

Article

HFDE–CKD: A Hyper-Fidelity Dynamic Ensemble Framework for Chronic Kidney Disease Diagnosis

Mahmoud Hassaballah ^{1,*}  and Mohamed Abdel Hameed ² ¹ Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, AlKharj 16278, Saudi Arabia² Department of Computer Science, Faculty of Computers and Information, Luxor University, Luxor 85951, Egypt; m.abdelhamid@fci.luxor.edu.eg

* Correspondence: mah.ali@psau.edu.sa

Abstract

Chronic kidney disease (CKD) is a global health disease that must be recognized and monitored early to prevent major consequences and improve the health of patients, yet modern gradient boosting algorithms often produce dangerous false positives due to overconfidence and static ensemble weighting. To address this, we propose the Hyper-Fidelity Dynamic Ensemble framework for CKD diagnosis (HFDE–CKD). Our approach utilizes a rigid quad-fold data isolation protocol and applies Standard SMOTE to the training manifold. By removing direct diagnostic identifiers (e.g., eGFR, creatinine, and albumin), the framework is forced to rely on underlying early indications. We quantify predictive doubt using Shannon entropy and inter-model variance, concatenating these measurements with calibrated probability vectors into a multidimensional uncertainty matrix optimized via a logistic meta-learner. Evaluated on an imbalanced dataset of 11,933 records, HFDE–CKD achieved an accuracy of 0.8596, a sensitivity of 0.8900, and an F1-score of 0.8979. Furthermore, HFDE–CKD’s superiority over the best-performing baseline was confirmed through McNemar’s test, which was statistically significant ($p < 0.001$). Emphasizing diagnostic safety, the model yielded a Brier score of 0.1012, an NLL of 0.3225, and a specificity of 0.7907, significantly reducing the risk of false positives. Furthermore, ablation testing confirmed the advantage of uncertainty injection, while SHAP analysis validated clinical dependency on early metabolic biomarkers. The proposed HFDE–CKD framework systematically leverages epistemic uncertainty to bypass the traditional sensitivity-versus-false-positive trade-off, providing a precisely calibrated, robust decision-support tool for early CKD screening.

Keywords: chronic kidney disease; hyper-fidelity; dynamic ensemble; heterogeneous learners; HFDE



Received: 4 June 2026

Revised: 9 July 2026

Accepted: 10 July 2026

Published: 30 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Chronic kidney disease (CKD) is a major public health issue. It is defined as a progressive loss of kidney function that may lead to renal failure if left untreated [1]. Early diagnosis of CKD is important to delay disease progression, avoid significant consequences, and minimize healthcare costs [2,3]. However, early diagnosis of mild, asymptomatic, and heterogeneous CKD remains difficult because its initial symptoms are difficult to detect [4]. Existing diagnostic techniques, such as laboratory testing and ultrasound, give only limited insight into kidney function and may easily overlook the nuanced changes that

signal the development of the illness [5,6]. If CKD is diagnosed early, clinicians have the opportunity to initiate treatment promptly to minimize disease progression, reduce patient discomfort, and improve survival. In this respect, knowledge-based systems may be of great use. These systems can mimic a clinician's cognitive processes, recognize risk factors, and help patients make timely treatment decisions by integrating clinical knowledge and patient-specific data [7]. Unlike traditional diagnostic tools [8], they can track patient data over time, recognize patterns, and recommend diagnostic or therapeutic measures even when nonspecific symptoms are present.

The increasing availability of advanced technology has made AI an important tool in diagnostics, especially for classifying disease. Knowledge-based expert systems are an important type of AI that stands out because they can accurately classify conditions and make their reasoning clear, which helps clinicians understand and support CKD diagnosis [9,10]. Machine learning (ML) algorithms have significantly improved prediction accuracy, allowing the prediction of disease development and the implementation of treatment measures. Despite this, conventional machine learning methodologies, such as support vector machines (SVMs) and decision trees, possess fundamental limitations. Their dependence on static feature-label relationships constrains their ability to perform automated feature selection, and they encounter difficulties handling the complexity of high-dimensional medical datasets [11]. These models are affected by overfitting, especially in cases where the data is limited or unbalanced. A significant weakness is their failure to accurately capture highly complex non-linear interactions among clinical indicators, which are essential for the early detection of CKD [12].

In recent years, deep learning (DL) architectures, such as Convolutional Neural Networks, have garnered interest because of their ability to extract complex features from high-dimensional biological data. In contrast to conventional ML techniques, DL techniques eliminate the need for extensive manual preprocessing and can reveal biologically significant patterns, maintaining robustness even with very small labeled datasets. Nevertheless, the use of deep learning in clinical practice is still limited because it needs greater computational resources and more extensive training datasets, thereby limiting its application in real-time healthcare environments and increasing the risk of data leakage and overfitting. Furthermore, its advanced feature learning does not consistently result in improved performance in relatively simple tasks, such as CKD detection [4,11,13].

CKD is a significant public health concern, as its early stages are often silent, leading to delayed diagnosis and increased risks of irreversible kidney damage [14,15]. Early detection is critical to slow disease progression, prevent complications, and reduce healthcare costs [3,16]. Common diagnostic tools, including routine laboratory tests and visual examinations, provide only a limited view of renal function and may not identify minor early changes [5,6]. These constraints can be addressed using knowledge-based systems and ML techniques that utilize multiple clinical parameters, monitor patient data continuously, identify risk factors, and provide prompt treatment [7,8]. All existing research is limited by accuracy and precision-based measurements. Although such biomarkers are useful, they may suffer from model confidence. A model might be fairly precise but not accurately calibrated, resulting in overconfident mistakes that are not acceptable in clinical practice. Moreover, medical datasets are always imbalanced, which causes traditional classifiers to prefer the majority class, raising the danger of missing important positive instances [17,18].

In particular, modern ML using explicit biomarkers like eGFR, serum creatinine, and BUN as predictive features can lead to significant target leakage in modern predictive frameworks that result in methodological errors in the early diagnosis of CKD [19,20]. On the other hand, gradient boosting ensembles achieving great classification accuracy tend to be overconfident in borderline clinical situations. Additionally, traditional static ensemble

approaches combine base learner outputs naively, ignoring both epistemic uncertainty and inter-model disagreement, which leads to weakness, large false-positive rates, and poor model calibration [21].

To overcome these clinical and computational limitations, we propose the Hyper-Fidelity Dynamic Ensemble framework for CKD diagnosis (HFDE–CKD). The proposed methodology introduces a clinical handicap by exchanging primary renal biomarkers, such as eGFR, urine-albumin, and serum creatinine, with secondary ones, such as age, phosphorus, bicarbonate, uric acid, and calcium, to discover hidden and early metabolic signals for CKD. To ensure numerical stability, min–max normalization is applied to isolate the most discriminative features using random forest out-of-bag (OOB) permutation errors as information gain [22]. This is followed by using probability vectors from heterogeneous base learners (MLP, XGBoost, LightGBM, and KNN) instead of static soft-voting to determine their epistemic contradiction. This data is used to train a logistic meta-learner to build a dynamic decision boundary capable of high-performance classification and probabilistic reliability for use as a proactive CKD screening tool.

The following is an outline of the proposed contributions and new features:

1. **Variance Feature Optimization:** We utilize a Random Forest OOB permutation method to generate a highly selective feature subspace, which minimizes noise and overfitting while keeping key signals.
2. **Target Leakage for Early Detection:** In order to reduce data leakage, we implement a novel clinical constraint by replacing fundamental GFR and their related biomarkers with secondary systemic indicators. Therefore, the proposed system is transformed from a repetitive diagnostic analyzer to an early-warning screening tool.
- Uncertainty-Injected Meta-Stacking:** We substituted random weighting methods and fixed soft-voting with a dynamic logistic meta-learner. This design employs predictive uncertainty to resolve inconsistent clinical scenarios by applying Shannon entropy and inter-model prediction variance as distinct characteristics.
3. **Probabilistic Calibration with Imbalance Awareness:** We develop a robust quad-fold data isolation approach combined with the Synthetic Minority Oversampling Technique (Standard SMOTE). The proposed HFDE–CKD enhances diagnostic safety through probability calibration and CKD-specific instance weighting, resulting in exceptional expected calibration errors (ECEs) and Brier scores while significantly minimizing false positives.
4. **End-to-End Clinical Interpretability:** We establish a comprehensive meta-stacking architecture using an experimental SHAP framework, demonstrating that HFDE’s diagnostic outcomes are affected by early metabolic indicators (e.g., age, phosphorus, calcium) rather than trivial diagnostic artifacts.

The rest of this paper is organized as follows: Section 2 reviews state-of-the-art methods, considering the diagnosis of CKD through numerous diverse methods. The proposed methodology is introduced in Section 3. Experimental results and related discussion are provided in Section 4. Finally, this paper concludes in Section 5.

2. Literature Review

Recent advances in kidney disease have shown the rising relevance of machine learning, deep learning, and optimization-based approaches in improving diagnostic accuracy, predicting outcomes, and assisting clinical decision-making.

In particular, machine learning models, including KNN, SVM, logistic regression, naive Bayes, random forest, and AdaBoost, were evaluated for reaching accuracy rates ranging from 74.1% to 99.1% for CKD prediction. The CNN-LSTM model attained and gained a high accuracy value of 99.1% in CKD diagnosis [23]. In Ref. [24], many ML classifiers were employed, revealing that gradient boosting achieved an accuracy of around 99.8%. In addition, earlier studies that are relevant to classical classifiers focus on high accuracy, precision, recall, and primary biomarkers as main indicators for CKD diagnosis [11,25]. Numerous studies have recognized intrinsic constraints in conventional machine learning methodologies with high accuracy [26–28]. Moreover, SVMs exhibit sensitivity to a dataset's size and have significant computational costs for large-scale issues, whereas random forests can suffer from overfitting as a result of feature estimation [9,29].

Deep learning techniques mitigate some challenges by naturally learning hierarchical features and diminishing dependence on manual feature engineering [11]. However, healthcare data frequently exhibit noise, high dimensionality, and imbalance, which hinder the learning process, especially for the early diagnosis of chronic kidney disease. In contrast, deep learning models show enhanced accuracy and scalability compared to ML methods [30–32]. In Ref. [33] they present challenges concerning computational costs, data preprocessing, and lower-level interpretability to address and highlight a trade-off between model complexity and clinical transparency.

Furthermore, optimization-based learning has attracted several research studies with respect to CKD. They have used Bayesian optimization with XGBoost and tested it against logistic regression, random forest, KNN, and SVM. They have shown that XGBoost has the greatest performance with an accuracy of 99.1% [34]. These models are useful to some level, but generally, their performance is limited when dealing with high-dimensional clinical data. An interpretation-based CKD prediction model is developed using the UCI dataset, and it applies min–max scaling, Z-score normalization, and Gini-based feature selection with primary biomarkers. Their method reached an accuracy of 98.4%, emphasizing the significance of interpretability in clinical situations [35]. A wide range of ML algorithms and ensemble approaches were used to explore the importance of performing class imbalance, and they found that rotation forest outperformed the other models [36].

Despite recently advancements, current ML and DL methods for CKD still have major drawbacks. Deep learning methods proficiently capture complex nonlinear correlations in medical data, which is fundamental for accurate diagnosis classification. Many proposed hybrid architectures are demonstrated by combining SVM with CNNs in order to establish a balance between computational efficiency and accuracy [37,38]. These models use the robust feature extraction of CNNs combined with the stable decision boundaries of SVMs, resulting in enhanced generalization across various datasets [39,40]. Nevertheless, hybrid models frequently prove to be computationally complex and challenging to scale, especially for complicated clinical datasets [41].

Specifically, three major limitations were present in many studies that achieved the diagnosis of CKD. Initially, target leakage was a significant issue that most existing frameworks could not handle or solve. They employed direct diagnostic indicators (e.g., eGFR and serum creatinine) as prediction variables. And this results in distribution leakage, which increased due to the insufficient imbalance of data, transforming the models into irrelevant diagnostic tools instead of early-warning systems [42]. Secondly, the majority of modern research is based on independent optimization or ensemble methodologies (e.g., fundamental soft or hard voting) that combine outcomes. These methodologies measure epistemic uncertainty, which leads to biased clinical situations. Third, the risk of false positives is increased, and the reliability of traditional models for safe clinical decision support

is limited as they frequently provide overconfident probabilities without evaluating the uncertainty of base learners [30,37].

Mainly, the proposed Hyper-Fidelity Dynamic Ensemble (HFDE) framework addresses the problem of trade-offs between discriminative capacity and probabilistic reliability by leveraging uncertainty meta-learning and dynamic probability calibrations. The HFDE–CKD architecture is provided via a multistage pipeline: imposing a high-level clinical handicap to prevent target leakage, converting clinical vectors into an optimized feature subspace, and producing a balanced manifold under data isolation. Ultimately, calibrated predictions from base learners are combined with epistemic uncertainty (Shannon entropy and variance). They are integrated through a logistic meta-learner, leading to an optimum decision threshold that maximizes both diagnostic sensitivity and F1 scores.

The proposed HFDE–CKD framework concurrently solves various issues, which has been mentioned earlier. It employs a rigid handicap constraint to prevent target leakage and takes advantage of random forest feature selection to identify early metabolic biomarkers. Then, the framework utilizes an uncertainty-injected meta-learner to dynamically integrate four heterogeneous base learners according to their epistemic uncertainty. At the end, the proposed HFDE–CKD framework addresses the methodological gap between isolated optimized models and static ensemble approaches, providing protection from data leakage and overconfidence for high-dimensional datasets. Also, it provides a highly reliable solution for the early screening and diagnosis of CKD.

3. The Proposed Methodology: HFDE Architecture

This section defines the primary phases of the Hyper-Fidelity Dynamic Ensemble (HFDE) architecture, as seen in Figure 1. The methodology employs a multistage procedure. Initially, to address target leakage, the main renal indicators (e.g., eGFR, serum creatinine, BUN, albumin) are removed from the raw dataset $\mathcal{D}$ to impose high constraints. Then, the data passes through a quad-fold isolation protocol, partitioning it into training ($\mathcal{D}_{train} = 60\%$), calibration validation ($\mathcal{D}_{val_calib} = 10\%$), evaluation validation ($\mathcal{D}_{val_eval} = 10\%$), and testing ($\mathcal{D}_{test} = 20\%$) sets to guarantee geometric isolation before the calibration phase. This is succeeded by numerical adjustments, including min–max normalization, which is exclusively based on $\mathcal{D}_{train}$ and subsequently applied to the next folds to maintain geometric isolation. Secondly, an optimum, low-dimensional feature subspace $\mathcal{X}_{opt}$ is derived by maximizing Gini impurity reduction using a non-parametric random forest assessment [43], which is important for mitigating the effect of multidimensional and decreasing variance. Third, the $\mathcal{D}_{train}$ set is subjected to SMOTE resampling [44] to ensure boundary separation in the presence of significant clinical class imbalance while preserving the validation geometry.

Subsequently, a highly heterogeneous set of base hypothesis functions $h_m \in \mathcal{M}$ —incorporating tree-based, neural, and distance-based topologies (KNN)—is trained across this optimized topological space. Fourth, we extract the calibrated predictions ($\hat{p}_m$) from all base estimators evaluated on $\mathcal{D}_{val_eval}$ and combine them with precise indicators of their uncertainty, namely, prediction variance (σ^2) and systemic Shannon entropy (H). Together, these metrics form a comprehensive, multidimensional feature space. Instead of relying on simple averaging, we feed this entire uncertainty matrix into the HFDE’s final manifold stacking, using ridge regularization combined with cost-sensitive CKD instance weighting to penalize false negatives for accurate predictions [45,46]. Unlike standard linear weighting, this uncertainty-injected meta-learner autonomously learns to suppress epistemic ambiguity and dynamically identify the failure modes of the individual base learners. Furthermore, the clinical decision boundary is established by locating the optimal threshold τ^* that maximizes the F1 score, ensuring maximal sensitivity for imbalanced

diagnostics [47,48]. Finally, the fully optimized architecture is evaluated on the independent test set $\mathcal{D}_{test}$ to generate final clinical classifications, $\hat{y}$. The probabilistic reliability of this uncertainty-injected meta-learner is rigorously validated against objective continuous penalty functions, including the Brier score, negative log-likelihood (NLL), and expected calibration error (ECE) [49,50]. All structural and functional mechanisms of this sequence are formalized in Algorithm 1.

Algorithm 1 HFDE: Hyper-Fidelity Dynamic Ensemble Algorithm

- 1: **Input:** Dataset $\mathcal{D}$, feature constraint K , four base learners $\mathcal{M} = \{\text{MLP, XGBoost, LightGBM, KNN}\}$.
 - 2: **Output:** Final classifications $\hat{y}$, calibrated probabilities $\hat{p}_{ens}$, optimal threshold τ^* , evaluation metrics.
 - 3: **Phase 1: Strict Data Isolation (Quad-Fold)**
 - 4: **Provide Clinical Handicap:** Remove GFR biomarkers (e.g., eGFR, creatinine, albumin) from $\mathcal{D}$.
 - 5: Partition dataset $\mathcal{D}$ into $\mathcal{D}_{train} = 60\%$, $\mathcal{D}_{val_calib} = 10\%$, $\mathcal{D}_{val_eval} = 10\%$ and $\mathcal{D}_{test} = 20\%$.
 - 6: Derive min-max scaling parameters on $\mathcal{D}_{train}$; apply for all splits.
 - 7: **Phase 2: Manifold Generation (Pre-Feature Selection)**
 - 8: Apply SMOTE resampling strictly to scaled $\mathcal{D}_{train}$ to achieve class balance.
 - 9: **Phase 3: Subspace Optimization**
 - 10: Compute Gini-impurity reduction on the *balanced* $\mathcal{D}_{train}$ to extract optimal subspace $\mathcal{X}_{opt}$ (top- K).
 - 11: Filter $\mathcal{D}_{train}$, $\mathcal{D}_{val_calib}$, $\mathcal{D}_{val_eval}$, and $\mathcal{D}_{test}$ to retain only $\mathcal{X}_{opt}$.
 - 12: **Phase 4: Base Model Training & Heterogeneous Calibration**
 - 13: **for** each learner $m \in \mathcal{M}$ **do**
 - 14: Train hypothesis function h_m on $\mathcal{D}_{train}(\mathcal{X}_{opt})$.
 - 15: Generate raw probability vectors p_{m, val_calib} .
 - 16: Fit continuous Platt Scaling (binomial GLM) on $(p_{m, val_calib}, y_{val_calib})$.
 - 17: Apply fitted calibration functions to generate $\hat{p}_{m, val_eval}$ and $\hat{p}_{m, test}$.
 - 18: **end for**
 - 19: **Phase 5: Uncertainty Injection & Manifold Stacking**
 - 20: Extract all base predictions: $\hat{p}_i = [\hat{p}_{1,i}, \dots, \hat{p}_{|\mathcal{M}|,i}]$
 - 21: Compute Shannon Entropy: $\mathcal{H}_i = -\sum \hat{p}_{m,i} \log(\hat{p}_{m,i})$
 - 22: Compute Prediction Variance: $\sigma_i^2 = \text{Var}(\hat{p}_i)$
 - 23: Form uncertainty meta-vector: $x_{meta,i} = [\hat{p}_i, \sigma_i^2, \mathcal{H}_i]$
 - 24: Compile meta-vectors into matrix $\mathbf{X}_{meta}$.
 - 25: Train Ridge-regularized Logistic Meta-Learner f_{meta} on $(\mathbf{X}_{meta}, y_{val_eval})$ utilizing CKD-focused instance weights.
 - 26: Generate final ensemble probabilities: $\hat{p}_{ens} = f_{meta}(\mathbf{X}_{meta, test})$.
 - 27: **Phase 6: Clinical Boundary Optimization**
 - 28: Locate τ^* maximizing F1-score on the independent $\hat{p}_{ens, val_eval}$ manifold.
 - 29: **for** each test instance $i \in \mathcal{D}_{test}$ **do**
 - 30: Project final diagnosis: $\hat{y}_i \leftarrow \mathbb{I}(\hat{p}_{ens,i} \geq \tau^*)$
 - 31: **end for**
 - 32: **Phase 7: Statistical Calibration Metrics**
 - 33: Compute Brier score, NLL, ECE, and McNemar's significance on $\mathcal{D}_{test}$.
 - 34: **Return** $\hat{y}$, $\hat{p}_{ens}$, τ^* , BS, NLL, ECE, Accuracy, Sensitivity, Specificity, F1-Score.
-

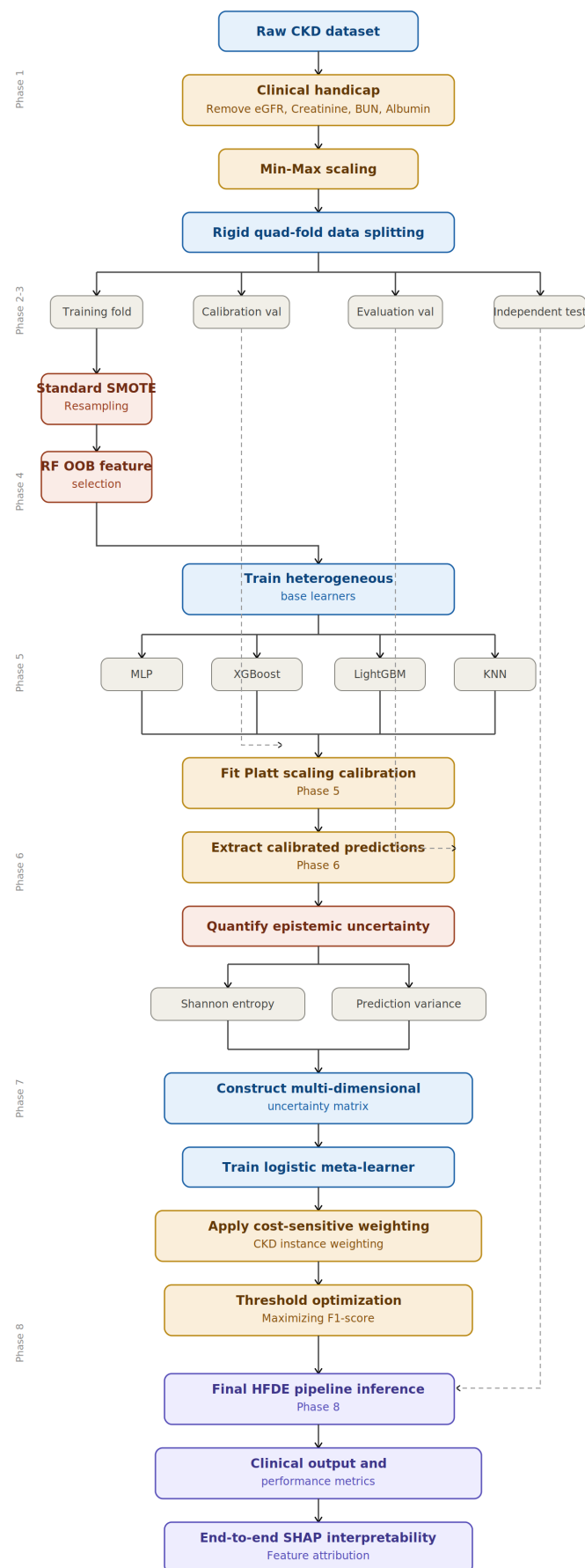


Figure 1. An overview of the proposed HFDE-CKD framework.

3.1. Data Isolation and Scaling Protocol

Data leakage is a primary cause of generalization failure in clinical machine learning prediction diseases, as it artificially impacts model performances [51]. To avoid this, the proposed HFDE-CKD imposes a clinical handicap by removing primary biomarkers (e.g., eGFR, serum creatinine, BUN, albumin) from the dataset so that the model learns early persistent indicators. Thus, the proposed HFDE-CKD initializes quad-fold data partitions: a training fold ($\mathcal{D}_{train}$), a calibration validation fold ($\mathcal{D}_{val_calib}$), a meta-stacking evaluation fold ($\mathcal{D}_{val_eval}$), and an independent testing fold ($\mathcal{D}_{test}$).

To achieve numerical stability, min-max normalization maps the continuous feature space to a narrow, limited interval $[0, 1]$. Crucially, the scaling parameters are locked and derived exclusively from the native $\mathcal{D}_{train}$ distribution. These locked parameters are subsequently applied to $\mathcal{D}_{val_calib}$, $\mathcal{D}_{val_eval}$, and $\mathcal{D}_{test}$, thereby completely preventing global distribution leakage as follows:

$$x_{scaled} = \frac{x - \min(\mathcal{D}_{train})}{\max(\mathcal{D}_{train}) - \min(\mathcal{D}_{train})} \quad (1)$$

3.2. Synthetic Manifold Generation Under Strict Isolation

Medical datasets inherently exhibit severe clinical class imbalance, causing conventional classifiers to mathematically bias toward the majority class (non-CKD). To resolve this, the proposed HFDE architecture intervenes in the latent data manifold using the Synthetic Minority Over-sampling Technique (SMOTE) [52]. After that, interpolation is implemented on the training set $\mathcal{D}_{train}$, while the validation and testing datasets remain in their unbalanced clinical conditions to maintain diagnostic validity and prevent data leakage. In the isolated $\mathcal{D}_{train}$ space, minority instances are represented, and vectors x_{new} are generated using interpolation along the line segments connecting k -nearest minority neighbors as follows:

$$x_{new} = x_i + \delta(x_{mn} - x_i), \quad \delta \sim \mathcal{U}(0, 1) \quad (2)$$

where x_i indicates a minority instance, and x_{mn} represents its k -nearest minority neighbors. Unlike hybrid resampling techniques that aggressively decimate majority class samples to artificially separate decision boundaries, HFDE relies exclusively on this standard SMOTE generation. This approach successfully balances the training density of the minority class while deliberately preserving the native, complex overlapping structures of the majority class, which often contain vital borderline clinical edge cases necessary for robust meta-learner generalization.

3.3. Subspace Optimization and Base Learner Training

Focusing on a compact set of meaningful clinical variables leads to more interpretable, robust, and accurate predictions, thereby supporting clinically decision-making for CKD patients. Accordingly, random-forest-based feature selection is adopted to preserve discriminative information while controlling a model's complexity. Let $\Delta G_j^{(t)}$ denote the reduction in Gini impurity contributed by clinical feature j at node split t . The global importance of feature j is illustrated as follows:

$$\mathcal{I}_j = \sum_{t=1}^T \Delta G_j^{(t)} \quad (3)$$

By defining a constraint to select only the top- K features that maximize $\sum_{j \in \mathcal{X}} \mathcal{I}_j$, the framework extracts a low-dimensional, highly discriminative subspace $\mathcal{X}_{opt}$. A highly heterogeneous set of probabilistic base learners, $m \in \mathcal{M} = \{\text{MLP}, \text{XGBoost}, \text{LightGBM}, \text{KNN}\}$, are then independently trained on $\mathcal{D}_{train}(\mathcal{X}_{opt})$. Incorporating distance-based (KNN) topologies alongside trees and neural networks reduces estimator variance, injects architec-

tural diversity, and mitigates overfitting for achieving high-dimensional medical datasets as follows:

$$\mathcal{X}_{\text{opt}} = \arg \max_{|\mathcal{X}|=K} \sum_{j \in \mathcal{X}} \mathcal{I}_j. \quad (4)$$

3.4. Heterogeneous Probability Calibration Routing

Standard ML architectures optimize for classification loss but frequently yield uncalibrated probability distributions. A critical vulnerability in modern ensembles is the application of global parametric calibration to diverse models, which do not produce normally distributed uncalibrated margins. HFDE solves this by establishing heterogeneous calibration routing constraints. The framework evaluates the architectural origin of the base learner to map its output through the mathematically appropriate calibration function. For models generating roughly Gaussian margin distributions (e.g., MLP, KNN), outputs are mapped through a continuous Platt scaling function by minimizing the negative log-likelihood on $\mathcal{D}_{\text{val}}$, where

$$\hat{p}_{m,i} = \frac{1}{1 + \exp(-(az_{m,i} + b))} \quad (5)$$

HFDE-CKD logically unifies the probability through a continuous scaling function for all base learners, ensuring that all prediction vectors are calibrated accurately and are similar before meta-integration.

3.5. Instance Dynamic Fusion and Manifold Stacking

The fundamental flaw of conventional heterogeneous ensembles is the assumption of uniform classifier competence across all local data manifolds. Deriving a static weighting vector from global metrics, such as PR-AUC, mathematically fixes ensemble parameters, preventing them from adapting to out-of-distribution (OOD) geometries at the patient level. The HFDE architecture breaks away from the limitations of standard static weighting or arbitrary pruning mechanisms. Instead, for any given patient i , we directly quantify the underlying epistemic uncertainty across all base learners using Shannon entropy ($\mathcal{H}$) and prediction variance (σ_i^2):

$$\mathcal{H}(\hat{p}_{m,i}) = -[\hat{p}_{m,i} \log(\hat{p}_{m,i}) + (1 - \hat{p}_{m,i}) \log(1 - \hat{p}_{m,i})] \quad (6)$$

$$\sigma_i^2 = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} (\hat{p}_{m,i} - \mu_i)^2 \quad (7)$$

where $|\mathcal{M}|$ represents the total number of base estimators, and μ_i is their mean predicted probability for instance i . Then, the proposed HFDE-CKD concatenates the calibrated predictions ($\hat{p}_{m,i}$) directly with the Shannon entropy ($\mathcal{H}$) and prediction variance (σ_i^2) to form a rich, multidimensional uncertainty matrix. This matrix is injected into a ridge-regularized logistic meta-learner. Cost-sensitive instance weighting is utilized to impose significant penalties on false negatives (missed CKD patients) and address clinical imbalance. So, the final stacking method exploits epistemic ambiguity and identifies multivariate failure scenarios that are systematically ignored.

3.6. Statistical Calibration Metrics

The proposed methodology assesses calibration quality by evaluating three metrics: Brier score (BS), negative log-likelihood (NLL), and expected calibration error (ECE) [53]. BS covers discrimination and calibration, with lower levels indicating superior probabilistic accuracy, hence targeting both overconfident and underconfident predictions. It measures

the average squared divergence between estimated probabilities and actual results, and it is defined as follows:

$$BS = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2, \quad (8)$$

where $\hat{p}_i \in [0, 1]$ signifies the calibrated ensemble probability for sample i , and $y_i \in \{0, 1\}$ represents the appropriate truth label. Furthermore, NLL assesses the probabilistic coherence of forecasts by imposing penalties on erroneous confident predictions. This characteristic makes NLL highly good at detecting overconfident errors, which are risky in clinical settings. NLL is defined as follows:

$$NLL = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{p}_i + \epsilon) + (1 - y_i) \log(1 - \hat{p}_i + \epsilon)], \quad (9)$$

where N represents the total number of samples. $y_i \in \{0, 1\}$ corresponds to the ground-truth label of the i -th instance. $\hat{p}_i \in [0, 1]$ indicates the predicted probability for the positive class, and $\epsilon = 10^{-15}$ is a small constant introduced for numerical stability to avert the evaluation of $\log(0)$. Moreover, ECE measures the divergence between predicted probability and actual results. A reduced ECE value signifies that the model's projected probabilities align more closely with empirical accuracy, indicating enhanced reliability and the confidence of probabilistic outputs. ECE is defined as follows:

$$ECE = \sum_{b=1}^B \frac{|B_b|}{N} |\text{acc}(B_b) - \text{conf}(B_b)|, \quad (10)$$

where B represents the total number of confidence bins that estimated probabilities throughout the interval $[0, 1]$. The set B_b comprises the samples for which their predicted probability resides inside the b -th bin, and $|B_b|$ indicates the total number of samples in that bin. The accuracy of samples in bin B_b is defined as follows:

$$\text{acc}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \mathbf{1}(y_i = \hat{y}_i), \quad (11)$$

where $\hat{y}_i = \mathbf{1}(\hat{p}_i \geq 0.5)$ is the predicted binary class label, and $\mathbf{1}(\cdot)$ is the indicator function. The mean predicted confidence for bin B_b is computed as follows:

$$\text{conf}(B_b) = \frac{1}{|B_b|} \sum_{i \in B_b} \hat{p}_i. \quad (12)$$

As a result, each calibration metric reflects a different aspect of probabilistic reliability. The BS examines overall probability accuracy, NLL penalizes overconfident wrong predictions. ECE measures how well expected confidence corresponds with observed accuracy [54]. The BS examines overall probability accuracies, and NLL penalizes overconfident wrong predictions. ECE measures how well expected confidence corresponds with observed accuracy [54]. Besides probabilistic calibration, understanding statistical significance is critical. To this end, the McNemar test is applied to examine the prediction accuracy of matched models. It is a non-parametric test applied to a 2×2 probability table of paired nominal data between two classifiers. The test statistic integrates Yates' continuity corrections. It is defined as follows [55]:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \quad (13)$$

where b represents the quantity of samples incorrectly assigned by the first model but correctly classified by the second. c represents the quantity of samples classified by the first model but misclassified by the second. So, the variance in performance is not related to random chance between two models.

3.7. F1-Manifold Threshold Optimization and Uncertainty

One suggested drawback of current methods is their dependence on fixed decision thresholds (e.g., $\tau = 0.5$). The proposed HFDE-CKD describes threshold selection as a restricted optimization problem. The framework provides a calibrated probability manifold to identify the appropriate threshold, τ^* , that optimizes precision and sensitivity [56]:

$$\tau^* = \arg \max_{\tau \in [0.05, 0.95]} \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Sensitivity}(\tau)}{\text{Precision}(\tau) + \text{Sensitivity}(\tau)} \quad (14)$$

3.8. Performance Evaluation Metrics

The following performance metrics were calculated based on a confusion matrix for the positive class ('ckd')[6]:

Accuracy:

$$\frac{(\text{TruePositive} + \text{TrueNegative})}{(\text{TruePositive} + \text{TrueNegative} + \text{FalsePositive} + \text{FalseNegative})} \quad (15)$$

Precision (Specificity):

$$\frac{\text{TruePositive}}{(\text{TruePositive} + \text{FalsePositive})} \quad (16)$$

Recall (Sensitivity):

$$\frac{\text{TruePositive}}{(\text{TruePositive} + \text{FalseNegative})} \quad (17)$$

F1 Score:

$$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (18)$$

The area under the ROC curve (AUC-ROC) quantifies the model's capacity to differentiate between classes at all classification levels. True Positive (TP): The model accurately forecasts the positive class. True Negative (TN): The model accurately identifies the negative class. False Positive (FP): The model incorrectly predicts the positive class when the real class is negative. False Negative (FN): The model incorrectly predicts the negative class when the actual class is positive [57].

4. Experimental Results and Discussion

The experiments were executed using MATLAB2024b on a system equipped with a 16 GB of RAM and an Intel Core i9 10th Generation CPU. In this section, we conducted a comprehensive experimental analysis of all heterogeneous base learners (MLP, XGBoost, LightGBM, and KNN) using the proposed HFDE-CKD architecture to evaluate their effectiveness in both No-SMOTE and Standard-SMOTE contexts. The experimental investigation was conducted to evaluate the prediction robustness, calibration reliability, and decision stability of the proposed architecture. We offer a comprehensive assessment of performance that incorporates both conventional classification metrics and uncertainty-aware calibration methodologies. Additionally, gains were observed in the normal metrics and in the calibrated functions, including BS, NLL, and ECE. Furthermore, we evaluate and analyze statistical significance and ablation analysis to ensure reliability for CKD in the

early stages. Finally, we conduct a SHAP interpretability analysis to evaluate the clinical feature importance, which led to verifying the robustness and consistency of the proposed HFDE–CKD framework with respect to clinical knowledge of CKD.

4.1. Dataset Description and Clinical Features

The dataset utilized in this study was constructed for CKD risk classification using open-source data from the National Health and Nutrition Examination. Additionally, the dataset known as CKD-NHANES 2021-2023 Kaggle [58] was used. Raw CDC post-pandemic releases can be problematic with missing variables, However, the dataset integrates data into 36 strong clinical and demographic characteristics, representing a complete population of 11,933 patient records. It was analyzed and pre-processed. Additionally, the dataset provides a highly reliable and consistent ground truth for ML evaluation by classifying the severity of kidney disease according to the most recent KDIGO recommendations 2024 and estimated glomerular filtration rate (eGFR). NHANES continues as an ongoing, nationally representative study that gathers cross-sectional data on the health and nutritional status of the U.S. population. The primary features and attributes extracted from this dataset are categorized and detailed in Table 1. To ensure clinical validity in our proposed HFDE–CKD early-warning framework, these variables are divided into secondary systemic biomarkers (used for predictive modeling) and primary diagnostic proxies (which are utilized for formal KDIGO staging but explicitly excluded from training to prevent target leakage).

Table 1. Description of the clinical features and attributes of the CKD-NHANES 2021–2023 dataset.

Category	Feature/Attribute	Data Type	Description & Clinical Relevance
Demographics	Age (Years)	Float64/Int64	Key risk factor for age-related renal decline.
	Gender	Object/Int64	Patient biological (male/female).
	Race/Ethnicity	Object/Int64	Patient racial or ethnic background.
	Socioeconomic Status	Float64	Measured by the income-to-poverty ratio.
Clinical & Vitals	Weight (kg) & BMI	Float64	Indicators for obesity-related kidney stress.
	Systolic Blood Pressure	Float64	Primary marker of hypertension.
	Diastolic Blood Pressure	Float64	Marker of vascular resistance.
	Pulse Rate	Float64	Resting heart beats per minute.
Secondary/Metabolic Biomarkers	Phosphorus	Float64	Early systemic marker of metabolic kidney dysfunction.
	Calcium	Float64	Tracks with phosphorus imbalance in early CKD.
	Bicarbonate	Float64	Low levels may indicate early metabolic acidosis.
	Uric Acid	Float64	High levels accelerate direct kidney damage.
	Fasting Blood Glucose	Float64	Primary indicator for diabetic kidney disease.
Primary Renal Proxies (Staging)	eGFR	Float64	The standard filtration metric for KDIGO staging.
	Serum Creatinine	Float64	Tracks muscle breakdown and filtration failure.
	Blood Urea Nitrogen (BUN)	Float64	Shows how well the kidneys excrete waste.
	Urine Albumin	Float64	Flags structural damage to kidney filters.
Target Variable	CKD Stage	Int64/Object	Official KDIGO stage (0-5) or binary CKD diagnosis.

In this section, we provide a baseline assessment on the No-SMOTE distribution to demonstrate a fundamental weakness in traditional gradient boosting ensembles. So,

the following subsections analyze and discuss the performance of the proposed HFDE–CKD framework using the No-SMOTE procedure.

4.2. Classification Metrics and Probabilistic Calibration

As illustrated in Table 2, the performance metrics with (Acc, Sens, Spec, and F1 Score) all heterogeneous base models (MLP, XGBoost, LightGBM, and KNN) and the proposed HFDE–CKD was evaluated on the No-SMOTE dataset. The outcomes indicate the significant weakness of gradient boosting methods in the presence of a clinical handicap (i.e., without eGFR, creatinine and albumin). Despite LightGBM’s and XGBoost’s ability to obtain high sensitivities of 0.9487 and 0.9427, respectively, they were at a low level of specificity: 0.5323 and 0.5777, respectively. Thus, these static base models suffered from probabilistic overconfidence, which leads to the production of a lot of false-positive alerts, resulting in significant alert loss in clinical situations. In contrast, the proposed HFDE–CKD framework achieved a trade-off dynamically using epistemic uncertainty. As mentioned in Table 2, the proposed HFDE–CKD obtained the maximum overall accuracy value of 0.8537 and a peak F1 score of 0.8931. Specifically, the uncertainty-injected meta-learner enhanced diagnostic safety, increasing specificity up to 0.7950 while maintaining a high sensitivity of 0.8794. Moreover, the proposed framework also showed its probabilistic superiority, with the lowest Brier score of 0.1006 and a very competitive NLL of 0.3117. This illustrates that HFDE–CKD is not just guessing by measuring predicted uncertainty, but instead, it delivers highly reliable, high-fidelity probabilities that are safe for proactive clinical decision support.

Table 2. Performance metrics of heterogeneous models and the proposed HFDE–CKD framework due to the No-SMOTE dataset.

Model	Accuracy	Sensitivity	Specificity	F1 Score	Brier Score	NLL	ECE
MLP	0.8474	0.9361	0.6451	0.8951	0.1061	0.3364	0.0440
XGBoost	0.8315	0.9427	0.5777	0.8861	0.1014	0.3215	0.0165
LightGBM	0.8218	0.9487	0.5323	0.8810	0.1077	0.3240	0.0264
KNN	0.8160	0.9361	0.5419	0.8761	0.1171	0.3671	0.0502
Proposed HFDE	0.8537	0.8794	0.7950	0.8931	0.1006	0.3117	0.0779

However, while HFDE demonstrates superior calibration under severe class imbalance, it must be highly stress-tested throughout a stable clinical manifold. So, we evaluate the proposed HFDE–CKD framework under the Standard SMOTE dataset in the next discussion.

As demonstrated in Table 3, the performance metrics (Acc, Sens, Spec, and F1 score) of all heterogeneous base models and the proposed HFDE–CKD was evaluated on the Standard SMOTE dataset. The outcomes indicate the significant weakness of gradient boosting methods in the presence of a clinical handicap. Despite LightGBM’s and XGBoost’s ability to obtain high sensitivities of 0.9456 and 0.9535, respectively, they were exhibited a low level of specificity: 0.4774 and 0.5185, respectively. Thus, these static base models suffered from probabilistic overconfidence, which led to the production of a lot of false-positive alerts, resulting in significant alert loss in clinical situations. In contrast, the proposed HFDE–CKD framework achieved a trade-off dynamically using epistemic uncertainty. As mentioned in Table 2, the proposed HFDE–CKD exhibited the maximum overall accuracy of 0.8596 and peak F1 score of 0.8979. Specifically, the uncertainty-injected meta-learner enhanced diagnostic safety, increasing specificity up to 0.7907 while maintaining a high sensitivity of 0.8900. Moreover, the proposed framework also showed its probabilistic superiority, with the lowest Brier Score of 0.1012 and a very competitive NLL of 0.3225. So, the proposed HFDE–CKD framework is not just guessing by measuring

predicted uncertainty but instead delivers reliable, high-fidelity probabilities that are safe for clinical decision support.

Table 3. Performance metrics of heterogeneous models and the proposed HFDE–CKD framework due to the Standard–SMOTE dataset.

Model	Accuracy	Sensitivity	Specificity	F1 Score	Brier Score	NLL	ECE
MLP	0.8529	0.9281	0.6826	0.8975	0.1029	0.3299	0.0188
XGBoost	0.8202	0.9535	0.5185	0.8803	0.1060	0.3268	0.0234
LightGBM	0.8022	0.9456	0.4774	0.8690	0.1071	0.3239	0.0241
KNN	0.7904	0.9607	0.4049	0.8641	0.1111	0.3474	0.0175
Proposed HFDE	0.8596	0.8900	0.7907	0.8979	0.1012	0.3225	0.0666

4.3. Sensitivity–Specificity Stabilization Analysis

As mentioned in Figure 2, the confusion matrices of four base learners (MLP, XGBoost, LightGBM, and KNN) and the proposed HDFE-CKD is evaluated. Clearly, models like LightGBM and XGBoost are strongly biased towards the positive class, as they correctly identify most of the real CKD patients but do so by substantially over-predicting the disease and yielding 340 and 307 false positives, respectively. On the other hand, MLP and KNN have the same diagnosis, yielding 258 and 333 false positives, respectively. Thus, the obtained results lead to probabilistic overconfidence and sensitivity and may cause significant alert loss in a real-world situation. As shown in Figure 2, the proposed HFDE–CKD approach provided significant results by leveraging epistemic uncertainty to reduce false positives and recover 149 and 578 true negatives, respectively, leading to a specificity of 79.5%. Moreover, by applying penalizing epistemic uncertainty, the proposed framework effectively halts the base learners from making predictions about ambiguous healthy profiles. Hence, the proposed HFDE–CKD framework has a reliable and trustworthy diagnostic scheme and can be an early screen tool for CKD.

As mentioned in Figure 3, the confusion matrices of four base learners (MLP, XGBoost, LightGBM, and KNN) and the proposed HFDE–CKD are evaluated. Clearly, models like LightGBM and XGBoost are strongly biased towards the positive class; they correctly identify most of the real CKD patients but do so by substantially over-predicting the disease, yielding 382 and 352 false positives, respectively. On the other hand, MLP and KNN exhibit similar diagnostic skews, yielding 232 and 435 false positives, respectively. Thus, these static base models suffer from probabilistic overconfidence, artificially inflating sensitivity, which would cause significant alert fatigue in a real-world clinical situation. As shown in Figure 3, the proposed HFDE–CKD approach provides significant improvements by leveraging epistemic uncertainty to reduce false positives to 153 and recover true negatives to 578, leading to a specificity of 79.1%. Moreover, by actively penalizing epistemic uncertainty, the proposed framework effectively halts the base learners from making predictions for ambiguous healthy profiles. Hence, the proposed HFDE–CKD framework offers a reliable and trustworthy diagnostic scheme to serve as an early screening tool for CKD.

4.4. ROC and Precision–Recall Analysis

As shown in Figure 4, receiver operating characteristic (ROC) and precision–recall (PR) curves were analyzed under both No-SMOTE and Standard SMOTE conditions. The obtained results of our proposed framework showed that ROC curves can frequently provide positive performance, with an AUC value of 0.927, which is related to the high proportion of true negatives according to the No-SMOTE scenario. Additionally, the PR curve is a significantly increasing measure. The proposed HFDE architecture effectively fixes this No-SMOTE data and achieves the maximum PR-AUC of 0.971. This indicates that the

uncertainty-injected meta-learner retains excellent prediction accuracy throughout the whole probability range, even without synthetic augmentation. In addition, after switching to the balanced Standard SMOTE distribution, traditional gradient boosting techniques like XGBoost and LightGBM suffer when forced to learn from the generated manifold that was adjusted, as illustrated in Figure 4. Thus, XGBoost's ROC-AUC is lowered from 0.927 to 0.919, and LightGBM follows suit. Meanwhile, the proposed HFDE-CKD framework depends on balanced geometry and improves performance by dynamically translating the variance between these weak base estimators, producing a peak ROC-AUC and PR-AUC of 0.931 and 0.972, respectively. So, the HFDE-CKD approach not only optimizes a single classification threshold but also incorporates uncertainty-aware fusion, which indicates superior discrimination between CKD and non-CKD cases without data leakage.

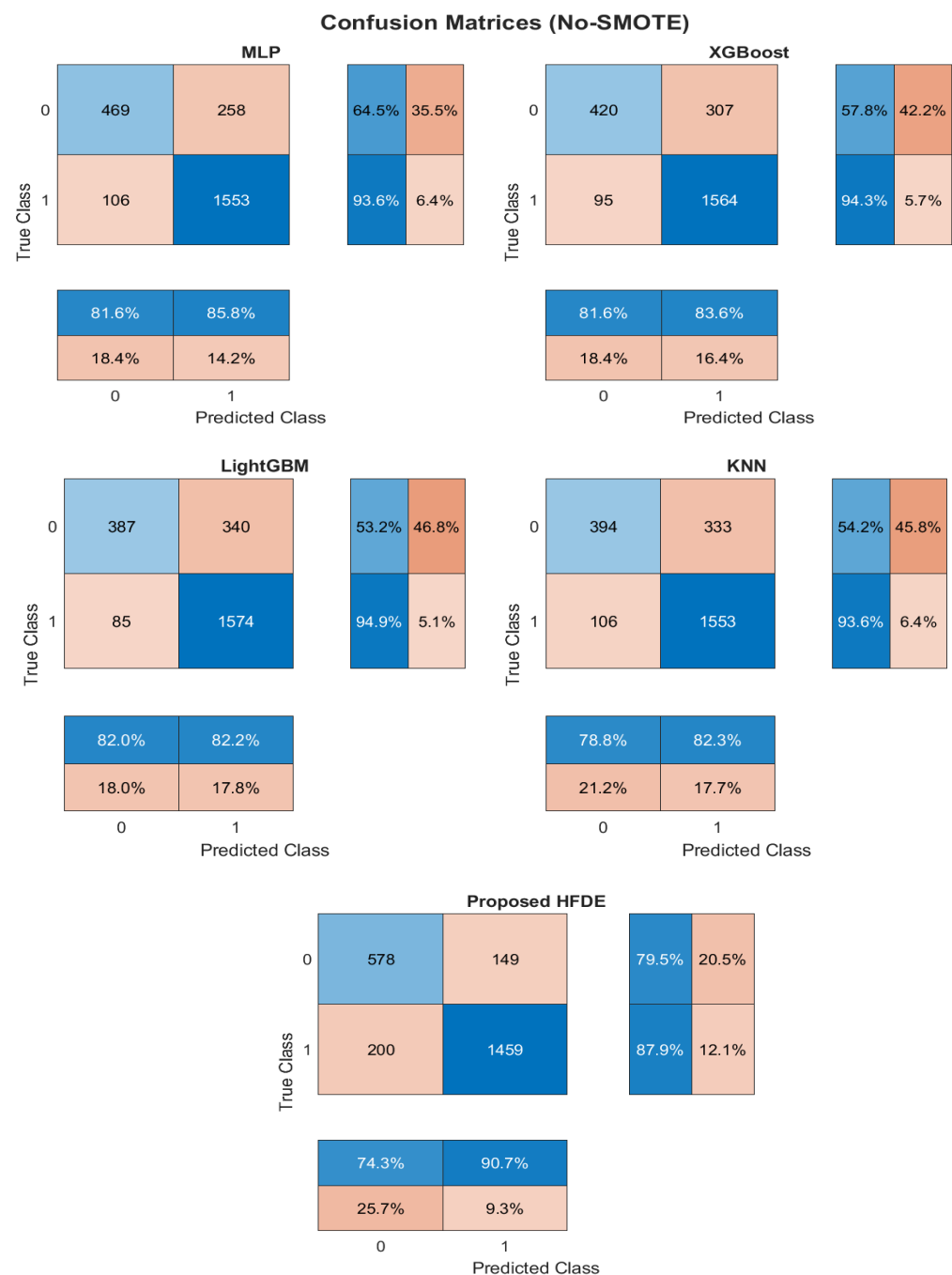


Figure 2. The confusion matrices of all base learner models and the proposed HFDE-CKD according to the No-SMOTE dataset.

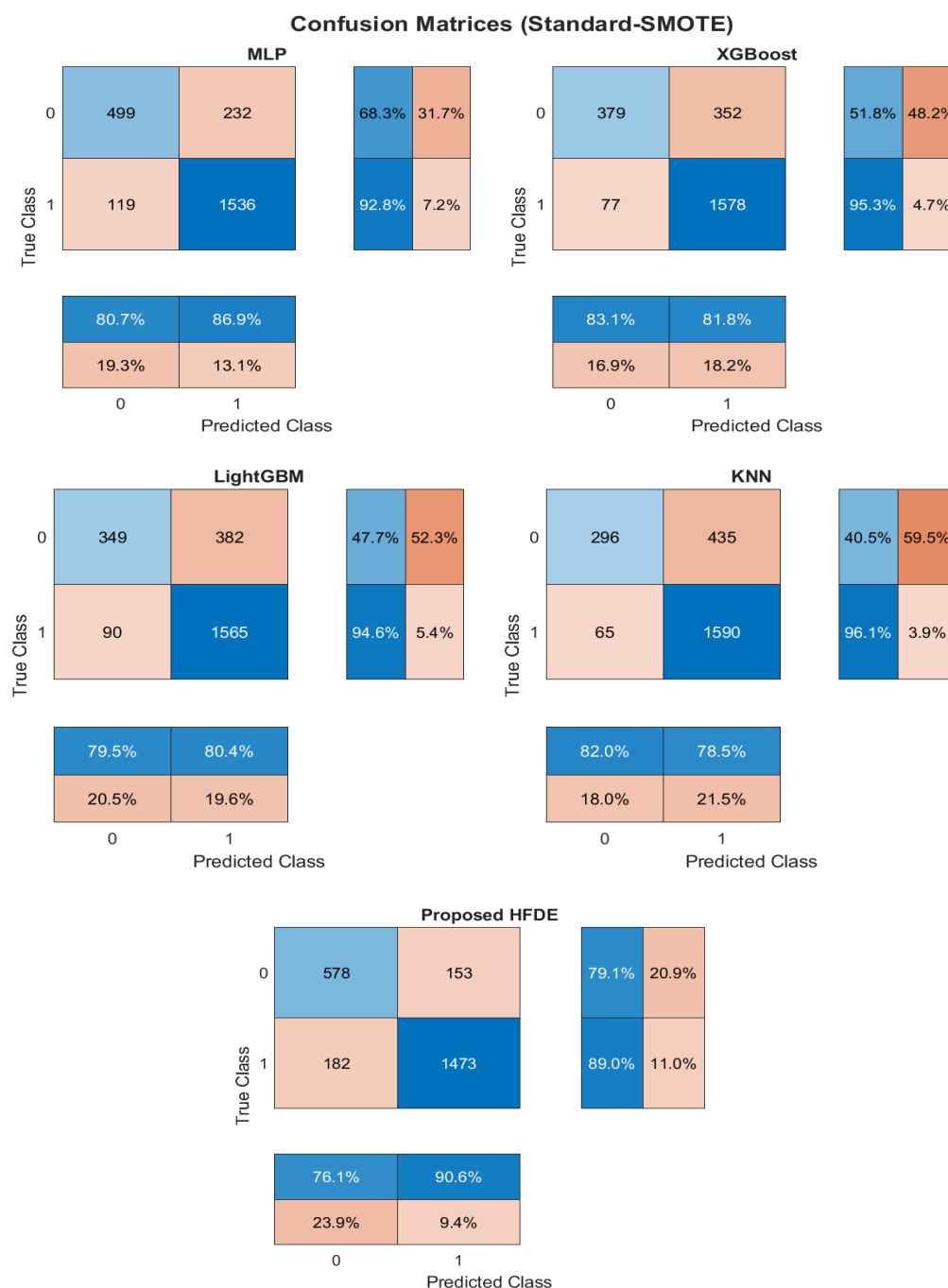


Figure 3. The confusion matrices of all base learner models and the proposed HFDE–CKD according to the Standard SMOTE dataset.

Therefore, these outcomes confirm that the proposed HFDE–CKD is highly reliable and effective in handling class separation, providing the strongest predictive stability among the evaluated classifiers.

4.5. Clinical Feature Correlation and Redundancy

As shown in Figure 5, a Pearson correlation matrix [59] was generated to evaluate the linear dependencies between clinical features within the No-SMOTE and Standard SMOTE datasets. Clearly, the feature importance rankings exhibit strong consistency across the baseline learners and the proposed HFDE framework. Despite minor variations in ordering, a stable subset of predictors consistently dominates the attribution profiles. The correlation matrix evaluates how strongly features are related to each other. High positive values

(red) indicate strong direct relationships, while negative values (blue) indicate inverse relationships. Most features show weak to moderate correlations, but strong associations are clearly visible among anthropometric attributes, such as body mass index (bmi) and weight (weight-kg), as well as among metabolic markers, including diagnosed diabetes (diabetes-diagnosed), insulin use (insulin-use), and diabetes medication (diabetes-pills). This highlights the inherent redundancy between certain physical and metabolic parameters and reinforces the importance of selecting only the most informative clinical variables for robust CKD predictions.

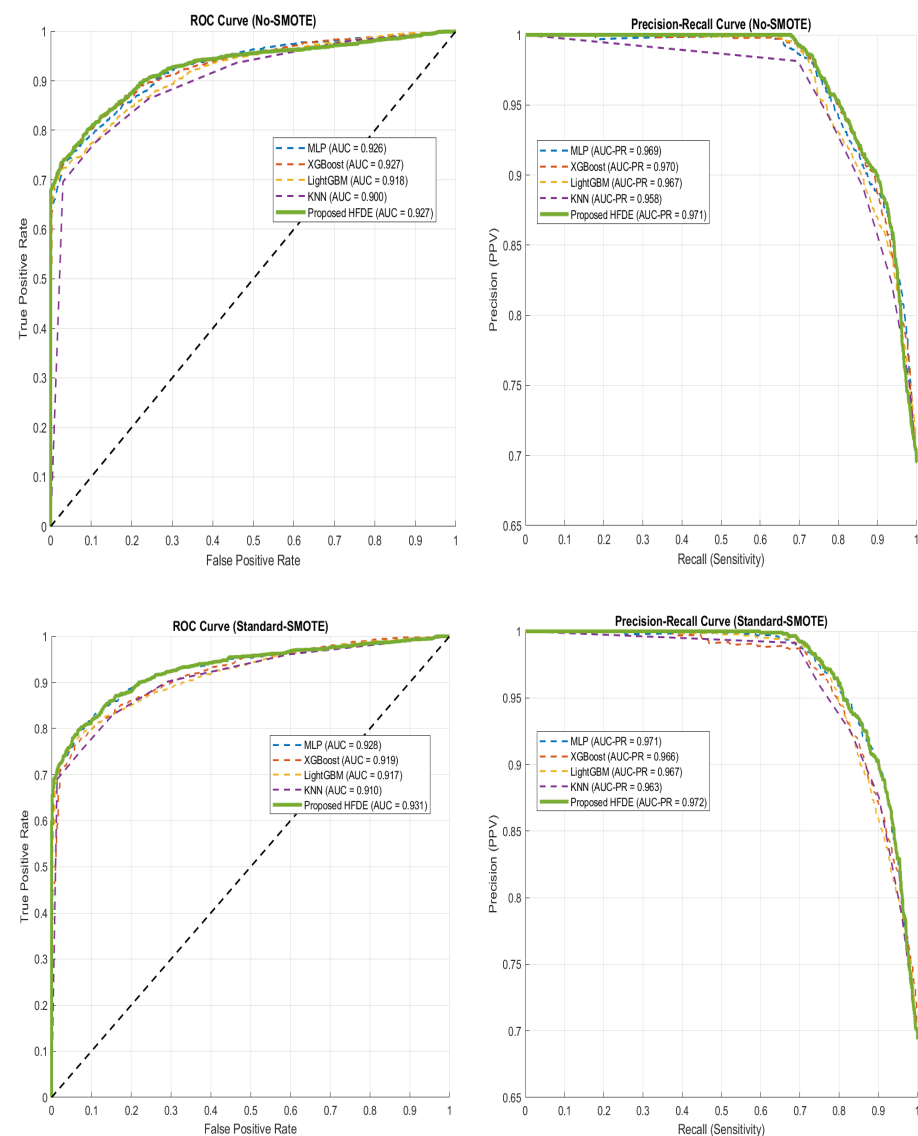


Figure 4. The RoC and PR-AUC curves of all base learners according to No-SMOTE and Standard SMOTE datasets.

As demonstrated in Figure 6, we illustrate the feature importance used in the underlying drivers of the proposed HFDE–CKD framework. It was evaluated using the OOB permutation error extracted from the random forest base estimators. The OOB delta error provides an unbiased estimate of feature relevance by measuring the drop in predictive accuracy when a specific variable is randomly shuffled. Consequentially, we compared clinical feature information gain under the No-SMOTE and Standard SMOTE features. The obtained results demonstrate that the age feature is overwhelmingly the strongest independent predictor of CKD across all data distributions, supported by uric acid and

calcium. Furthermore, the No-SMOTE model emphasizes late-stage metabolic collapse signs like phosphate and bicarbonate when assessing the unbalanced clinical manifold. Meanwhile, when trained on the balanced Standard SMOTE dataset, which includes more borderline and early-stage cases, the model prioritizes systemic and demographic vulnerabilities, increasing the predictive power of features like Non-Hispanic Black ethnicity. This shows that the proposed HFDE–CKD framework does not rely on a static rule set frequently; rather, it dynamically adjusts its feature priority to capture the most significant clinical signals depending on patient distribution complexity.

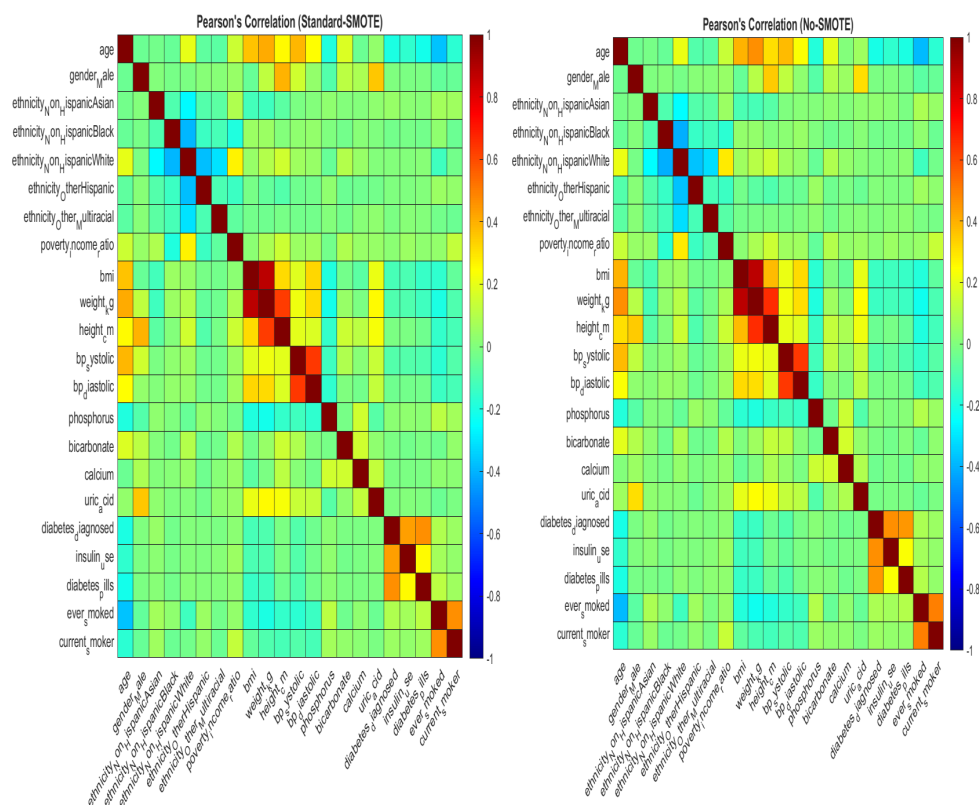


Figure 5. Feature selection analysis using a Pearson correlation matrix according to the No–SMOTE and Standard–SMOTE datasets.

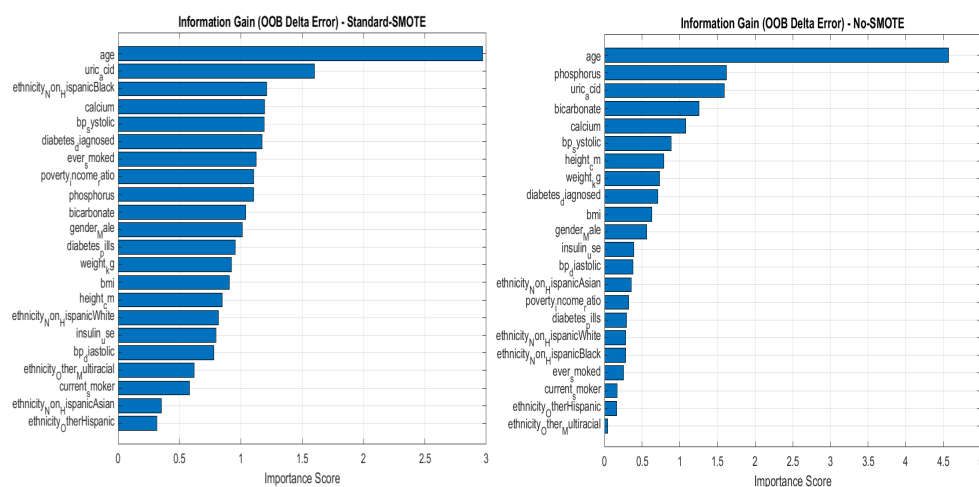


Figure 6. Feature selection analysis of the proposed HFDE–CKD framework using information gain according to the No-SMOTE and Standard SMOTE datasets.

4.6. Ablation Analysis and Statistical Significance

We performed an ablation study with the obtained results in Table 4 to evaluate and validate the independent usage of the uncertainty-aware fusion process. The proposed HFDE–CKD architecture was assessed on a simple baseline ensemble without the epistemic uncertainty quantification module. We also used McNemar’s test, which was utilized to evaluate the prediction differences between our proposed HFDE–CKD model and LightGBM, ensuring that the observed performance gains were not a result of random variations. The addition of the uncertainty module significantly improves the native, unbalanced No-SMOTE distribution, with the F1 score increasing from 0.9011 to 0.9332. Additionally, the McNemar test ($p < 0.001$) shows that the improvement is statistically significant. The variance matrix shows that HFDE effectively corrected 218 misclassifications by LightGBM (n_1), which shows the capacity of HFDE to prevent the probabilistic overconfidence of gradient boosting methods. Furthermore, McNemar’s test reveals a substantial and very significant shift in diagnostic behavior ($\chi^2 = 53.924$, $p < 0.001$). This HFDE–CKD design reclassified 240 patients that LightGBM missed in this manifold, but it only created 103 additional mistakes (n_2). This shows that while synthetic augmentation improves the stability of the global metrics of the baseline models, the theoretical uncertainty module actively reshapes the decision boundary at the patient level, trading arbitrary false positives for a much safer and more reliable clinical diagnostic footprint.

4.7. Local Interpretability and SHAP Analysis

As shown in Figure 7, SHapley Additive Explanations (SHAPs) [60] were created to simplify the HFDE–CKD framework decision-making process for high-risk queries according to No-SMOTE and Standard SMOTE datasets. SHAP measures exactly how much each clinical feature influenced the final decision, turning a complex algorithm into an interpretability framework that matches real-world medical knowledge. Across both the native No-SMOTE and Standard SMOTE datasets, age and phosphorus consistently serve as the primary drivers for a positive CKD diagnosis. Nevertheless, the adaptability of the HFDE–CKD architecture is its fundamental advantage. For instance, the model was significantly influenced by physical indicators such as weight and BMI when evaluating a patient in the No-SMOTE dataset, resulting in a positive prediction. In contrast, the model’s focus in the balanced Standard SMOTE dataset was to extensively consider demographic hazards, such as Non-Hispanic Black ethnicity and lifestyle history (for example, having ever smoked). This proves that the HFDE–CKD framework does not rely on rigid, one-size-fits-all rules. Instead, it creates a personalized, transparent risk profile for every patient, allowing clinicians to see exactly why a specific diagnostic alert was triggered.

Table 4. Ablation study and McNemar’s statistical test evaluating the impact of epistemic uncertainty injection under No-SMOTE and Standard SMOTE datasets.

Data Condition	Ablation Configuration	F1 Score	McNemar’s Test (HFDE vs. LightGBM)			
			n_1	n_2	χ^2 Statistic	p -Value
No–SMOTE	Baseline (No Uncertainty)	0.9011	218	142	15.625	<0.001 ***
	Proposed HFDE (With Uncertainty)	0.9332				
Standard–SMOTE	Baseline (No Uncertainty)	0.8981	240	103	53.924	<0.001 ***
	Proposed HFDE (With Uncertainty)	0.8999				

Note: n_1 denotes the number of cases where HFDE was correct and LightGBM was incorrect. n_2 denotes cases where HFDE was incorrect and LightGBM was correct. *** denotes statistical significance at $\alpha = 0.001$.

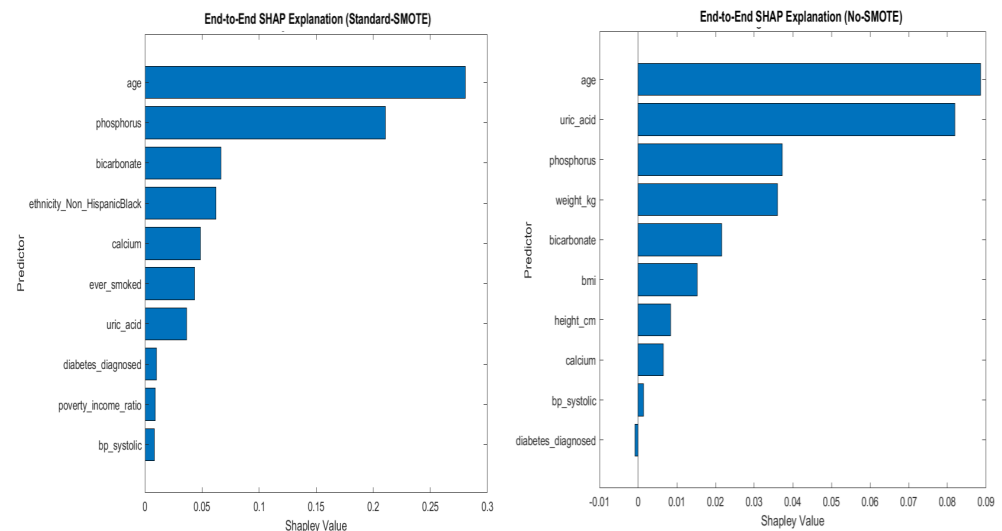


Figure 7. Local interpretability of the HFDE-CKD framework using SHAP according to the No-SMOTE and Standard-SMOTE datasets.

5. Conclusions

This study introduces the Hyper-Fidelity Dynamic Ensemble (HFDE-CKD) framework for chronic kidney disease (CKD) diagnosis to overcome probabilistic overconfidence and data leakage—major flaws in modern diagnostic models. We present a novel meta-learning framework that outperforms standard ensemble models by functioning as a dynamic cognitive gatekeeper, measuring the epistemic uncertainty of heterogeneous base learners (MLP, XGBoost, LightGBM, and KNN). The proposed HFDE-CKD outperforms conventional standalone algorithms on the NHANES dataset, particularly in clinical scenarios where late-stage renal indicators (e.g., eGFR, serum creatinine) are excluded. Evaluated across No-SMOTE and Standard SMOTE distributions, the framework effectively reduces epistemic uncertainty to enhance true negative recovery. Dynamic stabilization maintains a robust specificity of approximately 79%, significantly decreasing false alarms while preserving peak F1 scores up to 0.8979 and a PR-AUC of 0.972. The framework exhibits high clinical dependability through enhanced probabilistic calibration, achieving Brier scores and NLL as low as 0.1006 and 0.3117, respectively. Ablation analyses confirm statistical superiority (McNemar’s test, $p < 0.001$), demonstrating that the uncertainty-injected meta-learner corrects base learner misclassifications. In addition to robust predictive accuracy, the framework’s clinical transparency was validated by comprehensive Pearson correlation, information gain, and SHAP evaluations. Furthermore, the model effectively addressed feature redundancies and dynamically adjusted its diagnostic based on metabolic indicators during No-SMOTE distributions while adaptively transitioning to demographic vulnerabilities and lifestyle risks to assess borderline cases under the Standard SMOTE distribution. Finally, we conclude that the HFDE-CKD framework provides a precisely calibrated, robust, and reliable decision-support tool for the early screening of CKD diagnosis.

Future research should focus on integrating other data modalities such as medical imaging, longitudinal laboratory measurements, and genomic information to enhance diagnostic accuracy even further. Moreover, interpretability could be improved by incorporating causal modeling and explainable attention mechanisms that would provide clinicians with deeper insights into the impact of individual features on diagnosis decisions.

Author Contributions: Conceptualization, methodology, writing—review and editing, project administration, and funding acquisition, M.H.; methodology, software, validation, formal analysis, visualization, and writing—original draft preparation, M.A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Prince Sattam bin Abdulaziz University, Saudi Arabia, through project number (PSAU/2025/01/32904).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository the original data presented in the study are openly available in [kaggle] at [<https://www.kaggle.com/datasets/alitaqishah/ckd-nhanes-2021-2023-staged-kidney-disease/data>] (accessed on 9 July 2026)].

Acknowledgments: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through project number (PSAU/2025/01/32904).

Conflicts of Interest: The authors declare no conflicts of interests.

References

1. Iftikhar, H.; Hashem, A.F.; Qureshi, M.; Rodrigues, P.C. Clinical application of machine learning models for early-stage chronic kidney disease detection. *Diagnostics* **2025**, *15*, 2610. [[CrossRef](#)] [[PubMed](#)]
2. Wang, P.; Meng, Y.; Sun, Z.; Li, J.; Tao, H. Development of an interpretable machine learning model for predicting 4-year chronic kidney disease risk in elderly hypertensive patients. *Int. J. Med. Inform.* **2026**, *211*, 106320. [[CrossRef](#)] [[PubMed](#)]
3. Xiong, Z.; He, J.; Valkema, P.; Nguyen, T.Q.; Naesens, M.; Kers, J.; Verbeek, F.J. Advances in kidney biopsy lesion assessment through dense instance segmentation. *Artif. Intell. Med.* **2025**, *164*, 103111. [[CrossRef](#)] [[PubMed](#)]
4. Ramu, K.; Patthi, S.; Prajapati, Y.N.; Ramesh, J.V.N.; Banerjee, S.; Rao, K.B.; Alzahrani, S.I.; ayyasamy, R. Hybrid CNN-SVM model for enhanced early detection of Chronic kidney disease. *Biomed. Signal Process. Control* **2025**, *100*, 107084. [[CrossRef](#)]
5. Islam, M.S.B.; Sumon, M.S.I.; Sarmun, R.; Bhuiyan, E.H.; Chowdhury, M.E. Classification and segmentation of kidney MRI images for chronic kidney disease detection. *Comput. Electr. Eng.* **2024**, *119*, 109613. [[CrossRef](#)]
6. Anbazhagan, T.; Rangaswamy, B. Early prediction of CKD from time series data using adaptive PSO optimized echo state networks. *Sci. Rep.* **2025**, *15*, 6966. [[CrossRef](#)] [[PubMed](#)]
7. Kumar, S. Advancements in medical image segmentation: A review of transformer models. *Comput. Electr. Eng.* **2025**, *123*, 110099. [[CrossRef](#)]
8. Esmail, E.; Mohammed, A.A.; Mohammed, R.A.; Almuhaaya, B.; Fadhl, A.; Bazel, M.A. Kidney Disease Classification and Diagnosis: A Comprehensive Review of Current AI Techniques. In *Proceedings of the 4th International Conference on Emerging Smart Technologies and Applications*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–13.
9. Ma, F.; Sun, T.; Liu, L.; Jing, H. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. *Future Gener. Comput. Syst.* **2020**, *111*, 17–26. [[CrossRef](#)]
10. Rabie, A.H.; Saleh, A.I. Diseases diagnosis based on artificial intelligence and ensemble classification. *Artif. Intell. Med.* **2024**, *148*, 102753. [[CrossRef](#)] [[PubMed](#)]
11. Saif, D.; Sarhan, A.M.; Elshennawy, N.M. Deep-kidney: An effective deep learning framework for chronic kidney disease prediction. *Health Inf. Sci. Syst.* **2023**, *12*, 3. [[CrossRef](#)] [[PubMed](#)]
12. Khan, S.U.R. Multi-level feature fusion network for kidney disease detection. *Comput. Biol. Med.* **2025**, *191*, 110214. [[CrossRef](#)]
13. Jain, R.; Kukreja, V.; Chattopadhyay, S.; Verma, A.; Sharma, R. Deep Learning Insights: Evaluating the Efficacy of CNN-SVM for Accurate Kidney Disease Identification. In *Proceedings of the IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5.
14. Yang, C.; Tonelli, M.; James, M.T.; Tan, Z.; Bakker, W.M.; Gansevoort, R.T.; Vart, P. Incidence and Adverse Outcomes of Acute Kidney Disease: A Systematic Review and Meta-Analysis. *Am. J. Kidney Dis.* **2025**, *166*, 103153.
15. Kiremit, B.Y.; Şahin, D.Ö. Comparison of machine learning algorithms for predicting length of stay in chronic kidney disease patients. *Comput. Biol. Med.* **2025**, *196*, 110825. [[CrossRef](#)] [[PubMed](#)]
16. Houssein, E.H.; Sayed, A. A modified weighted mean of vectors optimizer for Chronic Kidney disease classification. *Comput. Biol. Med.* **2023**, *155*, 106691. [[CrossRef](#)] [[PubMed](#)]
17. Begoli, E.; Bhattacharya, T.; Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* **2019**, *1*, 20–23. [[CrossRef](#)]
18. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [[CrossRef](#)]

19. Ghosh, S.K.; Widatalla, N.; Khandoker, A.H. Machine Learning Framework for Early Detection of Chronic Kidney Disease Stages Using Optimized Estimated Glomerular Filtration Rate. *IEEE Access* **2025**, *13*, 78057–78072. [[CrossRef](#)]
20. Saif, D.; Sarhan, A.M.; Elshennawy, N.M. Early prediction of chronic kidney disease based on ensemble of deep learning models and optimizers. *J. Electr. Syst. Inf. Technol.* **2024**, *11*, 17. [[CrossRef](#)]
21. Prasad Reddy, T.B.; Gurav, S.; Sekar, R.; Satpute, B. Optimization assisted ensemble classification for prediction of chronic kidney disease. *Multimed. Tools Appl.* **2025**, *84*, 19551–19577.
22. Pinheiro, J.M.H.; de Oliveira, S.V.B.; Silva, T.H.S.; Saraiva, P.A.R.; de Souza, E.F.; Godoy, R.V.; Ambrosio, L.A.; Becker, M. The impact of feature scaling in machine learning: Effects on regression and classification tasks. *IEEE Access* **2025**, *13*, 199903–199931. [[CrossRef](#)]
23. Yildiz, E.; Cengil, E.; Yildirim, M.; Bingol, H. Diagnosis of chronic kidney disease based on CNN and LSTM. *Acadlore Trans. AI Mach. Learn.* **2023**, *2*, 66–74. [[CrossRef](#)]
24. Ghosh, P.; Shamrat, F.J.M.; Shultana, S.; Afrin, S.; Anjum, A.A.; Khan, A.A. Optimization of prediction method of chronic kidney disease using machine learning algorithm. In *Proceedings of the 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
25. Tirumalasetty, M.L.; Vuppuloori, R.S.R.; Tata, B.; Maddipati, V.G.R.; Navaneethan, J.; Kurra, U.C.; Kodepogu, K.R.; Gaddala, L.K.; Yalamanchil, S. Systematic Survey on Chronic Kidney Disease Prediction Using Different Machine Learning Techniques. *Rev. D'Intell. Artif.* **2023**, *37*, 1645. [[CrossRef](#)]
26. Elbedwehy, S.; Hassan, E.; Saber, A.; Elmonier, R. Integrating neural networks with advanced optimization techniques for accurate kidney disease diagnosis. *Sci. Rep.* **2024**, *14*, 21740. [[CrossRef](#)] [[PubMed](#)]
27. Rahman, M.M.; Al-Amin, M.; Hossain, J. Machine learning models for chronic kidney disease diagnosis and prediction. *Biomed. Signal Process. Control* **2024**, *87*, 105368. [[CrossRef](#)]
28. Dharmarathne, G.; Bogahawaththa, M.; McAfee, M.; Rathnayake, U.; Meddage, D. On the diagnosis of chronic kidney disease using a machine learning-based interface with explainable artificial intelligence. *Intell. Syst. Appl.* **2024**, *22*, 200397. [[CrossRef](#)]
29. Shamija Sherryl, R.; Jaya, T. Semantic multiclass segmentation and classification of kidney lesions. *Neural Process. Lett.* **2023**, *55*, 1975–1992.
30. Walse, R.S.; Kurundkar, G.D.; Khamitkar, S.D.; Muley, A.A.; Bhalchandra, P.U.; Lokhande, S.N. Effective use of naïve bayes, decision tree, and random forest techniques for analysis of chronic kidney disease. In *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems*; Springer: Singapore, 2020; pp. 237–245.
31. Nithya, A.; Appathurai, A.; Venkatadri, N.; Ramji, D.; Palagan, C.A. Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. *Measurement* **2020**, *149*, 106952. [[CrossRef](#)]
32. Yadav, D.C.; Pal, S. Performance based evaluation of algorithmson chronic kidney disease using hybrid ensemble model in machine learning. *Biomed. Pharmacol. J.* **2021**, *14*, 1633–1645. [[CrossRef](#)]
33. Bhuiyan, M.S.M.; Rafi, M.A.; Rodrigues, G.N.; Mir, M.N.H.; Ishraq, A.; Mridha, M.; Shin, J. Deep learning for algorithmic trading: A systematic review of predictive models and optimization strategies. *Array* **2025**, *26*, 100390. [[CrossRef](#)]
34. Egene, A.I.; Osaghae, E.O.; Basaky, F.D. Chronic kidney disease prediction model using Bayesian optimization and XGBoost machine learning algorithm. *FUDMA J. Sci.* **2025**, *9*, 161–171. [[CrossRef](#)]
35. Manju, V.; Aparna, N.; Krishna, K. Decision tree-based explainable AI for diagnosis of chronic kidney disease. In *Proceedings of the 5th International Conference on Inventive Research in Computing Applications*; IEEE: Piscataway, NJ, USA, 2023; pp. 947–952.
36. Dritsas, E.; Trigka, M. Machine learning techniques for chronic kidney disease risk prediction. *Big Data Cogn. Comput.* **2022**, *6*, 98. [[CrossRef](#)]
37. Singh, V.; Asari, V.K.; Rajasekaran, R. A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics* **2022**, *12*, 116. [[CrossRef](#)] [[PubMed](#)]
38. Komal, K.N.; Tulasi, R.L.; Vigneswari, D. An ensemble multi-model technique for predicting chronic kidney disease. *Int. J. Electr. Comput. Eng.* **2019**, *9*, 1321. [[CrossRef](#)]
39. Yang, W.; Ahmed, N.; Barczak, A. Comparative analysis of machine learning algorithms for CKD risk prediction. *IEEE Access* **2024**, *12*, 171205–171220. [[CrossRef](#)]
40. Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access* **2019**, *8*, 20991–21002. [[CrossRef](#)]
41. Singh, V.; Jain, D. A hybrid parallel classification model for the diagnosis of chronic kidney disease. *Int. J. Interact. Multimed. Artif. Intell.* **2023**, *8*, 14–28. [[CrossRef](#)]
42. Rahat, M.A.R.; Islam, M.T.; Cao, D.M.; Tayaba, M.; Ghosh, B.P.; Ayon, E.H.; Nobe, N.; Akter, T.; Rahman, M.; Bhuiyan, M.S. Comparing machine learning techniques for detecting chronic kidney disease in early stage. *J. Comput. Sci. Technol. Stud.* **2024**, *6*, 20–32. [[CrossRef](#)]
43. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]

44. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
45. Inam, S.A.; Rajput, H.; Umer, S. A hierarchical ensemble approach for multi-country PM10 forecasting using LightGBM and residual neural network. *Discov. Atmos.* **2026**, *4*, 1. [[CrossRef](#)]
46. Ramamurthy, K.; Gumber, S.; Abdelfattah, W.M.; Khan, N.A.; Mushtaq, Z.; Siddique, I. Human AI trust modeling in cognitive systems via ensemble learning and advanced feature engineering. *Discov. Artif. Intell.* **2026**, *6*, 366. [[CrossRef](#)]
47. Sarma, M.; Chatterjee, S. ‘Machine Learning’ multiclassification for stage diagnosis of Alzheimer’s disease utilizing augmented blood gene expression and feature fusion. *Discov. Appl. Sci.* **2025**, *7*, 636. [[CrossRef](#)]
48. Lin, J.; Li, Y.; Bian, C.; Fang, Q.; Li, C.; Yang, X.; Zhang, G. Feature selection by joint manifold structures for transformer fault diagnostic. *Electr. Power Syst. Res.* **2026**, *256*, 112906. [[CrossRef](#)]
49. Chidambaram, M.; Ge, R. Reassessing how to compare and improve the calibration of machine learning models. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025; Volume 2025, pp. 61542–61570.
50. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2017; pp. 1321–1330.
51. Apicella, A.; Isgrò, F.; Prevete, R. Don’t push the button! exploring data leakage risks in machine learning and transfer learning. *Artif. Intell. Rev.* **2025**, *58*, 339. [[CrossRef](#)]
52. Arunkumar, P.; Brindha, R.; Dharanika, C.; Elakkiyaa, B. Data Sampling Methods for Predicting CKD and Diabetes with Machine Learning. In *Proceedings of the 2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
53. Silva Filho, T.; Song, H.; Perello Nieto, M.; Santos Rodriguez, R.; Kull, M.; Flach, P. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Mach. Learn.* **2023**, *112*, 3211–3260. [[CrossRef](#)]
54. Arrieta Ibarra, I.; Gujral, P.; Tannen, J.; Tygert, M.; Xu, C. Metrics of calibration for probabilistic predictions. *J. Mach. Learn. Res.* **2022**, *23*, 1–54. [[CrossRef](#)]
55. Pembury Smith, M.Q.; Ruxton, G.D. Effective use of the McNemar test. *Behav. Ecol. Sociobiol.* **2020**, *74*, 133. [[CrossRef](#)]
56. Marengo, A.; Santamato, V. A Novel Machine Learning-Optimized Framework for Systematic Analysis of Foundation Models in Healthcare: Comprehensive Algorithm Optimization With Governance-Driven Predictive Modeling. *IEEE Access* **2025**, *13*, 210040–210088. [[CrossRef](#)]
57. Carrington, A.M.; Manuel, D.G.; Fieguth, P.W.; Ramsay, T.; Osmani, V.; Wernly, B.; Bennett, C.; Hawken, S.; Magwood, O.; Sheikh, Y.; et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 329–341. [[CrossRef](#)] [[PubMed](#)]
58. Shah, A.T. CKD-NHANES 2021–2023: Staged Kidney Disease. 2026. Available online: <https://www.kaggle.com/datasets/alitaqishah/ckd-nhanes-2021-2023-staged-kidney-disease> (accessed on 1 May 2026).
59. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [[CrossRef](#)]
60. Netayawijit, P.; Chansanam, W.; Sorn-In, K. Interpretable Machine Learning Framework for Diabetes Prediction: Integrating SMOTE Balancing with SHAP Explainability for Clinical Decision Support. *Healthcare* **2025**, *13*, 2588. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.