# Maximizing Sample Utilization in CKD Classification: Fusion and Alignment of Locally Trained Models with a Global Model

Ali Guran[1,*], Avishek Siris[1], Gary K.L. Tam[1], James Chess[2], Xianghua Xie[1]

*Abstract*— **Chronic Kidney Disease (CKD) is a significant global health issue that requires accurate classification for effective management. While machine learning techniques have shown promise in predicting CKD stages, obtaining comprehensive datasets with all relevant clinical features remains a major challenge. Existing works often face a trade-off between using small, complete datasets or larger, biased ones due to imputation. Additionally, data is often derived from various sources—such as blood and urine samples, along with demographic information—each requiring individual models that may vary in sample size. In this study, we explore this challenge by maximizing sample usage while minimizing reliance on imputation through the fusion of models trained on different data sources. Our approach integrates intermediate representations from these models with global representations from a Mixed Model, trained on a smaller but complete dataset that incorporates attributes from all sources. By leveraging cross-attention and self-attention mechanisms, our method improves staging accuracy and enhances model generalizability. This framework is particularly beneficial for clinical decision support when complete datasets are scarce, yet partial datasets are available, such as with CKD data from the UK SAIL databank.**

***Keywords*— *chronic kidney disease, multi-feature fusion, cross and self-attention, clinical decision support***

## I. Introduction

Chronic Kidney Disease (CKD) is a global health problem marked by gradual, irreversible kidney function loss, often progressing to end-stage renal disease if untreated. It is classified into five stages based on glomerular filtration rate (GFR), with early stages usually asymptomatic and later stages requiring dialysis or transplantation. Diabetes, hypertension, and glomerulonephritis are key causes, and its prevalence is rising due to aging populations and growing chronic disease rates. In 2017, CKD was the 12th leading cause of death globally, affecting around 700 million people [1]. Despite historical breakthroughs—such as dialysis and transplantation—many cases remain undiagnosed until advanced stages, highlighting the need for earlier detection and better diagnostic approaches.

Machine Learning (ML) and Deep Learning (DL) are revolutionizing clinical medicine by analyzing large datasets to detect patterns and make accurate predictions. By combining lab tests, clinical records, imaging, and genomics, they enhance diagnosis, personalize treatment, and improve monitoring. For example, changes in creatinine or urea levels can signal early disease progression. Benefits include automated data analysis, reduced human error, and continuous learning. These technologies support early disease detection, outcome prediction, and real-time decision-making, transforming patient care.

Existing research on CKD classification in the literature can largely be divided into two categories: binary and multi-class prediction tasks. Most of these studies focus on differentiating between CKD and non-CKD (binary classification), leveraging a variety of features and traditional machine learning methods [2]–[4]. In contrast, there is not many research on multi-class classification, which involves in predicting one of the five stages of CKD [5]–[7]. For the diagnosis of multi-stage CKD, new methods that can handle the intricacy of multi-class classification are required in order to provide better practical application and clinical relevance.

Predicting Chronic Kidney Disease (CKD) with high accuracy is hindered by incomplete and inconsistently collected multi-variate cross-sectional data, a challenge that is pervasive across healthcare datasets. Many clinical records contain missing values due to heterogeneous testing protocols, where patients undergo different diagnostic panels at varying times based on medical necessity, resource availability, or institutional practices. This inconsistency leads to significant gaps in feature availability, restricting the development of robust prediction models. To mitigate this issue, data imputation techniques are often employed to estimate missing values and preserve sample size. However, imputation carries the risk of introducing bias, distorting underlying patterns, and reducing model reliability, especially when large portions of data are inferred rather than observed [8]. On the other hand, excluding incomplete data ensures the integrity of feature relationships but drastically limits the number of usable samples, weakening model robustness and generalizability. This trade-off between data completeness and sample size is a critical barrier in CKD prediction, highlighting the need for innovative approaches that maximize the utility of partial yet valuable data while maintaining model accuracy and reliability.

In this work, we propose a novel framework to address the challenge of incomplete CKD datasets by training feature-specific models on different data facets and integrating them into a unified prediction system. CKD-related features in the literature typically fall into three key facets of data: (1) demographic data and creatinine, (2) other blood test results, and (3) urine sample data. While conventional approaches attempt to construct large datasets with complete feature sets, missing values often limit their practicality. Instead,

*Corresponding Author: 935538@swansea.ac.uk

[1]Department of Computer Science, Swansea University, Swansea, United Kingdom

[2]Wales Kidney Research Unit and Morriston Hospital, Swansea, United Kingdom

we explore an alternative research question: Can individual models be trained on different facets of CKD data, each with varying sample sizes, and then fused to form a globally effective model? By doing so, we can potentially maximize the utility of available samples and features, leveraging their diversity to improve model accuracy and generalization.

To investigate this, we develop three specialized models: one trained on the demographic facet (creatinine, age, gender), another on the blood test facet (albumin, sodium, potassium, urea, white blood cell count, red blood cell count), and a third on the urine sample facet (albumin in urine, urine albumin-to-creatinine ratio). In parallel, we train a mixed model on a smaller but fully comprehensive dataset encompassing all 13 features. We then extract local representations from the individual models and a global representation from the mixed model, integrating them using cross-attention and self-attention layers. This fusion mechanism enables the model to effectively capture both local and global feature relationships, enhancing multi-stage CKD classification performance despite the constraints of incomplete multivariate data.

Our experimental findings demonstrate that the proposed attention-based fusion model consistently outperforms traditional machine learning methods commonly employed in CKD classification, such as Random Forest, Decision Trees, Support Vector Machine and XGBoost. By leveraging both the large, partially complete dataset and the smaller, fully comprehensive dataset, our model is able to maximize the use of all the samples and features. This enables our model to learn more useful relationships between multiple features for multi-stage CKD classification, offering superior reliability and clinical relevance compared to existing techniques.

We summarize our main contributions as follows:

- To the best of our knowledge, we are the first to address the challenge of maximizing the utilization of both samples and features in CKD datasets, aiming to enhance model accuracy and generalization for multi-stage CKD classification.
- To tackle this challenge, we propose a robust framework that effectively fuses features from multiple models, each trained on different data facets (with varying features and sample sizes). We then explore the relationships between these models and align them to a global model, creating a local-global mechanism for more comprehensive predictions.
- Experimental results demonstrate superior accuracy and consistency compared to traditional techniques such as logistic regression, decision trees, and random forests, highlighting the effectiveness of our method in real-world clinical scenarios.

## II. RELATED WORKS

*1) CKD classification:* Recent advancements in artificial intelligence and machine learning have significantly improved the prediction of CKD. Various ML algorithms, including decision trees, support vector machines (SVM), artificial neural networks (ANN), and ensemble methods like random forests, have been applied to CKD diagnosis and progression prediction. Schena et al. conducted a systematic review of ML applications in CKD, emphasizing the ability of these models to integrate multiple predictors and capture complex, nonlinear interactions [9]. However, they also identified critical challenges, such as the lack of physician-validated datasets, insufficient external validation, and limited collaboration between nephrologists and computer scientists. Despite these challenges, ML-based approaches have demonstrated high predictive accuracy and potential for clinical decision support, highlighting the need for further research to enhance model interpretability and generalizability. In contrast to existing work that relies on well-prepared clinical datasets, our study addresses the challenge of maximizing the utilization of both samples and features. We develop individual models based on different facets of CKD data, each with varying feature sets and sample sizes, and then combine them into a local-global model. This approach aims to improve prediction accuracy and generalization by aligning and fusing the strengths of each model, leveraging the diversity of available data sources.

*2) Handling missing values:* Imputation techniques are commonly used to address missing data, but they have limitations that can compromise the reliability of results. A key issue is the assumption that data are Missing at Random (MAR), which may not hold for Missing Not at Random (MNAR) data, leading to biased outcomes [8]. Additionally, single imputation methods, such as mean substitution, underestimate variability and distort relationships between variables, weakening correlations and reducing data integrity [10]. While more advanced machine learning-based imputation methods are more robust, they require significant computational resources, making them less feasible for large datasets [10]. Unlike the approaches discussed above, our method prioritizes data integrity by avoiding imputation entirely. Instead, we leverage a local-global fusion strategy that optimizes the use of available data while maintaining trustworthiness, a critical consideration missing in many existing methods.

*3) Fusion methods:* Multi-modal attention models are widely used for integrating heterogeneous data, but most standard approaches assume complete data from each modality and are not designed to handle the common scenario of missing values across different views [11]. Federated learning (FL) is another approach to combine model outputs from distributed sources to protect privacy, but it presumes consistent features across datasets and does not address missing or incomplete data within a single cohort [12]. Several other relevant domains can be listed as representation-focused approaches (aligning or generating cross-modal representations), architecture-focused strategies (such as attention-based models, distillation, and graph learning), and model combination methods (including ensembles, dedicated models, and discrete schedulers) [13]. There is also recent interest in multi-modal learning with missing modality, which explores instance-based shared encoders and multi-stream data with missing modalities [11]. In [14], partial multi-view clustering is addressed by utilising common instances as anchors to reconstruct cross-view similarities, which are then fused into a single matrix to cluster incomplete multi-view data. Anchored-fusion is proposed by [15], which is a gene fusion detection method that targets a specific gene of interest (anchor) to improve sensitivity and accuracy, especially in low-quality or single-cell RNA-seq data. It aligns paired-end RNA-seq reads with a user-specified anchor gene to identify a potential fusion partner. In contrast to the above techniques, our method combines models trained independently on each data source (view-based) with a mixed model trained on the subset of patients with all attributes. We use the features from the mixed model to guide the fusion of latent features learned separately on partial datasets. This enables learning of the alignment between features from models trained with partial data with a representative complete data from the mixed model, while fully exploiting incomplete and variable size multi-source clinical data.

## III. DATASET

In our study, we utilized the Welsh Longitudinal General Practice Dataset (WLGP) [16] and the Welsh Results Reports Service (WRRS) [17] through the Secure Anonymised Information Linkage (SAIL) Databank [18]–[22]. The SAIL Databank is a trusted research environment in Wales that securely manages anonymized health and administrative data under strict privacy safeguards. By linking multiple datasets—from primary care to hospital and social records—SAIL enables robust, population-level research on health trends, outcomes, and interventions.

WLGP covers approximately 83% of the Welsh population and 80% of GP practices, providing detailed, longitudinal information on patient consultations, including symptoms, diagnoses, prescriptions, and test results, as well as social and environmental factors. This dataset can be linked within SAIL to other health and social data, making it a powerful resource for analyzing disease progression, treatment effectiveness, and broader population health patterns.

WRRS streamlines access to laboratory and pathology results across Wales, supporting both primary and secondary care. By integrating with the Welsh Clinical Portal (WCP), it offers a unified view of patient test data regardless of where tests were performed, reducing duplicate testing and improving patient safety. Authorized healthcare professionals can securely enter, view, and update results in real time, fostering greater efficiency and transparency in patient care.

We performed the following steps to build our complete dataset, and multiple subsets that maximizes available data samples for learning local models on various facets.

**Data Extraction and Initial Filtering:** We perform initial data extraction and filtering by first identifying patients labelled with any stage of CKD (from Stage 1 to Stage 5) in the WLGP dataset. We then removed patients who were missing essential demographic information, specifically age or gender. Next, we linked these patients to the WRRS dataset to retrieve their available laboratory test results, including creatinine, albumin, potassium, sodium, urea, red blood cell count, white blood cell count, albumin in urine, and urine albumin-to-creatinine ratio (uACR). These tests were selected based on their frequent use in CKD-related research, high feature importance reported in the literature, and validation through clinical expert consultation.

**Complete Dataset (CD):** We build our fully comprehensive and complete dataset, by excluding patients who did not have all relevant laboratory tests at the time of their CKD stage labels, as well as those who lacked any test results. During our review, we also removed biologically implausible negative values. Furthermore, we noted some patients whose CKD stages did not align with the eGFR ranges commonly referenced in the literature. To address this discrepancy while maintaining dataset integrity, we categorized patients into two groups: those with stage labels matching their eGFR values (in-range, InR) and those with stage labels inconsistent with their eGFR values (out-of-range, OoR). To mitigate potential biases and improve model generalizability, we implemented a stratified random sampling approach, selecting an equal number of InR and OoR patients for each CKD stage. Specifically, for each stage, we included approximately 130 InR patients and 130 OoR patients, maintaining a 1:1 ratio. This balanced sampling strategy resulted in approximately 260 samples per CKD stage, yielding a final dataset of around 1,300 samples that fulfilled all test feature inclusion criteria. We provide approximate figures due to the SAIL output policy, which restricts the disclosure of exact numbers [23]. All patient/sample numbers are approximate. Our complete dataset (CD) includes all tests: age, gender, creatinine, albumin, sodium, potassium, urea, wbcc, rbcc, albumin in urine and uACR. We consider this dataset for training and testing all models for multi-stage CKD classification.

**Subsets with Specified Features (S-1/2/3):** Recognizing that not all patients had the full complement of tests available at the same time, we created 3 different subsets to maximize data usage and allow focused analyses. Subset-1 (S-1), consisting of creatinine, age, and gender, produced about 30K labelled samples. Subset-2 (S-2), containing albumin, potassium, sodium, urea, rbcc, and wbcc, yielded roughly 15K samples. Finally, Subset-3 (S-3) comprising of albumin in urine and uACR tests, involved around 1.8K samples. These subsets were designed to facilitate a deeper investigation of specific laboratory parameters relevant to CKD classification. We consider the three subsets for training local models with vast data but specific features.

TABLE I
CREATED DATASETS. CD:COMPLETE DATASET, S:SUBSET

| Name | Total Size | InR Size | OoR Size |
|---|---|---|---|
| CD | 1.3K | 650 | 650 |
| S-1 | 30K | 15K | 15K |
| S-2 | 15K | 7.5K | 7.5K |
| S-3 | 1.8K | 900 | 900 |

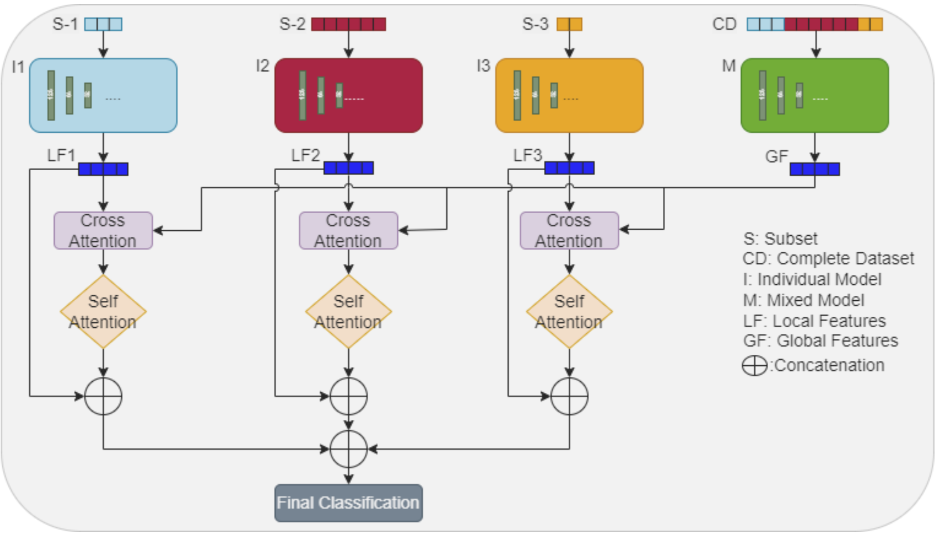

Fig. 1. Architecture Overview. The model consists of three individual models (I1, I2, I3) extracting local features (LF1, LF2, LF3) and one mixed model (M) extracting global features (GF). Local and global features are fused through cross-attention and self-attention mechanisms, followed by concatenation for final classification.

## IV. Methodology

*1) Existing Methods:* In this study we employ traditional machine learning models as our benchmark, which include Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. The four models were chosen as baseline approaches due to their proven effectiveness in classification tasks and their ability to capture different patterns within the data. These models provide valuable insights into the performance achievable when utilizing the complete feature set (trained on Complete Dataset), offering baseline results for comparison. However, training models on the complete dataset presented limitations due to the reduced sample size when all 11 features were used together. To overcome this, we maximize the use of sample data and features in the proposed method below.

*2) Our Proposed Method:* Our methodology addresses the challenge of limited, yet fully comprehensive CKD data by incorporating both large partial datasets and a smaller complete dataset. The overall pipeline includes three stages, Stage-1) we train individual models (I-models) on feature-specific subsets, Stage-2) we train a Mixed model on the smaller but fully multivariate dataset, Stage-3) finally an ensemble model fuses local (I-models) and global (Mixed model) representations via attention mechanisms for multi-stage CKD classification. This approach provides a global context by ensuring that local feature-specific insights are systematically integrated with complete multivariate patterns, maximizing predictive power despite limited fully comprehensive data.

**Stage-1:** We first train three separate individual models, each using a feedforward architecture designed to match its subset of features:

- I-Model1 (I1): The network has two fully connected layers with 32 and 16 units, each followed by ReLU activation and a dropout rate of 0.2. The 16-dimensional layer output acts as the local embedding. The S-1 was used to train I1.
- I-Model2 (I2): The network consists of three fully connected layers with 64, 32, and 16 units (ReLU + dropout = 0.2 after each). The 16-dimensional layer provides its local embedding. The S-2 was employed to train I2.
- I-Model3 (I3): The network has two-layer structure (32, then 16 units), with ReLU activation and 0.2 dropout per layer, culminating in a 16-dimensional embedding. The S-3 was utilized to train I3.

**Stage-2:** The Mixed Model (M) incorporates four dense layers with 128, 64, 32, and 16 units, each followed by ReLU activation and 0.2 dropout to reduce overfitting risk. The 16-dimensional layer outputs a “global embedding” that represents holistic interactions across all features. Despite the smaller dataset, this global perspective complements the specialized local embeddings, forming the foundation for the subsequent fusion process. The CD was used to train M.

**Stage-3:** In our ensemble model, each local embedding (query) from the individual models undergoes a cross-attention operation with the global embedding (key/value) generated by the Mixed Model. This mechanism allows the local features to incorporate global context, enhancing their representational power.

Following this, each cross-attended local embedding is further refined through a self-attention block, which captures internal correlations within the local representation. To preserve the original information, we apply a skip connection that adds the initial local embedding (from the corresponding individual model) back to the refined embedding. This ensures that the base representation is maintained while being enriched with global context. The resulting enriched embeddings are then concatenated to form a unified fused representation, which is then passed through a final classification module, consisting of two dense layers with 32 and 16 units respectively, each followed by ReLU activation and a 0.2 dropout rate. The final output layer contains 5 units corresponding to the CKD classification classes. This design ensures that each local model's specialized knowledge, enriched by global interactions, contributes effectively to the final five-class CKD prediction within a unified architecture.

During this stage, the individual models and mixed model are fine-tuned with the whole ensemble architecture using the CD. This allows the whole architecture to simultaneously enhance feature representation from each of the components through the interactions from fusion and attention mechanisms.

The architectural design choices of our individual and mixed models vary based on their input feature dimensions. Deeper networks, with more layers, are employed for models handling larger inputs (i.e., Individual Model, I2, with 3 layers and the Mixed Model with 4 layers, versus 2 layers in I1 and I3). This structure enables progressive dimensionality reduction, ensuring all models output a consistent 16-dimensional feature embedding that are then fused in the subsequent attention stages and concatenation.

*3) Implementation Details:* For all experiments, including the individual models, the mixed model, and our proposed architecture, we applied a consistent training setup. The Complete Dataset (CD) was split into 80% for training and 20% for testing. All features were normalized prior to model training to ensure uniformity across inputs. We employed the cross-entropy loss function for classification tasks and trained each model for 100 epochs with a batch size of 64. The learning rate was set to 0.001 for all models. This standardized configuration allowed for a fair comparison across all models.

## V. Results

We evaluated our CKD five-class classification approach by comparing it against a range of classical machine learning techniques as well as various configurations of our proposed attention-based model. The classical methods included Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. These allowed us to establish a clear baseline before introducing our novel architecture. We further conducted an ablation study on our proposed method to measure the individual and combined impacts of multiple input streams (I1, I2, I3) and different forms of attention (Cross and Self-Attention). Following the SAIL Output Review Policy, we cannot share the confusion matrices because some cells contain five or fewer cases, which could potentially expose sensitive patient information [23].

Table II reports the performance of four traditional machine learning (ML) algorithms—DT, SVM, RF, and XGBoost—on the (InR + OoR) test set under a 5-fold cross-validation scheme. Among these models, XGBoost yields the highest accuracy (67.56%), accompanied by the highest precision and recall (both 67.56%), as well as an F-measure of 67.46%. Random Forest follows closely with an accuracy of 64.12% and an F-measure of 64.06%. In contrast, Decision Tree exhibits the lowest accuracy (58.02%), although its recall (58.01%) is comparable to its precision (58.63%). The SVM model achieves moderate performance in terms of accuracy (61.07%) and recall (61.07%), with a precision of 60.96%. Specificity values across models range from 89.50% (DT) to 91.89% (XGBoost), indicating that even the least specific model demonstrates a relatively strong capability to correctly identify negative cases.

In Table II, we present results on the OoR test set, again using 5-fold cross-validation. The overall trends are consistent, with XGBoost outperforming the other algorithms in terms of accuracy (65.65%), recall (65.73%), and F-measure (65.40%), while also exhibiting the highest specificity (91.41%). RF ranks second, registering an accuracy of 61.83% and an F-measure of 61.73%. Meanwhile, SVM shows moderate results (57.25% accuracy) and DT performs the weakest overall, with an accuracy of 53.44% and a recall of 53.42%. Nonetheless, DT's specificity (88.35%) is not substantially lower than the other methods, suggesting some robustness in identifying negative cases despite its lower overall accuracy.

In order to leverage XGBoost's inherent ability to handle missing values, we combined our subsets—causing some null values to appear—into a single training set. Combining our subsets, we obtain approximately 32K samples with some null values for some features. As shown in Table III (under a 5-fold cross-validation scheme), when tested on the (InR + OoR) set, this approach yielded an accuracy of 77.86%, with a precision of 78.14% and a recall of 77.87%, culminating in an F-measure of 77.91%. Moreover, the model's specificity reached 94.46%, indicating robust performance in correctly identifying negative instances. Similarly, on the OoR test set (Table III), XGBoost achieved 75.76% accuracy, 76.13% precision, and 75.75% recall, for an overall F-measure of 75.77% and a specificity of 93.94%. These results suggest that allowing XGBoost to manage missing values directly can bolster classification performance by enabling the use of a larger dataset, without discarding records that lack certain features.

Table IV present the results for three individual models (I1, I2, and I3) and the Mixed model on the (InR + OoR) and OoR test sets, respectively. I1 achieves the highest accuracy, reaching 77.48% on (InR + OoR) and 74.05% on OoR, which may be attributed to its larger training sample (creatinine, age, and gender). By contrast, I2 shows the lowest accuracy—66.41% and 63.36%, respectively. I3, trained with albumin in urine and uACR, attains moderate accuracy (68.32% on (InR + OoR) and 65.65% on OoR), suggesting strong predictive power from these two features despite a smaller dataset. Lastly, while the Mixed model uses all 11 features, it operates with fewer overall samples, resulting in 69.47% and 67.94% accuracy, respectively.

Tables V summarize an ablation study evaluating different configurations of our Individual (I1, I2, I3) and Mixed (M) models along with Cross-Attention (C) and Self-Attention (S) modules. Each row represents a distinct combination of models (I1, I2, I3, and/or M) and attention strategies (CS).

TABLE II

Classic ML Models Results, Trained on CD dataset, Results on the (InR + OoR) / OoR test set (5-Fold)

| Models | Test Set | Accuracy | Precision | Recall | F-Measure | Specificity |
|---|---|---|---|---|---|---|
| DT | InR + OoR (CD) | 58.02 | 58.63 | 58.01 | 58.19 | 89.50 |
| | OoR (CD) | 53.44 | 55.29 | 53.42 | 54.04 | 88.35 |
| SVM | InR + OoR (CD) | 61.07 | 60.96 | 61.07 | 60.99 | 90.27 |
| | OoR (CD) | 57.25 | 57.75 | 57.24 | 57.37 | 89.31 |
| RF | InR + OoR (CD) | 64.12 | 64.17 | 64.13 | 64.06 | 91.03 |
| | OoR (CD) | 61.83 | 61.77 | 61.85 | 61.73 | 90.46 |
| XGBoost | InR + OoR (CD) | 67.56 | 67.56 | 67.56 | 67.46 | 91.89 |
| | OoR (CD) | 65.65 | 65.43 | 65.73 | 65.40 | 91.41 |

TABLE III

XGBoost trained on combined subsets (including some null values), Results on the (InR + OoR) / OoR test set (5-Fold)

| Models | Test Set | Accuracy | Precision | Recall | F-Measure | Specificity |
|---|---|---|---|---|---|---|
| XGBoost | InR + OoR (CD) | 77.86 | 78.14 | 77.87 | 77.91 | 94.46 |
| | OoR (CD) | 75.76 | 76.13 | 75.75 | 75.77 | 93.94 |

TABLE IV

Individual Models and Mixed Model Results on the (InR + OoR) / OoR test set (5-Fold)

| Models | Training Set | Test Set | Accuracy | Precision | Recall | F-Measure | Specificity |
|---|---|---|---|---|---|---|---|
| I1 | S-1 | InR + OoR (CD) | 77.48 | 77.45 | 77.48 | 77.43 | 94.37 |
| | | OoR (CD) | 74.05 | 74.22 | 74.07 | 73.83 | 93.51 |
| I2 | S-2 | InR + OoR (CD) | 66.41 | 66.46 | 66.42 | 66.43 | 91.60 |
| | | OoR (CD) | 63.36 | 63.53 | 63.30 | 63.11 | 90.84 |
| I3 | S-3 | InR + OoR (CD) | 68.32 | 68.33 | 68.33 | 68.32 | 92.08 |
| | | OoR (CD) | 65.65 | 68.85 | 65.61 | 65.51 | 91.42 |
| Mixed | CD | InR + OoR (CD) | 69.47 | 69.45 | 69.47 | 69.37 | 92.37 |
| | | OoR (CD) | 67.94 | 67.97 | 68.01 | 67.71 | 91.99 |

TABLE V

Ablation Study (5-Fold). All Models are Trained on CD dataset. I:Individual,M:Mixed, C:Cross-Attention, S:Self-Attention

| Models | Test Set | Accuracy | Precision | Recall | F-Measure | Specificity |
|---|---|---|---|---|---|---|
| I1+M+CS | InR + OoR (CD) | 75.19 | 75.21 | 75.19 | 75.18 | 93.80 |
| | OoR (CD) | 72.52 | 72.96 | 72.51 | 72.50 | 93.13 |
| I2+M+CS | InR + OoR (CD) | 67.94 | 67.97 | 67.94 | 67.94 | 91.98 |
| | OoR (CD) | 64.89 | 65.17 | 64.81 | 64.71 | 91.22 |
| I3+M+CS | InR + OoR (CD) | 69.47 | 69.46 | 69.46 | 69.45 | 92.37 |
| | OoR (CD) | 66.41 | 66.65 | 66.35 | 66.31 | 91.60 |
| I1+I2+M+CS | InR + OoR (CD) | 75.94 | 75.99 | 75.96 | 75.96 | 93.99 |
| | OoR (CD) | 74.05 | 74.44 | 74.05 | 74.01 | 93.51 |
| I1+I3+M+CS | InR + OoR (CD) | 77.10 | 77.18 | 77.10 | 77.11 | 94.27 |
| | OoR (CD) | 75.57 | 75.88 | 75.56 | 75.57 | 93.89 |
| I2+I3+M+CS | InR + OoR (CD) | 70.23 | 70.25 | 70.23 | 70.23 | 92.56 |
| | OoR (CD) | 66.67 | 66.90 | 66.64 | 66.65 | 91.67 |
| I1+I2+I3 | InR + OoR (CD) | 72.52 | 72.80 | 72.52 | 72.56 | 93.13 |
| | OoR (CD) | 70.23 | 70.76 | 70.23 | 70.21 | 92.55 |
| I1+I2+I3+M | InR + OoR (CD) | 73.28 | 73.59 | 73.29 | 73.34 | 93.32 |
| | OoR (CD) | 71.76 | 72.18 | 71.77 | 71.76 | 92.94 |
| I1+I2+I3+M+CS | InR + OoR (CD) | **80.15** | **80.15** | **80.16** | **80.14** | **95.04** |
| | OoR (CD) | **77.10** | **77.25** | **77.09** | **76.85** | **94.27** |

On the (InR + OoR) test set (Table V), the best-performing configuration is I1+I2+I3+M+CS, achieving 80.15% accuracy and 80.14% F-measure, with a specificity of 95.04%. This indicates that aggregating all individual models plus the Mixed model and stacking both cross- and self-attention modules yields a notable performance gain.

A similar trend emerges on the OoR test set (Table V), where I1+I2+I3+M+CS again attains the top accuracy of 77.10%, along with a precision of 77.25%, showing the advantage of combining all feature subsets and attention mechanisms. In contrast, using only one or two of the individual models (e.g., I2+M+CS) shows lower accuracy (64.89%) on the same set. Overall, these findings underscore that leveraging multiple feature subsets and attention strategies jointly can significantly enhance classification performance.

## VI. Conclusion

In this work, we explored a multi-branch modeling strategy for five-class CKD classification using both individual and mixed feature subsets. Our experiments show that individual models trained on larger or highly informative feature subsets can outperform models relying on smaller data sets, underscoring the importance of data availability and feature quality. At the same time, the Mixed model, while constrained by fewer overall samples, benefits from leveraging all features simultaneously. By integrating pretrained Individual and Mixed models through cross-attention and self-attention, and then fusing these representations, we effectively harness both local and global information. Furthermore, our ablation studies highlight the advantages of combining multiple feature sources with attention mechanisms, consistently yielding strong gains in accuracy, recall, and specificity. Overall, these findings affirm that a carefully orchestrated balance of feature selection, data utilization, and attention-based fusion can substantially enhance predictive performance for CKD classification.

Our study is limited in the following areas: Firstly, validation was performed using the SAIL Databank, primarily covering the Welsh population. This specific demographic and healthcare context may contain biases limiting the generalizability of our approach to other regions. Secondly, whilst our method is flexible, the requirement for a minimum complete dataset for the Mixed Model component might still be challenging to meet in some real-world applications. We plan to address these limitations in future research.

## ACKNOWLEDGMENT

Ali Guran received funding from the Turkish Ministry of National Education (Republic of Türkiye) through the Postgraduate Study Abroad Program. Gary Tam was supported by the International Mobility Award [62] and the Collaboration and Knowledge Exchange Support [92S], both provided by the CHERISH-DE Centre (EP/M022722/1). This work is partly supported by the EPSRC National Edge AI Hub (EP/Y007697/1). This project also benefited in part from a fee waiver fund provided by SAIL. This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research. The responsibility for the interpretation of the information we supplied is the authors' alone. All research conducted has been completed under the permission and approval of the SAIL independent Information Governance Review Panel (IGRP) project number 1220. For Open Access purposes, the author has applied a CC BY copyright licence to any Author Accepted Manuscript version resulting from this submission.

## References

[1] P. E. Stevens, S. B. Ahmed, J. J. Carrero, B. Foster, A. Francis, R. K. Hall, W. G. Herrington, G. Hill, L. A. Inker, R. Kazancıoğlu *et al.*, "Kdigo 2024 clinical practice guideline for the evaluation and management of chronic kidney disease," *Kidney international*, vol. 105, no. 4, pp. S117–S314, 2024.

[2] A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in *IEEE congress on evolutionary computation (CEC)*, 2018, pp. 1–9.

[3] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *IEEE International Conference on Healthcare Informatics (ICHI)*, 2016, pp. 262–270.

[4] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," *IEEE Access*, vol. 8, pp. 55 012–55 022, 2020.

[5] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019.

[6] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, p. 109, 2022.

[7] H. Ilyas, S. Ali, M. Ponum, O. Hasan, M. T. Mahmood, M. Iftikhar, and M. H. Malik, "Chronic kidney disease diagnosis using decision tree algorithms," *BMC nephrology*, vol. 22, no. 1, p. 273, 2021.

[8] R. A. Hughes, J. Heron, J. A. Sterne, and K. Tilling, "Accounting for missing data in statistical analyses: multiple imputation is not always the answer," *International journal of epidemiology*, vol. 48, no. 4, pp. 1294–1304, 2019.

[9] F. P. Schena, V. W. Anelli, D. I. Abbrescia, and T. Di Noia, "Prediction of chronic kidney disease and its progression by artificial intelligence algorithms," *Journal of Nephrology*, vol. 35, no. 8, pp. 1953–1971, 2022.

[10] M. Afkanpour, E. Hosseinzadeh, and H. Tabesh, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 188, 2024.

[11] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multimodal learning with missing modality via shared-specific feature modelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 878–15 887.

[12] L. C. Zwiers, D. E. Grobbee, A. Uijl, and D. S. Ong, "Federated learning as a smart tool for research on infectious diseases," *BMC Infectious Diseases*, vol. 24, no. 1, p. 1327, 2024.

[13] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," *arXiv preprint arXiv:2409.07825*, 2024.

[14] J. Guo and J. Ye, "Anchors bring ease: An embarrassingly simple approach to partial multi-view clustering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 118–125.

[15] X. Yuan, H. Wang, Z. Sun, C. Zhou, S. C. Chu, J. Bu, and N. Shen, "Anchored-fusion enables targeted fusion search in bulk and single-cell rna sequencing data," *Cell Reports Methods*, vol. 4, no. 3, 2024.

[16] SAIL, "Welsh longitudinal general practice dataset (wlgp) - welsh primary care version 21.0.0," 2021.

[17] ——, "Welsh results reports service (wrrs) version 7.0.0," 2021.

[18] D. V. Ford, K. H. Jones, J.-P. Verplancke, R. A. Lyons, G. John, G. Brown, C. J. Brooks, S. Thompson, O. Bodger, T. Couch *et al.*, "The sail databank: building a national architecture for e-health research and evaluation," *BMC health services research*, vol. 9, pp. 1–12, 2009.

[19] K. H. Jones, D. V. Ford, C. Jones, R. Dsilva, S. Thompson, C. J. Brooks, M. L. Heaven, D. S. Thayer, C. L. McNerney, and R. A. Lyons, "A case study of the secure anonymous information linkage (sail) gateway: a privacy-protecting remote access system for health-related research and evaluation," *Journal of biomedical informatics*, vol. 50, pp. 196–204, 2014.

[20] R. A. Lyons, K. H. Jones, G. John, C. J. Brooks, J.-P. Verplancke, D. V. Ford, G. Brown, and K. Leake, "The sail databank: linking multiple health and social care datasets," *BMC medical informatics and decision making*, vol. 9, pp. 1–8, 2009.

[21] S. E. Rodgers, J. C. Demmler, R. Dsilva, and R. A. Lyons, "Protecting health data privacy while using residence-based environment and demographic data," *Health & place*, vol. 18, no. 2, pp. 209–217, 2012.

[22] S. E. Rodgers, R. A. Lyons, R. Dsilva, K. H. Jones, C. J. Brooks, D. V. Ford, G. John, and J.-P. Verplancke, "Residential anonymous linking fields (ralfs): a novel information infrastructure to study the interaction between the environment and individuals' health," *Journal of Public Health*, vol. 31, no. 4, pp. 582–588, 2009.

[23] SAIL Databank, "Sail-pol-024 output review policy," PDF, 2022. [Online]. Available: https://saildatabank.com/wp-content/uploads/2022/08/SAIL-POL-024-Output-Review-Policy-v1.2-3.pdf