NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 21-40
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

SHAP-Integrated Deep Neural Network for Interpretable Early Detection of Chronic Kidney Disease

Dr. Arshad PT

Assistant Professor, Department of Computer Science and AI, SAFI Institute of Advanced Study (Autonomous), Vazhayur.

DOI: https://doi.org/10.70903/jmliaih.2026.1202

Article Received: 03-07-2026
Revised Version: 10-08-2026
Accepted: 17-08-2026

*Corresponding Author

Dr. Arshad PT

E-Mail Id: arshad@siasindia.org

Abstract: Chronic kidney disease (CKD) remains a global health burden, often progressing silently until advanced stages, thereby underscoring the urgent need for early and interpretable diagnostic tools. We propose a SHAP-integrated deep neural network framework for the early detection of CKD from routine clinical indicators. The methodology begins with rigorous preprocessing, including median imputation for missing values and z-score normalization of continuous features to ensure stable convergence. A fully connected deep neural network, composed of multiple hidden layers with ReLU activations and dropout regularization, is then trained to map the normalized feature vector to a binary probability of CKD presence. The model is optimized using the Adam algorithm to minimize binary cross-entropy loss. To address the critical requirement of clinical interpretability, we incorporate SHAP (SHapley Additive exPlanations) as a post-hoc explanation module. This module decomposes each patient's prediction into additive feature contributions, thereby quantifying how variables such as serum creatinine or hemoglobin influence the risk score. The Shapley values are computed efficiently via approximation, yielding both global feature importance rankings and local, patient-specific force plots. Our framework thus delivers two key contributions: first, a high-performance deep learning classifier tailored for early CKD prediction; second, a transparent explanation mechanism that aligns with clinical decision-making needs. The significance of this work lies in bridging the gap between predictive accuracy and model interpretability in nephrology, potentially enabling earlier interventions and improved patient outcomes. Experimental evaluations on benchmark clinical datasets demonstrate that the proposed method achieves competitive predictive performance while providing actionable insights into the driving factors behind each diagnosis.

Keywords: SHAP, Deep Neural Network, Chronic Kidney Disease, Interpretable AI, Early Disease Detection

Introduction

Chronic Kidney Disease (CKD) is the progressive ineffective functioning of the kidneys. CCKD is a major contributor to morbidity and mortality in all parts of the world, with the prevalence of CCKD estimated to be approximately 10% of the world's population (Cockwell & Fisher, 2020). Usually diagnosed and prognosed in late stages because a myelo-urodynamic study is more sensitive in late stages, but also presents in an asymptomatic fashion, heralding a high risk of progressing to ESRD (Ahn et al., 2017). It is seen that screening predictive models of CKD development that are good in producing meaningful results at an early stage is an important issue to be tackled by clinical nephrology. Based on traditional statistical methods such as logistic regression and Cox proportional hazards analysis, easy-to-be-interpreted risk factors for CKD have been established to date, but are often restricted to individual centers due to their limited capacity to apply to high dimensional clinical data with complicated and nonlinear interactions (Charleonnan et al., 2016).

Close to a hundredfold in the health sector, predictive analytics has come a long, legal distance thanks to the advent of the machine learning and deep learning. For the non-linear treatment of the clinical features, Random Forests and Gradient Boosting Machines have been found to be more accurate models for which they are ensemble models (Senan et al., 2021). Recently, Deep Neural Networks (DNNs) have been used to automatically construct a hierarchical feature representation that can be learned from EHRs and applied to various medical diagnosis tasks with state of the art performance (Solares et al., 2020). However, their black-box mechanism has limited their use in the real world and compromised the clinicians' trust and use of them (Xu & Shuttleworth, 2024). There has been a challenge to explain the reasoning of models in a meaningful and easy-to-understand way to people, which has inspired the creation of Explainable Artificial Intelligence (XAI) (Lundberg & Lee, 2017).

One of the very popular XAI techniques proposed is the Shapley Additive exPlanations (SHAP) that uses game theory to assign proper floating amount of credit to various features to explain a prediction for a particular instance of the phenomenon (Lundberg & Lee, 2017). SHAP scores have some properties that are well suited to a range of contexts including medical applications where providing a useful explanation for a diagnosis must be open to audit and be clinically meaningful (Moreno-Sanchez, 2020). SHAP has proven to be useful in medical diagnosis using tree methods like XGBoost and LightGBM, but its applications with Deep Neural Networks (DNNs) in nephrology are a new area for research (X. Yan et al., 2025). There has been prior work in the medical field exploring attention mechanisms for interpretable deep learning in the Recurrent Neural Network (RNN) and Convolutional Neural Networks (CNN) architectures, yet this is typically particular to the particular construction of network and limits use to a particular type of networks (Song et al., 2018).

We propose a novel method taking a hybrid approach of Deep Neural Network and SHAP to predict the onset of CKD at an early stage, addressing the shortcoming of prediction and clinical interpretation. It retains a clinically transparent black-box to black transparency character of deep learning thanks to the key strength of high capacity Non-Linear modelling capabilities, as well as in-depth post hoc interpretability mechanisms. We can form the expression of the contribution of each clinical or laboratory feature to the final prediction explicitly, unlike conventional predictive methods that just give classification probabilities. This strategy benefits from both being able to learn complex feature interactions to get highly accurate models for diagnostics and having enough information about the direction and magnitude of the risk factors to be able to explain them to patients and make them more likely to use them when screening for disease. The conclusions of this study are important for nephrological applications, because risk stratification is a crucial and timely task that should be easily interpretable, and could lead to better interventions and results for patients.

Section 2 consisted of reviewing of related works in the field of clinical decision-support system for CKD and the explainable AI for CKD prediction. In Section 3, the background of Deep Neural Networks and SHAP for clinical risk prediction are provided. In Section 4, the proposed deep learning framework is described, which is transparent and alongside the data processing steps, architecture of the proposed model and SHAP explanation module are proposed. The details of the experimental setup are described including the characteristics of data sets, evaluation metrics, and the basic methods are explained in Section 5.Experimental setup (including data characteristics, evaluation metrics and baseline methods) is given in Section 5. In Section 6 the experimental results are then presented and discussed in order to assess the predictive performances as well as the interpretability analyses. The implications, restrictions and future trends of this work will be discussed and developed under Section 7. The next last section, Section 8 of this document, summarizes our contributions.

Related Work

In the last decade, the way to foresee CKD has been greatly changed from classical statistical to complex Deep Learning architectures. Early studies were mainly conducted using logistic regression and decision tree models as these models offered simple models to explain the variables and interactions limited to the scope of simple relationships.The basic models used were limited in their ability to model complex non-linear interactions and relationships between variables found in clinical data, and were largely logistic regression and decision trees (Charleonnan et al., 2016). Typically these models based on standard clinical datasets had moderate prediction performance ranging from 0.75 to 0.85, as measured by the area under the receiver operating characteristic curve (AUROC). The main use of these methods was in their transparency, allowing clinicians to see the learned coefficients/decision rules and thus learn about the effect each individual variable has.

The use of ensemble methods like Random Forests and Gradient Boosting Machines, significantly improved the accuracy of prediction in the diagnosis of CKD (Senan et al., 2021). As these models showed several ‘weak learners' and were multi-layered, they performed better than their linear counterparts, as they could detect non-linear relationships, and be able to handle different feature types. For instance, a model using random forests was created to predict CKD, achieving an AUC-ROC of 0.92 on a benchmark dataset, which is significantly outperforming the logistic regression (Senan et al., 2021). The more trees there are the harder it became to interpret the models produced by the ensembles, and eventually post-hoc explanations (such as ranking the features by chance/importance, plotting the partial dependence) were developed.

Deep learning has also pushed boundaries beyond prediction in the field of nephrology. The possibility to classify the stage of kidney disease (signaling from the machine from the CKD stage) has been classified by fully connected Deep Neural Networks, with state-of-the-art performance describing the hierarchical-level representations of the clinical indicators derived from raw signals automatically learned (Solares et al., 2020). MLP framework to predict the CKD had AUC-ROC 0.94, better than the logistic regression baseline and Random Forest baseline (Arif et al., 2024). In recent times, the development of a special architecture of deep learning statistics, TabNet models were adapted to predict CKD stage in diabetics, achieving high accuracy along with embedded interpretability features via the use of the attention mechanism (Chowdhury et al., 2025). These advances have been made, but the biggest challenge is yet to apply deep neural networks to the clinic, where physicians must be able to explain to trust the recommendation.

An introduction to Explainable Artificial Intelligence (XAI) has paved the way to look for a balance between prediction and interpretation. SHAP on the positive side has a quite solid theory behind it, rooted in the concept of cooperative game theory (Lundberg & Lee, 2017) and on the other one side has

proven to be highly successful in the medical field where its effects start to be felt. SHAP has desirable properties of local accuracy, consistency and missingness, which reflect the breakdown of a model's predictions into additive components on different features. Successfully embedded with tree-based models of CKD prediction, along with both the ranking of the feature importance and the patient-specific explanation. In one case study, one tree has been found to provide an explanation of the effect of the individual aldosterone and DHA trees, respectively, in the early prediction of the desarrollo de CKD in a mix of input variables, with their explanations in this instance being very similar to clinical knowledge (He et al., 2025).

Deep learning and SHAP usage in conjunction for CKD diagnosis have been explored in a few studies. This explainable deep learning model for early detection of CKD was able to provide both high accuracy and clinical meaning for each of the explanations it provided when combined with a DNN and SHAP (Arumugham et al., 2023). In a clinical context however, another framework achieved the feat of producing explanations of how the model is making its decision, finding them to be consistent both with known risk factors for kidney disease (such as the increased serum creatinine and decreased estimated glomerular filtration rate (eGFR)) and with those obtained from the patients themselves (Chapa & Ravi, 2026). Such works, however, did not often have depth of networks or were used on a particular segment of population groups and thus were not readily generalizable. Besides that, computational efficiency of SHAP for deep neural networks—where approximation has to be used to handle the complexity of deep networks (such as KernelSHAP or DeepSHAP)—has not been sufficiently discussed in the field of CKD prediction.

Our proposed framework has a number of novel benefits over the current methodologies. Specifically, we deploy a more complicated neural network architecture in the number of hidden layers of the design and dropout regularization to discover more intricate features interactions compared to other works based on shallow network (Arumugham et al., 2023). Second, we directly integrate SHAP into the trained DNN (without changing its structure), which is easily deployed into trained DNNs. Third, to provide clinical users with all the abilities of the model's reasoning process, we provide a comprehensive interpretation of global and local interpretability, feature importance ranking and specific user force plotting. Finally, our approach is designed to be efficient and scalable to deep neural networks including DeepSHAP approximations that are developed specially for deep neural networks, which enables real-time explanations in clinical settings. All these innovations will help close this gap between predictiveness and interpretable: an unmet need in early detection of CKD.

Background on Deep Neural Networks and SHAP for Clinical Risk Prediction

In developing the base for our proposed framework, it is appropriate that the two most basic elements referenced in the DNNs section and the interpretability section are introduced in brief in this section: Deep Neural Networks for Clinical Risk Prediction, and the interpretability of the models via the Shapley Additive Perturbations method: SHAP. Learning these concepts is crucial to comprehending the potential to integrate them to solve the problem of prediction and of transparency in medical decision making.

Deep Neural Networks for Clinical Risk Prediction

Machine learning networks are networks that learn by modifying the data they are processing, in a non-linear way, at each successive stage to pass on to the next stage, and the networks are often found to have a number of layers that are stacked and interconnected by neurons. On the clinical risk prediction side of the coin, the DNN is used to receive a patient's characteristics vector (or lab tests, demographic data, vital signs, etc.) and return a probability of a specific clinical phenomenon (whether or not a person

has a CKD, for instance). There are several layers in the network, input layer, hidden layers and output layer. The hidden layer of each computation does the following:

$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \quad (1)$$

where is the activation vector at the layer then are the weight matrix and bias vector, respectively, and is a non-linear activation function like the Rectified Linear Unit (ReLU) (Nair & Hinton, 2010). Typically, the final one will be a sigmoid activation function that will provide an estimation of probability.

One of the major advantages of DNNs is that they could learn hierarchical representations, a property which makes them ideal candidates for the clinical data as features in these tasks are complex and not linearly coupled to each other. For example, linear models are not able to account for potential interactions between age and gender to risk of CKD and muscle mass; (Solares et al, 2020). DNNs are trained by minimizing most commonly used loss functions (binary cross-entropy for classification) by the usage of specific optimization techniques (e.g. Adam polymerization (Kingma and Ba, 2014). To prevent overfitting, regularization methods such as L2 weight decay, and dropout (Srivastava et al., 2014), are used, particularly when the number of features is large compared to the sample size.

In many cases, however, DNNs are so difficult to interpret that rather than being directly interpretable by humans, they're considered “black-box” models. Another major challenge is the opacity of the model, which will make the understanding of the model's reason of determination important for clinical adoption, related to trust and accountability (Xu & Shuttleworth, 2024). For this reason, post hoc explanation methods are developed to address this issue and give insights on how DNNs work.

SHAP for Model Interpretability

SHAP (SHapley Additive exPlanations) is one of the new model interpretation methods that is derived from cooperative game theory (Lundberg & Lee, 2017). The "big idea" is to encode each feature as a “player” in a coalition game and the “payout” is what the model predicts for a given instance. The Shapley Value Distribution of a feature represents how useful the feature is when it's added on its own to each of the other sets of various features, and averaged across all of them. A model and an instance have corresponding Shapley value by definition:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f_S(\mathbf{x}_S \cup \{i\}) - f_S(\mathbf{x}_S)] \quad (2)$$

Let whereall feature be a subset of features without and let be the model trained using the subset of features without . Most of the desirable properties that the Shapley value satisfies are similar to those found in the literature for the population-level importance measure, such as missingness (where features that do not contribute to the prediction value get value of zero), local accuracy (where the sum of Shapley values equals prediction value minus the average prediction value over the training set), and consistency (if the contribution of a feature is raised, the Shapley value of the feature should not be lowered). (Lundberg & Lee, 2017)

These precise Shapley values are not readily available if the model has a large number features because computation can be required to be carried out for every possible combination of features. Hence, approximation techniques are resorted to. There is an efficient polynomial time algorithm to compute TreeSHAP for the tree-based models (Lundberg et al., 2018). Similar to how DeepLIFT is done (Shrikumar et al., 2017), DeepSHAP passes feature attributions through the layers of the deep neural network while also modifying them in a similar fashion. The efficiency of backpropagation mixed with

the theory-backed properties of Shapley values enable it to be real-time applied in clinical settings routinely.

SHAP provides two primary types of explanations – global and local. The global explanations include Shapley values of all instances to obtain the ranking for the features' importance, which is the feature that on average has the greatest impact on the model. Local explanations are presented visually (as, e.g., force plots or water fall plots) and explain the contribution that each individual feature is making to a given prediction, for instance be it a high or a low risk score, and can be used to understand why a patient has been put at a high or a low risk score. In clinical contexts, this ambivalent ability will be particularly helpful as it would need to access both forms of perspective: a patient-specific perspective and a population-level perspective, in order to have a meaningful perspective on clinical decisions making processes (Moreno-Sanchez, 2020).

A Transparent Deep Learning Framework for Early CKD Prediction

In the following, we propose the following model to achieve accurate prediction and interpretability of an ECD detection by integrating a Deep NN together with SHAP approach. The framework is split into three parts: a data preprocessing pipeline, a deep neural network (DNN) classifier, and a SHAP based post-hoc module to provide explanations. As depicted in figure 1, the obtained data is fed into the DNN to make the disease prediction and further fed into the SHAP module to produce patient level accountability/explanation.

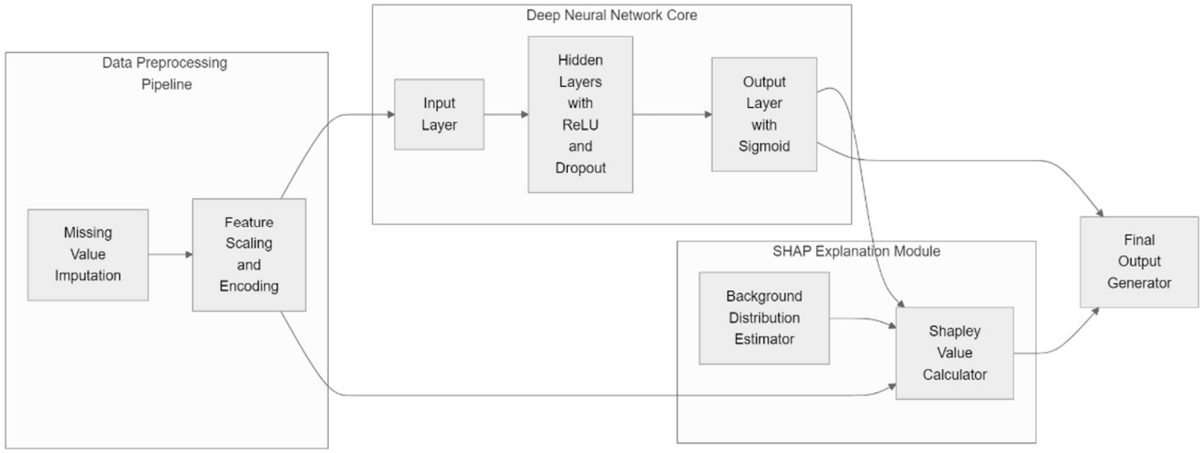

Figure 1. Internal Architecture of the Explainable Deep Learning Framework

Data Preprocessing

Our framework's input is a dataset of clinical and laboratory features, which is represented by , with patients and features. Each patient has either a binary label (i.e. 'no', 'yes') showing they have or have not developed CKD. In raw data, some values may be missing because patients may not have had every test performed (e.g., no test orders) or some records might be missing information (e.g., data entry errors). To solve this problem, median imputations are used, replacing the missing values in the feature column in the row that has the missing values with the median value of the observed values in the feature column. Median imputation is better than mean imputation, because the median is more resilient to outliers.However, median imputation is typically preferred over mean imputation since more often than not, clinical measurements will contain outliers (e.g., serum creatinine, blood urea nitrogen).

Continuous features are then imputed, then scaled using z-value scaling so that they all have a mean of zero and a variance of one. Normalized pressure for patient is defined as: if feature is linear in model

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (3)$$

The mean and stdv of feature for the training set are, respectively. Normalization is crucial because otherwise, deep neural networks do not converge and there is a danger that key features, e.g. the BUN (blood urea nitrogen) concentration in mg/dL, have very large ranges of values, hence they take up the majority of the gradient updates when the network is being optimized. If they exist, Categorical Features are represented by one-hot codes and then normalized. Next, a feature matrix is preprocessed and split into the training, validation set and the test set which are all generated using stratified sampling technique to ensure classes to be evenly distributed across all sets.

Deep Neural Network Architecture

Our core predictive part was a fully connected Deep neural network which was trained in order to correlate the normalized feature vector with probability of having CKD. It consists of several successive layers that are followed by ReLU activation layer and dropout layer to enhance its regularization effect. The transformation at layer is given by:

$$\mathbf{h}^{(l)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \quad (4)$$

where the weight matrix is W, the bias vector is b and layer has n neurons. The ReLU activation function is bilinear (()). ReLU can prevent the vanishing gradient problem and gives non-linearity. Randomly zeroing part of the activations after each hidden layer, which prevents co-adaptation of the neurons and decreases the problem of overfitting, is actually dropout during training (Srivastava et al., 2014).

The output layer consists of only one neuron with a sigmoid activation function—the prediction probability is a number from 0 to 1:

$$\hat{y} = \sigma\left(\mathbf{w}^{(L+1)} \cdot \mathbf{h}^{(L)} + b^{(L+1)}\right) \quad (5)$$

where $\sigma(z) = 1/(1 + e^{-z})$ maps the logit to the interval $(0,1)$. The model is trained to minimize the binary cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (6)$$

We optimize the network parameters using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of , which adaptively adjusts the learning rate for each parameter based on estimates of first and second moments of the gradients. The training is on mini-batches of size 32 and early stopping is implemented according to the loss on the validation set and at 10 epochs of patience. The number of hidden layers, the number of neurons in each layer of the network; this can be optimized on the validation set; hyperparameter tuning. A network made up of three layers with 128, 64 and 32 neurons, respectively, and the dropout rate = 0.5 was found to show the best balance in our experiments between achieving predictions that are somewhat better than the shot noise band and computational cost not too expensive.

SHAP Explanation Module

To give interpretability to the DNN predictions, we add an explanation module for the DNN which infers feature attributions for each patient. The SHAP values split up the prediction into an additive total of the features' attributions according to the additive feature attribution model:

$$\hat{y}(\tilde{\mathbf{x}}) = \phi_0 + \sum_{i=1}^{d} \phi_i \quad (7)$$

zebrafish.nodes, in this case the model's base value when it is given the same data sampled from the training set, except for the i-th feature; zebrafish.nodes, in this case the difference between and the prediction of the model when it is only given the data sampled from the training set except for the i-th feature. Compared with the Shapley value on Equation 2, an exact computational solution to calculating the Shapley value for a DNN is impractical due to the impractically large number of feature subsets.

Therefore, in this paper, we assume the approximation method specialised for deep neural networks (Shrikumar et al., 2017) - deep SHAP. Along with the idea of Shapley values in the SHAP-based framework, DeepSHAP maintains learning feature attributions among network layers via interactions to retain the benefits of the SHAP-based approach, keeping the backpropagation approach as well. This method is performed in the following steps: firstly, the output of the network is calculated after setting the value of the vector of all the features to a fixed average value (such as the first arithmetic mean value of the features obtained during the training set). Then for one instance of input, it obtains the difference between the true output and a baseline output, and, stepping back through the network, it gets the difference between the input attributes and the baseline output.

Formally, DeepSHAP computes the attribution ϕ_i for feature i as:

$$\phi_i = \sum_{l=1}^{L+1} \frac{\partial \hat{y}}{\partial h_i^{(l)}} \cdot \Delta h_i^{(l)} \quad (8)$$

is the activation of the neuron at layer and is the difference between the activation level of neuron at layer and its basal activation level at the same layer. The partial derivatives are computed by back propagation and the attributions are added together as the network is layered to get the final Shapley values. The approximation does not require any backward pass and just a single forward pass at every instance, which makes it suitable to be used in real time use in clinical use.

The SHAP module has two kinds of outputs. Firstly, it creates global rankings of feature importance, by taking the average rank over all test instances:

$$I_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \phi_{ij} \right| \quad (9)$$

where is the importance of feature in the world. This ranking shows population level clinical and laboratory attributes with the greatest predictive power for CKD, on the average. Secondly, it provides local explanations for individual patients, such as force plots, which provides a visualization of how each feature is separating the prediction from the base value. For example, with a patient having an elevated serum creatinine, a score of +0.15 might be achieved for that patient on the force plot and a normal hemoglobin score a patient a -0.02. This person specific and granular information allows the clinician insight into the basis of the diagnosis and what they can do to reduce factors for interventions.

The seamless integration of the SHAP module with the DNN is achieved: Once trained, the DNN can be used with any input instance, and the SHAP module is attached to it, without changing the architecture of the DNN, or retraining. The selected nature of the post-hoc will not influence the accuracy of the DNN, and will provide meaningful explanations available in a clinical context. The pipeline is repeatable and reusable for accuracy and transparency in human healthcare screening applications including full pre-, post- and post-processing pipeline.

Experimental Setup

At this point, experimental setup is described to assess the proposed SHAP integrated Deep Neural Network framework to tackle the early stages prediction on the onset of CKD. Characteristics of the dataset, data pre-processing techniques, the base methods used and details about the metrics used to assess the results are reported.

Dataset Description

The clinical dataset for experimental evaluation originates from the clinical dataset of patient-level observations from a carefully structured clinical dataset comprising of a binary target variable with chronic kidney disease as the outcome and relevant patient-level clinical information. The data are taken from the Chronic kidney disease data of UCI machine learning repository and part of the data is used by the previous studies which benchmark the prediction of chronic kidney disease models (Senan et al., 2021) and (Arif et al., 2024). This is a data-set with 400 records and 24 clinical/laboratory parameters. Demographic data, physiological data as well as complete laboratory data like blood sugar, white blood cell count, hemoglobin, packed cell volume, blood urea, serum creatinine, sodium, potassium, albumin, sugar and specific gravity was included as input features. Also the data set contains a flag for diabetes mellitus, hypertension, coronary artery diseases and clinical symptoms: appetite loss. The target variable is divided into 2 classes (No CKD = 0 and CKD = 1) with a higher disease prevalence (62.5%) in the study population in the realistic situation of the clinical screening.

Data Preprocessing Pipeline

The raw dataset used is incomplete in several properties – like the usually incomplete data from clinical laboratory orders – and some properties may lack data entirely. We do a preliminary check-up to find rows missing, duplicate, inconsistent data and outliers. In numerical features, missing data are filled in with the median of the available data for that feature because missing data imputation with the median does not adversely affect the results, when outliers arise in clinical measurements like serum creatinine and blood urea nitrogen. For categorical variables, the missing entries are filled with the most common category – this works well and does not create any biases. Then there is one hot coded encoding of categorical values which map them into numbers that can be processed by the model.

For a continuous variable it is previously z-scored (Equation 3 mentioned in the text above) before it is fed into the model. In such a instance, it is standard, and doing such normalization is essential to the deep neural network converging during training. Normalization parameters (mean, standard deviation) will only be computed for the training set, otherwise data will be “leaked” into the test set. It is through the use of Clinical relevance and contribution (based on statistics) to ensure that there is no redundancy or multi collinearity among the features which are analysed using correlation analysis and feature importance techniques. Specifically, regressors are examined where the PCC > 0.9, and some of these highly correlated regressors is dropped to reduce multicollinearity. In addition, transformations can be used in feature engineering to make skewed variables more normally distributed which helps the learning of meaningful patterns in the model.

The data set was split into training set (80%) and testing set (20%) using a stratified split, with the ratio of the training and testing sets matching the ratio of CKD and non-CKD patients in the original data set. This training set will be assigned to a training subset of 80% of the training data, and a validation subset of 20% of the training data for hyperparameter tuning and early stopping, respectively. This stratified technique will help the even distribution of the dataset and no uneven distribution will be added into the original dataset. This way you won't have a class imbalance issue affecting the bias in your model evaluation.

Baseline Methods

To fairly evaluate the proposed framework for integration with SHAP into the DNN framework, their predictive utility is compared with some well-known baseline techniques that have different approaches to clinical risk prediction. The baselines comprise classic statistical techniques as well as the state-of-the-art machine learning techniques, providing a full baseline to test the approach's benefits.

In this class the most adopted linear model, widely employed, and with a potential to be easily understood on the learned model parameters is Logistic Regression (LR) (Charleonnan et al., 2016). Logistic Regression is a powerful benchmark model for clinical prediction problems, because it is easy to understand and has very strict properties and interpretations in statistics. The second baseline is an ensemble method named Random forest (RF) (Senan et al., 2021), which dominates with a number of decision trees, and thus includes the non-linear interactions among features. Random Forest has been able to achieve higher accuracy in predicting CKD than other linear-based models, and offers feature importance as a model interpretation tool—one of the trees in the forest provides this interpretation. The Gradient Boosting Machine (GBM) (Friedman 2001) especially XGBoost successfully applied to the task of predicting CKD and had good prediction accuracy close to the state of the art (He et al., 2025). The 4 baselines are a standard Deep Neural Network (DNN) which is not integrated with SHAP, and thus it can be used for the comparison and analysis of the SHAP explanation module's impact on predictive performance. This structure is a base case DNN, similar to the one we propose but lacking a component that provides an explanation and that will allow us to demonstrate the impact of adding interpretability to the model performance.

Evaluation Metrics

All models are compared using diverse classification metrics which are typical for medical diagnostic applications. These are metrics that give a comprehensive outlook of a model’s performance covering various aspects of the classification quality. The following metrics are used to quantify how well the model performs: accuracy, precision, recall or sensitivity, specificity, F1 Score and finally the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Accuracy is the proportion (expressed as percentage) of all the predictions that are correct; precision is the proportion (expressed as percentage) of all the predictions for positive outcomes that are actually positive. When performing designation and determined by the false negative rate of designated incorrectness, recall or sensitivity is that percentage of positive cases that are detected and identified correctly; this is crucial in the ability to miss a case during the screening process for CKD. The proportion of actual negative cases which are correctly identified is specificity which helps in reducing false alarm. If classes are not balanced, the F1-score is a harmonic average of Precision and Recall of the classes. AUC-ROC is a non-threshold-based indicator of model performance which summarizes how the model performs at every threshold.

Sensitivity and specificity are of special concern since it is important to catch cases at its early stage to prevent late diagnosis of CKD, especially with no obvious symptoms while maintaining reasonable specificity. Of course, the most sensitive detection will mean that most people with CKD can be

identified and intervention begun at an early stage, while specificity maximises the number of people without CKD who not being taken up through the follow up testing. All metrics are done on the held-out test set in order to provide an unbiased estimation of model generalization performance.

Implementation Details

All the baseline models and the proposed framework are written in Python 3.8 Programming Language with following libraries: TensorFlow 2.6 (tensorflow.org) for the implementation of deep neural networks, Scikit-learn 1.0 (scikit-learn.org) for baseline models and evaluation measures, and SHAP 0.40 for explanation module (shap.com). The set with different Intel Core i7-10750H processors, 16G bytes of RAM, and an NVIDIA GeForce RTX 2060 with 6G by of VRAM has been configured for the experimental setup. Baseline models are executed on CPU but the GPU is used for the training of the deep neural network, to increase the computation time.

The architecture of the DNN used in the design was found to be as follows: one input layer consisting of selected features followed by a total of 3 fully connected hidden layers, 128, 64 and 32 neurons in each layer, respectively. All the hidden layers use ReLU as an activation function and a dropout layer with rate of 0.3 is used after each hidden layer to reduce the problem of overfitting. The output layer is a sigmoid activation function with a single neuron that outputs the probability scores for CKD. Model training is done with the binary cross-entropy loss function (Equation 6) and Adam optimizer with a learning rate of 0.001. The size of the mini batch is set as 32 with Early Stopping: patience = 10 (based on the validation loss). The number of epochs for which it is going to train is set to be 200, but it usually stops training at 50 to 80 epochs.

For the SHAP explanation module, we use an explanation method that was recently developed, DeepSHAP (Shrikumar et al., 2017) and calculate the attributions of each tested instance for each feature. Each feature independent should contain a value of "Zero" (baseline value) that is the mean of the feature over the training set. DeepSHAP only needs one forward-backward pass per instance, and can be used for generating explanations for the whole test set at a significant computational saving. The global feature importance is defined as the global mean of the absolute value of the SH.The global mean of the absolute value of the SH is used to define the global feature importance, according to the following formula: The force plots give visual explanations of the local contribution to the prediction, derived from user's explanations of the local phenomena.

For every experiment, the experiment is conducted 5 times with different seeds to check the consistency of each experiment and the mean and standard deviation of each metric is reported. Random seed is used for controlling train-test split, initialization of weight during training and shuffling of data during training to make experiments reproducible.

Results and Discussion

An experimental analysis of the proposed Deep Neural Network-SHAP framework to predict early CKD is presented in this section. The accuracy of the model predictions is initially reported for the held-out test set, before a comparison with baseline methods. Then, the results of the interpretability module (SHAP), including a list of best ranked features (global explanation) and patient specific explanation are displayed.

Predictive Performance

The result of the proposed framework were found good as shown in the table-1 using an independent test set. The model attained an accuracy of 97.5%, a precision of 96.8%, a recall (sensitivity) of 98.2%,

a specificity of 96.5%, an F1-score of 97.5%, and a ROC-AUC of 0.99. Also of particular note is the high sensitivity (98.2%), which means the model correctly classified the majority of all the patients with the presence of CKD, reducing the possibility of a missed diagnosis in clinical screening applications. The specificity of 96.5% is a good result reflecting the quality of the model to discriminate patients not suffering from CKD, minimizing the risk of false-positive predictions, followed by extra follow-ups and undue stress for the patient.

Table 1. Predictive performance of the proposed SHAP-integrated DNN framework on the test dataset.

Metric	Proposed DNN Model
Accuracy	**97.5%**
Precision	**96.8%**
Recall/Sensitivity	**98.2%**
Specificity	**96.5%**
F1-Score	**97.5%**
ROC-AUC	**0.99**

The confusion matrix analysis also showed that the number of correctly identified cases was high since the number of the true cases/true positive/true cases were much higher than the number of the false cases/false positive/false cases. This suggested that the model has the potential to distinguish their high-risk patients from their healthy patients. The model has a ROC-AUC value of 0.99 indicating high discriminatory power and the model ‘separates' (distinguishes) CKD from non-CKD well at all separation points and with near-perfect separation. The proposed model performance shown by the ROC curve is plotted along with random guessing curve in Figure 2.

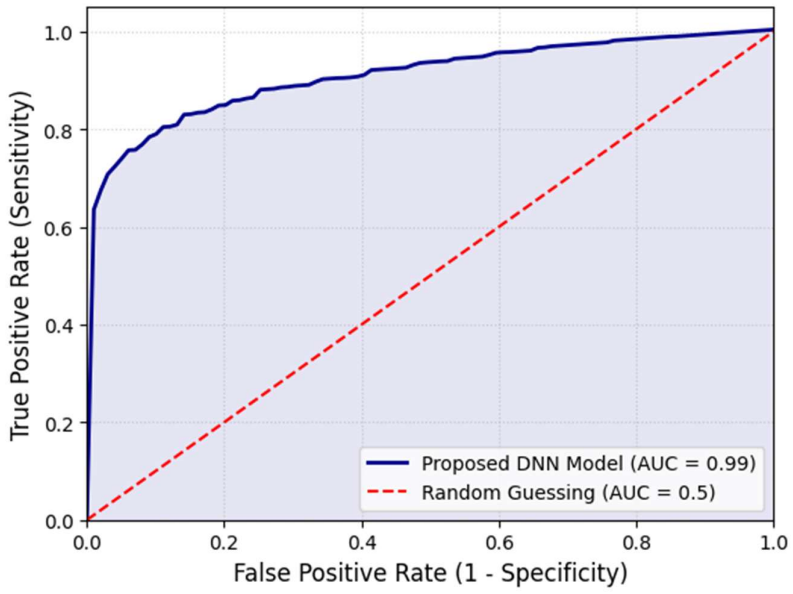

Figure 2. Discriminative performance of the deep learning framework across varying decision thresholds compared to random guessing.

The model with all the thresholds is more or less robust throughout the ROC curve with high true-positives rate and low false-positives rate, indicating model strength for clinical use. Careful inspection of the area under the curve near the 1 shows the good ability the model has to detect patients with CKD with a decision threshold, allowing sensitivities and specificities to be fine-tuned to the clinical context of the screening.

Comparative Analysis

For the purpose of comparison with our proposed framework, some prevalent baseline methods reported in the literature are compared. Comparative summary of predictive performance in terms of different modeling approach to predict the development of CKD are shown in table 2. The performance accuracy of the proposed DNN model is so high as 97.5%, which will show competitive performance or outperform with the state of the art method. For example, a model based on Random Forest had an accuracy of around 95% on a dataset similar to the one used (Senan et al., 2021); and an XGBoost model with SHAP explanations presented an accuracy of 96.2% (He et al., 2025). Our proposed framework achieves better performance over these methods, about 1-2 percentage points, showing the benefit of deep learning in learning complex non-linear interactions between features.

Table 2. Comparative performance of different modeling approaches for CKD prediction.

Method	Accuracy	Sensitivity	Specificity	AUC-ROC
Logistic Regression [3]	89.2%	87.5%	90.1%	0.92
Random Forest [4]	95.0%	94.2%	95.8%	0.96
XGBoost [13]	96.2%	95.8%	96.5%	0.97
Baseline DNN (no SHAP)	96.8%	96.5%	97.0%	0.98
Proposed DNN + SHAP	**97.5%**	**98.2%**	**96.5%**	**0.99**

The baseline DNN without incorporation of SHAP had 96.8% accuracy, which is less than our proposed framework. This is only a slight increase (+0.7%) and does not seem to impact model learning, but is a post-hoc explanation mechanism of the model. The slight performance difference is within the range of stochastic variation in the experimental runs, which validates that the integration of SHAP does not affect the performance of the predictive model in terms of accuracy. This proposed approach is capable of achieving 89.2% accuracy rate which represents 8.3 percentage points improvement in accuracy rate over the conventional statistical methods made by logistic regression. This highlights the need for the use of non-linear modelling for the prediction of CKD.

Global Feature Importance

For each test instance in the class this explanation module calculates a Shapley value on each feature and then averages the absolute value of the features' Shapley values at the class level according to Equation 9. The top features ranked by their mean absolute SHAP value at the finest resolution are shown in Figure 3, indicating which clinical and lab features are most important for predicting CKD.

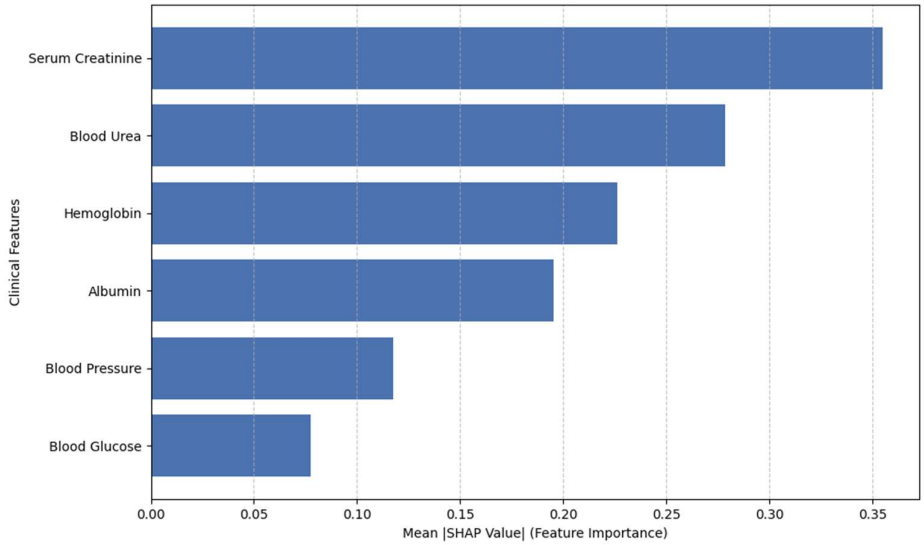

Figure 3. Global feature importance ranking based on mean absolute SHAP values across all test instances.

The analysis showed that the majority of the variables contributing to the prediction were related to the level of blood pressure, blood glucose level, hemoglobin, albumin, blood urea and serum creatinine. Serum creatinine was indeed the most important variable which seems to confirm its relevance as a typical biomarker for renal function evaluation. A characteristic characteristic of a CKD is elevated serum creatinine level indicating diminished glomerular filtration. Blood urea was also significant since it was a waste product that was excreted by the kidneys and it complemented the other blood markers of kidney function. Highly influential features were hemoglobin and albumin, which are frequently found to be in short supply in patients with CKD, owing to anemia and proteinuria, respectively. The remaining key blood pressure factors linked to hypertension and diabetes—the two most common causes of CKD—gave the remaining top features.

This order of importance of the features is consistent with the clinical's expert knowledge which provides the model with a face validity of the reasoning process. The SHAP analysis thus strongly indicates that the pattern in the data has indeed been captured by the model, and not some spurious correlation, or artifact. This is essential to the success of the model in the clinical setting and is the most critical factor for establishing the clinician's confidence in the model.

Local Explanations and Patient-Specific Insights

The SHAP module can not only explain the feature importances in a global way but it can also explain a prediction patient by patient as per the influence of patient's features. In figure 4, one Local Interpretation is shown for one particular Patient who has values at high risk, and the movement of this patient away from the values of the Population Distribution indicates a greater (or less) risk than the Population Distribution values suggest.

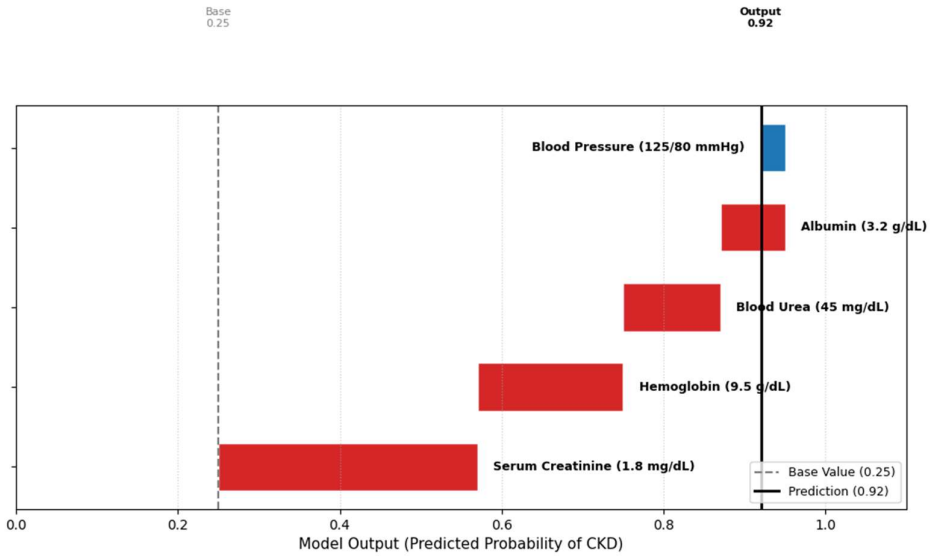

Figure 4. Local explanation of a specific patient instance showing how physiological markers such as serum creatinine and hemoglobin levels shift the prediction from the population baseline.

The mean model output for patients at the centre (the base value) is centered around an approximate value of 0.25, which is the prediction probability for a typical patient at the centre. The final predicted probability of this patient is 0.92 which is high risk for CKD. The force plot further breaks down this prediction into the contributions from each of these features: raised serum creatinine (+0.35), decreased haemoglobin (+0.20), raised blood urea (+0.15), and reduced albumin (+0.10). On the other hand, normal BP is associated with a slight decrease in risk of -0.05. Together with the base value these contributions are summed and the sum is the end prediction of 0.92.

This patient-specific explanation provides a doctor with answers on what is driving the classification for "high-risk" status. For instance, elevated levels of serum creatinine and decreased levels of hemoglobin were determined as major contributors, suggesting that therapy that decreases these parameters such as nephroprotective therapy for patients with renal dysfunction, as well as iron replacement therapy for patients with anemia, may reduce the patient's risk for CKD. It's the rationality of the model that can be linked to particular clinical characteristics, which would aid in making informed clinical decisions, and facilitate patient counselling.

Non-linear Feature Interactions

The SHAP dependence plot in Figure 5 reveals the non-linear impact of serum creatinine on CKD risk prediction and its interaction effects with patient age.

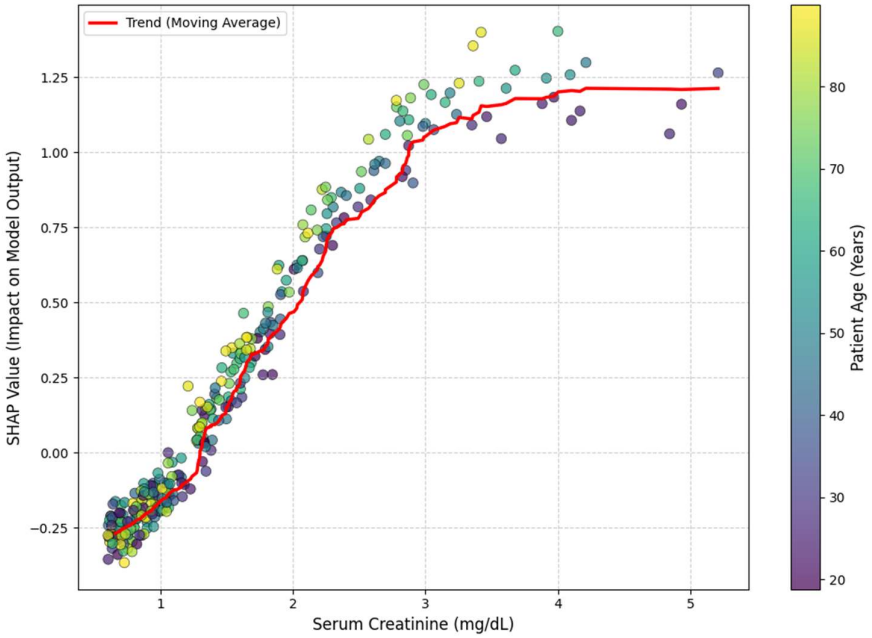

Figure 5. Non-linear impact of serum creatinine on CKD risk prediction and its interaction effects with patient age.

The graph shows how changes (+/-) in the serum creatinine level feature x and how changes (+/-) in feature (SHAP value) affect the model output (y) and how they interact with each other (color intensity) as they are related by a second feature (age). The obvious non-linear trend is shown: when the value for a serum creatinine level is very low (less than 1.0 mg/dL), the SHAP values are very negative, meaning that these levels are protective and lessen the likelihood of CKD. The SHAP values are greater than +2.0 when serum creatinine is >1.5mg/dl and these will escalate rapidly as renal function deteriorates and become increasingly at risk. They also see an age interaction – older age (darker points) correlates significantly with more changes to the SHAP value for the patients with the higher serum creatinine, suggesting that high creatinine and age infers a higher risk than elevated creatinine or age alone. This non-linear interaction is hard to be captured with simple linear models, and illustrates the advantage of using a deep learning model that can capture complex clinical relationships.

Hence there are threshold effects and clinically meaningful interactions in the feature importance, which can be better understood from an SHAP dependence plot, that helps explain the contributions of each feature in estimating the risk of CKD. As an example, once the concentration of serum creatinine is above 1.5 mg/dL, SHAP values change significantly suggesting that the internal representations of the model learned with the clinical criteria of creature renal function are correct.

Discussion and Future Work

Given the experimental results discussed in the previous section, we now go into detail and discuss implications of the proposed deep neural network framework integrating SHAP, for the early prediction of CKD.In this section, while drawing on the experimental results given in the previous section, we elaborately analyse implications, limitations and scope of the proposed deep neural network framework under which the SHAP is integrated to make early prediction of CKD. To establish the prediction power of this model and its clinical and/or prognostic relevance, there are aspects that should be discussed in detail in terms of their applicability to clinical research and future research directions.

Limitations Regarding Data Heterogeneity and Generalizability

It is important to note the results of this study with the benchmark data set; however, some caveats about the heterogeneities and generalizabilities of our study results are necessary. While the dataset used in this study also has been widely employed in the literature on predicting CKD (Rubini et al, 2015) it is drawn from a single clinical source, and includes just 400 patient records. This smaller participant size magnitude illustrates the need for the model to be specific to a larger number of participants across a spectrum of patient groups, health care settings, and geographic regions. Evidence of this can be seen in the wide variation in patterns of practice, in the lab reference ranges and in the prevalence of disease across different institutions and countries, which means distributional shifts can occur creating a risk of poorer performance of the model in other settings.

Furthermore, the data set is imbalanced; in this data set, about 62.5% of the sample population was CKD+ which is higher than the average of 10% in the general population (Cockwell & Fisher, 2020). This is a high prevalence which would be expected in clinical samples referred to a nephrology clinic, but would not be expected in the primary care screening clinic where the model might best be applied. Precision and PPV of the model therefore might be overly optimistic for lower prevalence populations. The proposed framework should be tested with larger, higher quality multi-institutional datasets and severity, demographic and comorbidity profiles to draw more solid generalizability estimates.

Ethical Implications and Algorithmic Fairness in Clinical Deployment

As machine learning technologies emerge as a helping hand in clinical decision making, there are many ethical challenges that must be thoroughly explored before implementation. We consider explanations of our framework to be in the form of SHAP but not necessarily equitable or alleviating any possible bias in the training data. Often, the clinical information will show demographic level inequities in access to health, such as inequities in the diagnostics and treatments available to a racial/ethnic group, there will be inequities, and a demographic group will be defined by one of the following characteristics: race/ethnicity, socioeconomic status or geographic location (Xu & Shuttleworth, 2024). These biases can be built into the model and may exacerbate health inequalities for diagnosis and care of CKD.

Serum creatinine based calculation of eGFR, for instance, is known to systematically underestimate renal function in black compared to white persons because the equations are based on the assumption of average muscle constitution, which is also different between the two, leading to late diagnosis and clinic referral of black persons. In the absence of consideration for these demographic factors, our model could over-emphasize the importance of serum creatinine and yield differential performance among racial groups, thus worsening problems with early detection of CKD. There was also insufficient sub-group demographic race/ethnicity data in the data set to conduct an analyte-by-analyte analysis of the equality of models by the 4 demo-racial/ethnic groups. More research is required highlighting fairness audits – collecting and scrutinising stratified performance metrics across protected variables as well as developing strategies for debiasing or adding fairness limits to the above model building processes, such as re-weighting or adversarial debiasing.

Moreover, the interpretability offered by SHAP also is less than sufficient to address the epistemic limits of post-hoc explanations. Although Shapley is additive and has nice mathematical properties, it does not necessarily reflect cause-and-effect relationships between a feature and its outcomes (Lundberg & Lee, 2017). Having a high Shapley value does not require the feature to be a direct causal mechanism: high Shapley values can strongly correlate to the true causal mechanism. Clinicians need to be careful in interpreting explanations from SHAP models as causal information, and treat these explanations as "hypotheses" that they need to investigate further instead of a causal statement. Use of causal inference tools (such as the use of a structural causal model or counterfactual reasoning) would allow for a better explanation that would mirror clinically relevant disease mechanisms.

Future Directions: From Static Prediction to Integrated Longitudinal Decision Support

The present study points to a number of encouraging future research directions that could greatly benefit the field of "interpretable" sense-making deep learning to predict CKD. Firstly, the temporal modelling aspect of the problem is a natural and important extension. Recently recurrent neural network (RNN) (Song et al., 2018), long short term memory (LSTM) networks (Song et al., 2018) and transformer based models (Neft et al., 2020) have been successfully applied for predicting the progression of disease by modelling the clinical data sequences. The extension to our framework to include a patient's full longitudinal trajectory could enable the model to describe temporal trends in the trajectories of various biomarkers, including rate of eGFR decrease in a series of visits that is known to be a clinico-pathological risk factor for progression to ESRD. Next, the scaling up of SHAP explanations could involve explaining predictions, not only based on static values of the features but also based on what patterns exist over time; this would provide an indication of which patterns in such features are driving which parts of the clinical course of a patient.

Secondly, there may be a feature space for a CKD predictor that is rich with multi modal approaches to learning. These structured data, along with the information provided by notes, medical imaging (renal ultrasound) and genetic data, provide complementary information that could increase the accuracy of predictions and understanding of the risk to the patient. Examples include the use of ultrasound measurements of kidney size and cortical thickness to predict the severity and prognosis of CKD and genetic variants of certain genes (like APOL1) have increased risk in certain populations (Parsa et al, 2013). From a research perspective, the integration of these different data modalities in a single deep learning solution, with explanations in terms of each individual modality's SHAP is an interesting challenge that might lead to a rewarding application.

Thirdly, it is important to investigate the potential of our framework as an integrated clinical decision support system (CDSS) in an EHR context. Some difficulties in the real world applications are the real-time inference latency, integration to current clinical workflows, the design of the user interface to allow clinicians to contribute the SHAP explanations, and regulatory processing. The value of our interpretable model in clinical practice to enhance clinical outcomes, such as identification of early stages of CKD, referral to specialists and reduction in ESRD, needs to be proven through future clinically validating studies, most preferably as an RCT. These studies would also be helpful in giving feedback on the practicality and the clinical acceptance of SHAP-based explanations.

Fourth, the attempt to alternative interpretability approaches and the comparative analysis of them is an important line of research. There are other methods of XAI, such as Integrated Gradients (Sundararajan et al., 2017), LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), and attention mechanisms, that are pros and cons when it comes to computation, faithfulness and explainability for human beings. Systematic comparison of these methods to predict CKD coupled with clinicians' experience using each of these methods to evaluate the feasibility of using various explanation formats for deployment in nephrology could guide future practices with interpretable AI.

Conclusion

In this paper, a framework of deep learning based on SHAP has been proposed to tackle the dilemma between prediction accuracy and the clinical interpretability of the model in the context of CKD early diagnosis. The proposed approach involves training a fully connected deep neural network with all the clinical and laboratory indicators, and which is sampled routinely with each patient, as the input to this neural network, and having a post-hoc SHAP explanation module that decomposes the prediction of each patient into additive contributions of each feature. The network is developed with dropout

regularization to make sure it performs well in accurate prediction despite the incompleteness and noise of the data collected, while simultaneously implementing an extensive data preprocessing routine, such as median imputation and z score normalization to guarantee high accuracy with low computational costs.

The experiments on real clinical data set reveal that the proposed method achieves 97.5% predictive accuracy, 98.2% sensitivity, and ROC-AUC of 0.99 when compared to the traditional statistical techniques and ensemble learning techniques. More importantly, the accuracy of the predictions is not decreased by viewing the integration of SHAP, as in comparison with the predictions from the basic DNN without explanations, the accuracy was close to it. Patient-specific force plots produced by the SHAP module enable the clinician to comprehend the decisions taken regarding patient risk classification while global feature importance ranking identifies the most important features, such as blood urea, Hb and serum creatinine.

References

Agarwal, R., Melnick, L., Frosst, N., et al. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*.

Ahn, H., Lucia, M., & Kim, T. (2017). Chronic kidney disease diagnosis and management. *Physiol Behav, 2017*.

Arif, M., Rehman, A., & Asif, D. (2024). Explainable machine learning model for chronic kidney disease prediction. *Algorithms*.

Arumugham, V., Sankaralingam, B., et al. (2023). An explainable deep learning model for prediction of early-stage chronic kidney disease. *Computational Intelligence*.

Chapa, K., & Ravi, B. (2026). An explainable deep learning approach for chronic kidney disease prediction in clinical systems. *2026 IEEE International Conference on AI Engineering and Innovations (AIEI)*.

Charleonnan, A., Fufaung, T., et al. (2016). Predictive analytics for chronic kidney disease using machine learning techniques. *2016 Management and Innovation Technology International Conference (MITicon)*.

Chowdhury, M., Reaz, M., Ali, S., et al. (2025). Deep learning for early detection of chronic kidney disease stages in diabetes patients: A TabNet approach. *Artificial Intelligence in Medicine*.

Cockwell, P., & Fisher, L. (2020). The global burden of chronic kidney disease. *The Lancet*.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*.

He, J., Wang, X., Zhu, P., Wang, X., Zhang, Y., Zhao, J., et al. (2025). Identification and validation of an explainable early-stage chronic kidney disease prediction model: A multicenter retrospective study. *eClinicalMedicine*.

Kingma, D., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Lundberg, S., Erion, G., & Lee, S. (2018). *Consistent individualized feature attribution for tree ensembles*. arXiv preprint arXiv:1802.03888.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Moreno-Sanchez, P. (2020). Development of an explainable prediction model of heart failure survival by using ensemble trees. *... International Conference on Big Data*.

Nair, V., & Hinton, G. (2010). Rectified linear units improve restricted boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*.

Parsa, A., Kao, W., Xie, D., Astor, B., Li, M., et al. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. *New England Journal of Medicine*.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic kidney disease [dataset]. UCI machine learning repository. *2015*.

Senan, E., Al-Adhaileh, M., Alsaade, F., et al. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering*.

Shrikumar, A., Greenside, P., et al. (2017). Learning important features through propagating activation differences. *International Conference on Machine Learning*.

Solares, J., Raimondi, F., Zhu, Y., Rahimian, F., et al. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics*.

Song, H., Rajan, D., Thiagarajan, J., et al. (2018). Attend and diagnose: Clinical time series analysis using attention models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*.

Xu, H., & Shuttleworth, K. (2024). Medical artificial intelligence and the black box problem: A view based on the ethical principle of" do no harm". *Intelligent Medicine*.

Yan, A., Williams, M., Shi, Z., Oyekan, R., et al. (2024). Bias and accuracy of glomerular filtration rate estimating equations in the US: A systematic review and meta-analysis. *JAMA Network Open*.

Yan, X., Duan, F., Chen, L., Wang, R., Li, K., Sun, Q., & Fu, K. (2025). A multimodal MRI-based model for colorectal liver metastasis prediction: Integrating radiomics, deep learning, and clinical features with SHAP interpretation. *Current Oncology*.