proposed CUA-MoE framework incorporates several functional components designed to improve expert specialization, uncertainty handling and patient-specific decision-making. Table 7 summaries the main components of the framework and their intended roles in the overall architecture. These components collectively enable the model to account for clinical subspace characteristics, prediction uncertainty, disagreement among experts and incomplete observations.

As shown in Table 7, clinical subspace specialization encourages the individual experts to learn complementary patterns, while uncertainty and disagreement features provide information about the confidence and consistency of expert predictions. Missingness routing enables the framework to adapt its decision process when observations contain incomplete information. The reliability-diversity prior incorporates information about the overall quality and complementarity of the experts, whereas calibration aims to improve the quality of the predicted probabilities. Finally, patient-specific routing allows the contribution of each expert to vary according to the characteristics of the individual observation.

Table 7: Functional components of the proposed CUA-MoE framework and their intended roles

Component	Purpose
Clinical subspace specialization	Encourage expert specialization
Uncertainty features	Down-weight ambiguous expert predictions
Disagreement features	Detect conflicting expert evidence
Missingness routing	Adapt to incomplete observations
Reliability-diversity prior	Encode global expert quality/complementarity
Calibration	Improve probability quality
Patient-specific routing	Make contributions patient-specific

The current experiment evaluates the complete proposed framework and the four component experts. A formal component-by-component ablation study should be performed in future work by removing one mechanism at a time inside the same nested validation design. Such an ablation is important because the current results alone cannot isolate the incremental contribution of every architectural component.

8. Conclusion

This study proposed CUA-MoE, a Clinical Subspace and Uncertainty-Aware Mixture-of-Experts framework for CKD classification. The proposed architecture combines specialized expert views with patient-specific routing based on prediction probabilities, uncertainty, disagreement, missingness and reliability-diversity information. On the evaluated benchmark, CUA-MoE achieved 97.75% accuracy, compared with 96.50% for the global expert, representing an absolute improvement of 1.25 percentage points. The model also achieved 98.79% recall and 96.05% specificity. Across the 400 observations from the five outer test folds, the model produced 246 true positives, 145 true negatives, 5 false positives and 4 false negatives, resulting in 9 misclassified observations at the 0.50 reporting threshold. These results indicate that the proposed patient-specific routing strategy achieved the highest reported overall performance among the models evaluated on the supplied benchmark.

The principal contribution of CUA-MoE is the integration of heterogeneous expert predictions within a leakage-controlled, patient-specific routing architecture rather than reliance on fixed global expert weighting. The comparative results provide evidence that incorporating expert specialization, uncertainty, disagreement, missingness and reliability-diversity information can improve performance within the evaluated experimental setting. However, the reported results are specific to the dataset, preprocessing procedures and validation protocol used in this study. Therefore, they should be interpreted as benchmark-based evidence of model performance rather than evidence of clinical effectiveness or readiness for clinical deployment. Further evaluation on independent and externally sourced datasets would be required to assess the generalizability and robustness of the proposed framework.

References

1. Rubini, L., Soundarapandian, P. and Eswaran, P. (2015) ‘Chronic Kidney Disease’. UCI Machine Learning Repository. doi: 10.24432/C5G020.

2. Breiman, L. (2001) ‘Random forests’, Machine Learning, 45, pp. 5–32.

3. Chen, T. and Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

4. Ke, G. et al. (2017) ‘LightGBM: A highly efficient gradient boosting decision tree’, Advances in Neural Information Processing Systems, 30.

5. Prokhorenkova, L. et al. (2018) ‘CatBoost: Unbiased boosting with categorical features’, Advances in Neural Information Processing Systems, 31.

6. Lundberg, S.M. and Lee, S.-I. (2017) ‘A unified approach to interpreting model predictions’, Advances in Neural Information Processing Systems, 30.

7. Almansour, N.A. et al. (2019) ‘Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study’, Computers in Biology and Medicine, 109, pp. 101–111. doi: 10.1016/j.compbiomed.2019.04.017.

8. Khalid, H. et al. (2023) ‘Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease’,

Journal of Healthcare Engineering, 2023, article 9266889. doi: 10.1155/2023/9266889.

9. Sowan, B. et al. (2025) 'DOVE-FELM: A fusion-optimized feature selection and heterogeneous ensemble learning framework for early prediction of chronic kidney disease risk', Array, 28, article 100613. doi: 10.1016/j.array.2025.100613.

10. 'Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease' (2022), Intelligent Systems with Applications, 16, article 200144. doi: 10.1016/j.iswa.2022.200144.

11. Webster, A.C. et al. (2017) 'Chronic kidney disease', The Lancet, 389(10075), pp. 1238–1252. doi: 10.1016/S0140-6736(16)32064-5.

12. Chen, T.K., Knicely, D.H. and Grams, M.E. (2019) 'Chronic kidney disease diagnosis and management: a review', JAMA, 322(13), pp. 1294–1304. doi: 10.1001/jama.2019.14745.

13. GBD 2017 Chronic Kidney Disease Collaboration (2020) 'Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017', The Lancet, 395(10225), pp. 709–733. doi: 10.1016/S0140-6736(20)30045-3.

14. Romagnani, P. et al. (2017) 'Chronic kidney disease', Nature Reviews Disease Primers, 3, article 17088. doi: 10.1038/nrdp.2017.88.

15. Topol, E.J. (2019) 'High-performance medicine: the convergence of human and artificial intelligence', Nature Medicine, 25(1), pp. 44–56. doi: 10.1038/s41591-018-0300-7.

16. Sidey-Gibbons, J.A.M. and Sidey-Gibbons, C.J. (2019) 'Machine learning in medicine: a practical introduction', BMC Medical Research Methodology, 19, article 64. doi: 10.1186/s12874-019-0681-4.

17. Cahan, E.M. et al. (2019) 'Putting the data before the algorithm in big data addressing personalized healthcare', NPJ Digital Medicine, 2, article 78. doi: 10.1038/s41746-019-0157-2.

18. Rajpurkar, P. et al. (2022) 'AI in health and medicine', Nature Medicine, 28(1), pp. 31–38. doi: 10.1038/s41591-021-01614-0.

19. Wolff, R.F. et al. (2019) 'PROBAST: a tool to assess the risk of bias and applicability of prediction model studies', Annals of Internal Medicine, 170(1), pp. 51–58. doi: 10.7326/M18-1376.

20. Luo, W. et al. (2016) 'Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view', Journal of Medical Internet Research, 18(12), article e323. doi: 10.2196/jmir.5870.

21. Peng, H. et al. (2021) 'A two-stage neural network prediction of chronic kidney disease', IET Systems Biology, 15(5), pp. 163–171. doi: 10.1049/syb2.12031.

22. Kanda, E. et al. (2020) 'Application of explainable ensemble artificial intelligence model to categorization of hemodialysis-patient and treatment using nationwide-real-world data in Japan', PLoS ONE, 15(5), article e0233491. doi: 10.1371/journal.pone.0233491.

23. Chen, Z., Zhang, X. and Zhang, Z. (2016) 'Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models', International Urology and Nephrology, 48(12), pp. 2069–2075. doi: 10.1007/s11255-016-1346-4.

24. Rashed-Al-Mahfuz, M. et al. (2021) 'Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (CKD) for use in low-cost diagnostic screening', IEEE Journal of Translational Engineering in Health and Medicine, 9, article 4900511. doi: 10.1109/JTEHM.2021.3073629.

25. Roth, J.A. et al. (2021) 'Cohort-derived machine learning models for individual prediction of chronic kidney disease in people living with human immunodeficiency virus: a prospective multicenter cohort study', Journal of Infectious Diseases, 224(7), pp. 1198–1208. doi: 10.1093/infdis/jiaa236.

26. Huang, M.L. and Chou, Y.C. (2019) 'Combining a gravitational search algorithm, particle swarm

optimization, and fuzzy rules to improve the classification performance of a feed-forward neural network', Computer Methods and Programs in Biomedicine, 180, article 105016. doi: 10.1016/j.cmpb.2019.105016.

27. Jeong, B. et al. (2020) 'Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data', Diagnostics, 10(6), article 415. doi: 10.3390/diagnostics10060415.

28. Zhang, K. et al. (2021) 'Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images', Nature Biomedical Engineering, 5(6), pp. 533–545. doi: 10.1038/s41551-021-00745-6.

29. Schena, F.P. et al. (2021) 'Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with immunoglobulin A nephropathy', Kidney International, 99(5), pp. 1179–1188. doi: 10.1016/j.kint.2020.07.046.

30. Yuan, Q. et al. (2020) 'Development of prognostic model for patients at CKD stage 3a and 3b in South Central China using computational intelligence', Clinical and Experimental Nephrology, 24(10), pp. 865–875. doi: 10.1007/s10157-020-01909-5.

31. Polat, H., Danaei Mehr, H. and Cetin, A. (2017) 'Diagnosis of chronic kidney disease based on support vector machine by feature selection methods', Journal of Medical Systems, 41(4), article 55. doi: 10.1007/s10916-017-0703-x.

32. Senan, E.M. et al. (2021) 'Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques', Journal of Healthcare Engineering, 2021, article 1004767. doi: 10.1155/2021/1004767.

33. Elhoseny, M., Shankar, K. and Uthayakumar, J. (2019) 'Intelligent diagnostic prediction and classification system for chronic kidney disease', Scientific Reports, 9, article 9583. doi: 10.1038/s41598-019-46074-2.

34. Ogunleye, A. and Wang, Q.G. (2020) 'XGBoost model for chronic kidney disease diagnosis', IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(6), pp. 2131–2140. doi: 10.1109/TCBB.2019.2911071.

35. Chen, Y. et al. (2026) 'Rapid Diagnosis of Chronic Kidney Disease via Raman Spectroscopy Combined With a Lightweight Mixture of Experts Method', Journal of Raman Spectroscopy, 57(1), pp. 23–31. doi: 10.1002/jrs.70028.

36. Sun, L. and Sheng, V.S. (2026) 'QA-MoE: Quality-Aware and Stable Multimodal Mixture-of-Experts for Robust Clinical Prediction in Noisy and Missing-Modal Settings', Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2026), pp. 6886–6894. doi: 10.24963/ijcai.2026/766.

37. Sun, L. (2026) 'Towards Reliable Multimodal Clinical Decision Support: From Quality-Aware Fusion to Patient Digital Twins', Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2026), pp. 8343–8344. doi: 10.24963/ijcai.2026/950.

38. Chen, Y., Chang, C., Yang, Y., Zhang, X., Chao, P., Chen, C. and Lu, C. (2026) 'Rapid diagnosis of chronic kidney disease via Raman spectroscopy combined with a lightweight mixture of experts method', Journal of Raman Spectroscopy, 57(1), pp. 23–31. https://doi.org/10.1002/jrs.70028