Deep Learning Identifies Intelligible Predictors of Poor Prognosis in Chronic Kidney Disease

Ping Liang, Jiannan Yang, Weilan Wang, Guanjie Yuan, Min Han, Qingpeng Zhang, *Senior Member, IEEE*, and Zhen Li

***Abstract*—Early diagnosis and prediction of chronic kidney disease (CKD) progress within a given duration are critical to ensure personalized treatment, which could improve patients' quality of life and prolong survival time. In this study, we explore the intelligibility of machine-learning and deep-learning models on end-stage renal disease (ESRD) prediction, based on readily-accessible clinical and laboratory features of patients suffering from CKD. Eight machine learning models were used to predict whether a patient suffering from CKD would progress to ESRD within three years based on demographics, clinical,and comorbidity information. LASSO, random forest, and XGBoost were used to identify the most significant markers. In addition, we introduced four advanced attribution methods to the deep learning model to enhance model intelligibility. The deep learning model achieved an AUC-ROC of 0.8991, which was significantly higher than that of baseline models. The interpretation generated by deep learning with attribution methods, random forest, and XGBoost was consistent with clinical knowledge, whereas LASSO-based interpretation was inconsistent. Hematuria, proteinuria, potassium, urine albumin to creatinine ratio were positively associated with the progression of CKD, while eGFR and urine creatinine were negatively associated. In conclusion, deep learning with attribution algorithms could identify intelligible features of CKD progression. Our model identified a number of critical, but under-reported features, which may be novel markers for CKD progression. This study provides physicians with solid data-driven evidence for using machine learning for CKD clinical management and treatment.**

***Index Terms*—Interpretable deep learning model, machine learning, chronic kidney disease.**

Manuscript received 15 October 2022; revised 19 March 2023; accepted 3 April 2023. Date of publication 12 April 2023; date of current version 3 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 81771801, 82071889, 81970591, and 82270725, in part by the Innovation and Technology Fund of Innovation and Technology Commission of Hong Kong under Grant MHP/081/19, and in part by the National Key Research and Development Program of China, Ministry of Science and Technology of China under Grant 2019YFE0198600. *(Ping Liang and Jiannan Yang contributed equally to this work.) (Corresponding authors: Min Han; Qingpeng Zhang; Zhen Li.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (Application No.TJ-IRB20210517 and performed in line with the Declaration of Helsinki).

Ping Liang, Guanjie Yuan, and Zhen Li are with the Department of Radiology, Tongji Hospital, Huazhong University of Science and Technology, Wuhan 430030, China (e-mail: pinglianglp@163.com; ygjforever98@163.com; zhenli@hust.edu.com).

Qingpeng Zhang is with the Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, and the Musketeers Foundation Institute of Data Science, University of Hong Kong, Hong Kong SAR, China (e-mail: qpzhang@arizona.edu).

Jiannan Yang and Weilan Wang are with the School of Data Science, City University of Hong Kong, Hong Kong SAR, China (e-mail: jiannan.yang@my.cityu.edu.hk; weilwang@cityu.edu.hk).

Min Han is with the Department of Nephrology, Tongji Hospital, Huazhong University of Science and Technology, Wuhan 430030, China (e-mail: minhan@tjh.tjmu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2023.3266587, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2023.3266587

I. Introduction

APPROXIMATELY 10.8% (10.2–11.3) of the population in China are suffering from chronic kidney disease (CKD) [1]. With population aging and the rising prevalence of chronic diseases such as diabetes, hypertension, and obesity, the number of people suffering from CKD is anticipated to increase in the next few years [2]. CKD carries a high risk of complications (such as cardiovascular events), as well as death. In early-stage CKD, non-drug therapies (such as diet and lifestyle adjustments) and specific drugs (such as angiotensin-converting enzyme inhibitors or Angiotensin II receptor blockers) are commonly introduced to preserve kidney function [3]. However, due to the robust compensatory ability of the kidney, most people have no apparent symptoms in early stage of the disease [4]. Once CKD progresses into end-stage renal disease (ESRD), sufferers develop typical renal insufficiency symptoms. By this stage, the available treatments are largely limited, including hemodialysis, peritoneal dialysis, or kidney transplantation [5].

Early diagnosis and prediction of CKD progress within a given duration are critical to ensure personalized treatment, which could improve patients' quality of life and prolong survival time. However, due to the heterogeneity of CKD sufferers and the impact of confounding factors, it is difficult to predict when CKD will progress into renal failure [6]. Inaccurate prediction of CKD progression may lead to delays in treatment for people who are likely to progress to renal failure, and unnecessary treatment for people whose condition may not progress.

Percutaneous kidney biopsy is helpful to determine the pathological types of CKD, guide treatment, and identify the degree of fibrosis, which is the gold standard in defining prognosis [7]. However, percutaneous kidney biopsy is an invasive procedure that may induce bleeding, infection, or other damage. Non-invasive biomarkers such as estimated glomerular filtration rate (eGFR) are used to detect the progress of CKD and provide an

individualized prognosis which in turn can provide clinicians with roadmaps of early intervention [8], [9].

Logistic regression and Cox proportional hazards regression models are the most commonly used clinical methods to predict CKD progression using non-invasive biomarkers [10], [11]. Tangri et al. used Cox proportional hazards regression models to establish the Kidney Failure Risk Equation based on age, sex, eGFR, and urine albumin/creatinine ratio (UACR) to predict CKD progression [12]. However, these studies were based on linear assumptions, and these models performed relatively poorly in the validation cohort. Furthermore, these studies mainly focused on people with advanced CKD, while ignoring the much larger group of people with early-stage CKD. Thus, establishing methods that provide more accurate predictions for people with earlier-stage disease is critical for personalized treatment.

In the past few decades, machine learning and deep learning technologies have been widely used in many fields, such as translation and face recognition [13]. Some progress has been made in medical research [14], [15], especially for the progression prediction of CKD. For example, deep learning models can be used to identify CKD and type 2 diabetes from fundus images combined with clinical data [16]. They can also be applied to predict the risk of diseases whilst still in early stage [17]. Specifically, Thomas et al. [18] developed a random forest model for the progression of CKD and achieved an AUC-ROC of 0.88 (95% CI 0.87-0.89). Francesco et al. [19] developed an artificial neural network to predict ESRD in patients with IgAN (Immunoglobulin A Nephropathy) and achieved an AUC-ROC of 0.82. However, although deep learning models could significantly improve prediction performance, they are nearly all black-box models, that humans cannot understand how the input features are being organized by the models to make predictions.

In this paper, we aimed to apply a deep neural network (DNN) and compare it with classic machine learning models to predict CKD progression for people at different stages of the disease, based on demographic variables, laboratory and blood biochemical indicators, and comorbidity information. In addition, we introduced advanced attribution algorithms to enhance the intelligibility of DNN, and compared their outputs with those from other intelligible machine learning models. Our models and intelligibility analysis may assist clinicians to formulate more appropriate management and treatment plans to delay the progression of CKD and reduce patient burden.

II. Materials and Methods

A. Study Population and Data Processing

This research was approved by the Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (No.TJ-IRB20210517) in accordance with the Declaration of Helsinki, and the informed consent was waived. We retrospectively analyzed data of 2382 people diagnosed with CKD from January 2009 to December 2020. The database includes basic demographic and clinical characteristics such as liver and kidney function, blood routine test results, and comorbidity information. We excluded patients based on the following criteria: (1) those missing greater than 30% values; (2) those people younger than 18 years old; (3) people with only one admission record or whose observation period was less than six months, or lost of follow-up; (4) people with acute renal insufficiency or congenital kidney diseases. Finally, 1765 people were included in this study. The flowchart of the study cohort is presented in Fig. 1.

The factors (input features) affecting the progression of CKD disease can be roughly divided into three categories: (1) basic demographics such as sex, age; (2) systemic comorbidity such as hypertension, diabetes, urolithiasis, hyperlipidemia; and (3) basic laboratory biochemical tests. For each subject, comorbidity information was collected from diagnosis records between the first diagnosis date and the date 30 days after the first kidney-related diagnosis record. For each laboratory biochemical test, we kept the earliest record of each test. The missing values of one subject's biochemical tests were replaced by the mean value calculated from the subjects that belong to same CKD stages. The basic statics of the features considered in this study are shown in Table I, data are presented as mean $\pm$ standard deviation (SD), or n (percentage).

B. ESRD Definitions

In this study, the end of CKD progression within three years (the positive label, denoted as 1) was end-stage renal disease (ESRD). We considered the possibility of false negative results due to both too short and too long follow-up times. Too short follow-up times might miss patients who will eventually progress to ESRD (only 20% subjects have progressed to ESRD within 1 a in our dataset), while long follow-up times might include mostly patients who have already progressed to ESRD (89% of subjects progressed to ESRD within 5 years, and more than 77% of those subjects had already progressed to ESRD 2 years prior). ESRD was defined as the initiation of renal dialysis treatment (including peritoneal dialysis and hemodialysis) or kidney transplantation, or eGFR was reduced by 50% over the observation period of an individual from the first time it was recorded. The eGFR values were calculated by using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation based on creatinine clearance [20].

C. CKD Stage Definitions

We divided subjects into four groups according to their first eGFR values. People with eGFR between 15 to 60 ml/min/1.73 m^2 may develop ESRD at an estimated rate of 1.5% per year [21]. Thus we classified CKD stage 1 (eGFR$\geq$90 ml/min/1.73 m^2) and stage 2 (eGFR, 60-89 ml/min/1.73 m^2) into one group in our study, and compared them with stage 3 (eGFR, 30–59 ml/min/1.73 m^2), stage 4 (eGFR, 15–29 ml/min/1.73 m^2), and stage 5 (eGFR$<$15 ml/min/1.73 m^2).

D. Deep Learning Model With Intelligible Mechanisms

Deep neural networks (DNN) [22] are artificial neural networks with multiple layers between the input features and output predictions. Each linear layer is connected by non-linear activation functions to learn non-linear relationships between the input features. In this study, we utilized a four-layer neural

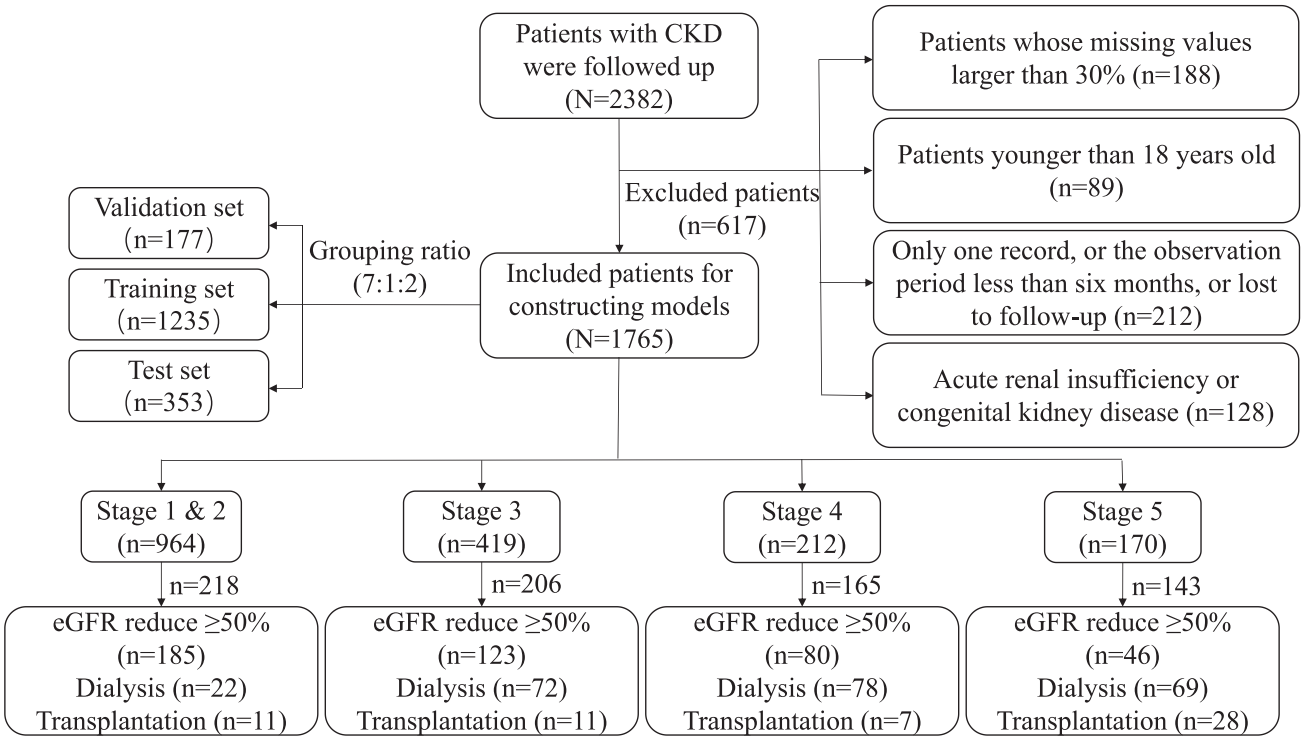

Fig. 1. The flowchart of the study cohort.

TABLE I
STATICS OF DEMOGRAPHICS CHARACTERISTICS, COMORBID CONDITIONS, AND LABORATORY TEST DATA OF THE COHORT

Characteristics	Total	Stage 1 and 2	Stage 3	Stage 4	Stage 5	P value
No. of participants	1765	964	419	212	170	< 0.0001
Demographics						
Age (year)	44.89±15.46	40.64±14.56	50.12±14.99	51.41±15.18	47.92±14.36	<0.0001
Sex Male (%)	0.59	0.56	0.65	0.64	0.60	0.008
End-stage ratio (%)	0.41	0.23	0.49	0.78	0.84	<0.0001
Follow time (days)	1463.83±769.88	1654.94±733.42	1422.51±736.45	1017.95±659.10	1038.07±764.50	<0.0001
Comorbid conditions						
Diabetes	606 (34%)	254 (26%)	170 (41%)	102 (48%)	80 (47%)	<0.0001
Hypertension	637 (36%)	199 (21%)	216 (52%)	131 (62%)	91 (54%)	<0.0001
Laboratory data						
eGFR ($mL/min/1.73m^2$)	66.30±37.37	94.87±22.75	46.14±8.57	22.38±4.15	8.72±3.63	<0.0001
Serum creatinine (mg/dL)	165.57±190.98	79.05±48.78	138.84±37.38	255.53±70.36	609.85±322.33	<0.0001
BUN (mg/dL)	8.68±5.95	5.62±2.40	8.74±3.51	13.61±4.71	19.74±8.42	<0.0001
Uric-acid (mg/dL)	387.20±112.70	351.89±95.14	423.09±104.73	445.21±110.05	426.65±149.64	<0.0001
WBC count (mm^3)	8.21±9.88	8.28±7.69	8.81±15.96	7.76±5.74	6.90±2.58	0.17
RBC count (mm^3)	37.47±291.81	46.20±258.64	42.84±449.91	9.52±30.92	9.65±49.15	0.21
Hemoglobin (g/dL)	125.23±26.56	135.62±22.64	121.54±23.08	109.82±21.30	94.71±25.85	< 0.0001
Albumin (g/dL)	33.12±9.84	32.36±10.45	34.01±8.80	33.13±8.61	35.18±9.70	0.008
AST (U/L)	23.85±35.50	24.57±22.39	22.26±12.96	27.54±86.34	19.02±21.97	0.08
ALT (U/L)	23.90±43.84	27.38±52.82	19.95±16.19	21.97±48.60	16.29±19.33	0.001
Sodium (mg/dL)	138.72±11.85	139.11±9.41	138.91±12.14	139.53±3.03	135.05±23.78	0.001
Potassium (mg/dL)	4.28±0.66	4.11±0.46	4.27±0.58	4.73±0.70	4.72±1.12	<0.0001
Calcium (mg/dL)	1.75±1.09	1.76±1.10	1.82±1.05	1.78±1.04	1.54±1.12	0.04
Phosphorus (mg/dL)	1.09±0.42	1.02±0.36	1.06±0.33	1.24±0.41	1.36±0.72	<0.0001
Chloride (mg/dL)	102.98±9.36	102.88±7.47	103.38±9.47	105.01±4.19	100.03±18.32	<0.0001
Cholesterol (mg/dL)	5.88±2.84	6.42±3.03	5.57±2.54	5.32±2.50	4.25±1.73	<0.0001
Triglyceride (mg/dL)	2.37±1.99	2.42±2.03	2.46±2.08	2.46±1.87	1.71±1.47	0.0003
Glucose (mg/dL)	27.29±10.68	28.24±8.08	27.39±11.06	26.30±13.73	22.85±15.81	<0.0001
Urine albumin-to-creatinine ratio (mg/g) Median (IQR)						
0-299	123.90 (72.35-205.18)	126.30 (72.20-206.00)	101.60 (65.75-180.03)	154.00 (114.85-199.90)	132.60 (98.03-174.75)	0.88
≥300	2002.80 (900.50-4000.20)	1819.90 (842.15-3951.30)	2081.60 (865.95-4163.60)	2285.10 (1352.88-3751.48)	2012.35 (926.98-3398.97)	0.70
Outcome						
Dialysis	257 (15%)	24 (2%)	79 (19%)	83 (39%)	71 (42%)	<0.0001
Kidney Failure	239 (15%)	24 (2%)	75 (18%)	80 (38%)	60 (47%)	<0.0001
Transplantation	57 (3%)	11 (1%)	11 (3%)	7 (3%)	28 (16%)	<0.0001

network with BatchNorm [23] and Dropout [24] modules for better performance. Each layer is as follows:

$$O_l = ReLU(\text{Dropout}(\text{BatchNorm}(\text{Linear}(I_l)))), \quad (1)$$

where I_l, O_l are the input and output of layer l, respectively, and $ReLU$ denotes the Rectified Linear Unit activation function. The last layer is a sigmoid layer for a binary prediction as shown in (2).

$$O_{last} = \sigma(Linear(I_{last})) \quad (2)$$

where I_{last} and O_{last} are the input and output of the last layer, respectively, and σ denotes the sigmoid activation function.

Since deep learning models are mostly black box models which lack interpretability, we introduce several attribution algorithms [25] to enhance the intelligibility of DNN. These algorithms computed the gradient of the model's prediction concerning each feature to show how the output value changes, given small changes due to perturbations of input features. Here we applied four attribution algorithms, including Integrated Gradients [25], DeepLIFT [26], Gradient SHAP [27], and Feature Ablation [28]. All four algorithms generated a score for a specific feature based on the trained model, denoting the contribution of this feature to the positive label. Specifically, high positive scores can be interpreted as positively related to ESRD. Instead, high negative scores would be interpreted as a higher value of this feature, a lower possibility of progressing to ESRD.

a) Integrated Gradients: Integrated Gradients is an axiomatic attribution proposed by Mukund et al. [25] It represents the path integral of the gradients along the path from the baseline x' to the input x. In this study, the baseline is the average or mode value for each feature of the subjects with label 0. And the integral is approximated by Riemann Sum or Gauss Legendre quadrature rule as follows

$$IntegratedGrads_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (3)$$

where x denotes the input and x' is the baseline; i is the ith feature; F is the neural network; α is the scaling coefficient.

b) DeepLIFT: DeepLIFT [26] is similar to Integrated Gradients which is a back-propagation-based approach that attributes changes to inputs based on the differences between the inputs and baselines. DeepLIFT uses multipliers m to "blame" specific neurons for the difference in output as follows

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x} \quad (4)$$

where Δx is the difference between the input neuron x and baseline, and Δt is the difference between target neuron and reference. C is then the contribution of Δx to Δt.

c) Gradient SHAP: Gradient SHAP [27] is SHAP-based method which adds Gaussian noise to the inputs multiple times and then selects a random point along the path between baseline and input. The final SHAP values are defined as follows:

$$\text{SHAP} = E[gradients \times (x - x')] \quad (5)$$

where the $gradient$ is defined as the gradient of outputs with respect to the selected random points.

d) Feature Ablation: Feature Ablation [28] is a perturbation-based approach. It replaces each input with a given baseline, and computes the difference in output.

E. Experimental Settings

Suppose the features of an individual i are denoted as x_i, with label y_i, where $y_i = 1$ represents that this person progressed to ESRD within three years, or vice versa. All our models aimed to learn a function $\hat{y}_i = f(x_i|y_i)$, where $\hat{y}_i$ denotes the probability of whether one person would progress to ESRD within three years. We formulated the following loss function for our DNN model:

$$\mathbb{L} = \sum_{i \in N} \mathcal{L}(y_i, \hat{y}_i) + \frac{\lambda_1}{2}||\Theta||_1 + \frac{\lambda_2}{2}||\Theta||_2^2 \quad (6)$$

where $\mathcal{L}(y_i, \hat{y}_i) = -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$ is the binary-cross-entropy loss, $\hat{y}_i$ is the predicted label, Θ denotes the model parameters to be learned. The second and third terms are the l_1 and l_2 regularizes to prevent over-fitting, where λ_1 and λ_2 are the hype-parameters for l_1 and l_2 regularizations, respectively.

F. Baselines

We applied seven machine learning models, which were classified into four categories.

a) Linear model: We selected Logistic Regression (LR) [29], Ridge Regression Classification (RRC) [30], and Least Absolute Shrinkage and Selection Operator (LASSO) [31]. These are all linear models but with different regularizations on the parameters. Specifically, LR applies a logistic function to model binary dependent variables without any regularizations, which is widely used in medical research [32], [33]. LASSO introduces an l_1 regularization to perform both variable selection and avoid over-fitting problem [31]. In contrast, RRC utilizes an l_2 regularization which extends the robustness of the model but lacks variable selection ability [30].

b) Support Vector Machine (SVM) model: SVM [34] aims to learn a non-linear relationship in the kernel space to make a classification. In detail, the SVM model depends on a kernel trick to implicitly map the features into high-dimensional feature spaces. In this study, we introduce two kernel tricks, a Gaussian kernel (SVM-RBF) and a linear kernel (SVM-Linear).

c) Decision Tree model: Here we introduce two widely-used decision tree models, Random Forest (RF) [35] and XGBoost model [36]. XGBoost is a scalable end-to-end tree boosting system that is fast and accurate and used in many medical tasks.

G. Tuning of Parameters

We utilized grid search to find the best setting for each model, which is conducted by optimizing the area under the receiver operating characteristic curve (AUC-ROC) metric on the validation set. The ratio of training, validation, and test set

TABLE II
PERFORMANCE METRICS OF ALL THE MODELS

Model	Accuracy	Precision	Recall	AUC-ROC	PR-AUC	F1 score
DNN	**0.8440 (0.0083)**	0.7732 (0.0134)	**0.7575 (0.0268)**	**0.8991 (0.0100)**	**0.8281 (0.0183)**	**0.7584 (0.0169)**
LR	0.8141 (0.0158)	0.7400 (0.0307)	0.6521 (0.0392)	0.8561 (0.0211)	0.7500 (0.0282)	0.6923 (0.0255)
LASSO	0.8206 (0.0128)	0.7664 (0.0354)	0.6372 (0.0438)	0.8636 (0.0163)	0.7624 (0.0248)	0.6945 (0.0273)
RRC	0.8258 (0.0137)	0.7850 (0.0193)	0.6309 (0.0377)	0.8667 (0.0183)	0.7647 (0.0242)	0.6989 (0.0253)
SVM-RBF	0.8160 (0.0129)	0.7563 (0.0303)	0.6346 (0.0475)	0.8734 (0.0150)	0.7801 (0.0211)	0.6884 (0.0232)
SVM-Linear	0.8273 (0.0120)	0.7862 (0.0280)	0.6362 (0.0327)	0.8627 (0.0182)	0.7641 (0.0247)	0.7026 (0.0219)
RF	0.8361 (0.0160)	**0.8031 (0.0336)**	0.6525 (0.0438)	0.8874 (0.0107)	0.8056 (0.0184)	0.7186 (0.0260)
XGBoost	0.8331 (0.0147)	0.7738 (0.0374)	0.6816 (0.0422)	0.8883 (0.0128)	0.8042 (0.0289)	0.7235 (0.0272)

The average performance over 10 trainings was reported. The values in the brackets denote standard deviations. The bold values indicate the best performance.

is 7:1:2. Specifically, for our deep learning model, the number of neurons of each layer is 97-194-97-1, where 97 is consistent with the number of input features. Each layer is connected by a $ReLU$ activation function with a dropout rate equal to 0.3. For the two regularization terms, λ_1 and λ_2 are set to 0.005 and 0.001, respectively. The learning rate is set to 0.001, and all trainable parameters are optimized by the Adam algorithm with a batch size of 128, and the number of epochs is set to 50 for training. For the linear models, the l_1 regularization parameter is set to 0.03 for LASSO and the l_2 regularization parameter is set to 0.65 for the RRC model. For two SVM models, the c values are set to 1.0 for both SVM-RBF and SVM-Linear, and $\gamma = 0.4$ specifically for the SVM-Linear model. For the decision tree models, the number of estimators is 180 for RF and 170 for XGBoost, respectively, and the maximum depth of XGBoost is set to 4. All models are implemented with Python 3.7, PyTorch 1.6.0, NumPy 1.19.1, and scikit-learn 0.23.2.

III. RESULTS

A. Study Cohort

In total, 1765 people suffering from CKD were recruited into this study cohort. Detailed demographic characteristics, comorbid conditions, and laboratory data are shown in Table I. There were significant differences within the four groups in age, follow-up duration, and 17 biochemical test values such as serum creatinine, hemoglobin.

B. Performance of Deep Learning

Performance was evaluated by five metrics: accuracy, precision, recall, AUC-ROC, the area under the precision-recall curve (AUC-PR), and F1 score. Each model was trained ten times, and the average performance and standard deviation were reported. In general, as shown in Table II, the DNN model outperformed all baselines given all the metrics except for Precision, which reached a mean AUC-ROC value of 0.8991 (1.2% higher than the second-best finding). Furthermore, the DNN model also achieved much higher recall and PR-AUC metrics compared with other models, indicating that the DNN model can identify the patients who will progress to ESRD within three years more precisely and sensitively. Combining these findings, the DNN model shows a better performance in capturing the non-linear relationships within the input features, and generates a better prediction.

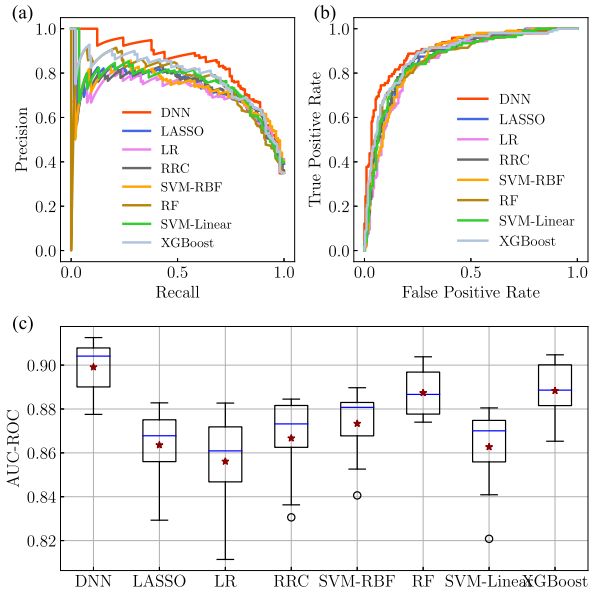

Fig. 2. (a) The Precision-Recall curves of all the models. (b) The ROC curves of all the models. (c) The box plots of the AUC-ROC metric of all the models. The blue lines and red stars denote the median and mean values, respectively.

For the other models, two decision tree models – XGBoost and RF performed second-only after the DNN model. All the linear models (LASSO, LR, RRC, and SVM-Linear) performed more poorly than the non-linear models, which indicates that the non-linear functions better describe the relationships between the predicted features and the outcome. Compared within three linear regression classification models, RRC performs best, LASSO followed, and LR last, in line with the number of regularization terms. The SVM-RBF, which utilizes a Gaussian kernel, generates a higher AUC-ROC value among two SVM-based models.

To test the robustness of each model, we trained all the models ten times with different random seeds, and the box plots of AUC-ROC values are shown in Fig. 2(c). The DNN model is the most robust (with the slightest standard deviation of 0.0100) of all the models. Except for the two decision tree models, other machine learning models have similarly larger standard deviation values.

C. The Features of the ESRD Driver

Of all the models used in this study, the DNN model with attribution algorithms, LASSO, Random Forest, and XGBoost

can generate a score of each feature which denotes its importance to the positive prediction (ESRD). The score is shown as the normalized contribution weights in our study. As shown in Fig. 3, DNN with attribution algorithms and LASSO can generate both positive and negative contribution weights, where the positive weights denote that the higher value of this feature, the higher risk to drive ESRD. In contrast, two decision tree models can only generate contribution weights without directions which indicate how much the feature contributes to the positive label regardless of directions.

In this study, we define "critical features" as the features with the top 20 highest contribution weights (absolute value) to ESRD. The critical features identified by the DNN model are consistent with other intelligible/explainable machine learning models. For example, 14 features such as hematuria and eGFR are captured by both DNN-DeepLIFT and machine learning models. In addition, we performed a Cox proportional hazards regression analysis on our dataset, which revealed that four crucial features: Potassium, Ucr (urine creatinine), Proteinuria, and eGFR (CKD-EPI), exhibit the same logical relationships as those identified by the DNN-DeepLIFT (as detailed in the Supplementary Information). Two decision tree models generate almost the same top 20 critical features (14 in common), where a difference exists in the order of these features. For example, eGFR is identified as second-important in RF, while it is the most important in XGBoost. Hematuria and Ucr (urine creatinine) are identified as the most significant positive and negative critical features by DNN-DeepLIFT, respectively, whilst LASSO, monocytes (%) and red cell distribution width (RDW) are identified as the most significant positive and negative critical features, respectively. Serum creatinine and eGFRare identified as the most critical features by RF and XGBoost, respectively. Comparison within the four attribution algorithms shown in Fig. 3(e), except for the GradientSHAP algorithm, the other three algorithms generate similar results. For example, all three algorithms identified Ucr as the most significant negative critical feature to ESRD. However, GradientSHAP assumes that Ucr is positive, which is not consistent with clinical knowledge [37].

Combining the critical features identified by all these models, we conclude the positive critical features to the progression of CKD are: hematuria, potassium, proteinuria, urine albumin to creatinine ratio (ACR), cystatin C. Negative critical features for the progression of CKD are eGFR and Ucr.

To further validate the interpretability of the DNN model for the individuals with different etiologies, we divided all the individuals into four categories: individuals with CKD caused by hypertension, diabetes, urolithiasis, or chronic glomerulonephritis separately. The identified critical features are shown in Fig. 4. We found that hematuria is the most important independent risk predictor for the progression of diabetic nephropathy (DN) and urolithiasis. Bicarbonate was the most important independent risk factor for predicting the deterioration of renal function in hypertensive patients with renal insufficiency. Furthermore, bicarbonate, hematuria and proteinuria are the most important independent risk factors for the progression of primary glomerulonephritis.

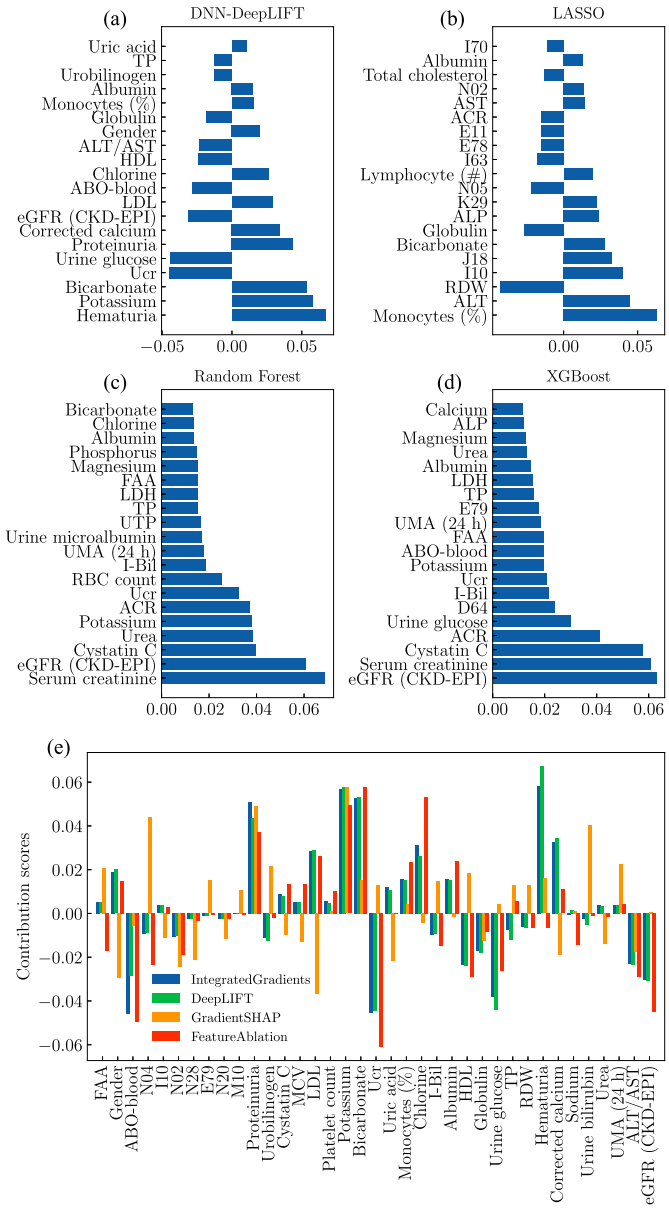

Fig. 3. Top 20 important features identified by DNN with Integrated Gradients (a), LASSO (b), Random Forest (c), and XGBoost (d). The contribution weights (absolute value is larger than 0.01) of all the features generated by five attribution algorithms for the DNN model (e). The full names of the abbreviation of these features are as follows: TP: total protein; ALT/AST: the ratio between the concentrations of the enzymes aspartate transaminase (AST) and alanine transaminase, aka alanine aminotransferase (ALT); HDL: high-density lipoprotein; LDL: low-density lipoprotein; Ucr: urine creatinine; ACR: urine albumin to creatinine ratio; ALP: alkaline phosphatase; RBC: red blood cell; RDW: red cell distribution width; ALT: alanine transaminase; FAA: first admission age; LDH: lactate dehydrogenase; UTP: urinary total protein; UMA (24 h): urine microalbumin in 24 hours; I-Bil: indirect bilirubin; MCV: mean corpuscular volume; RDW-CV: red cell distribution width cv. The references of ICD-10 codes are: I70: Atherosclerosis of aorta; E11: type 2 diabetes mellitus; E78: disorders of lipoprotein metabolism and other lipidaemias; I63: cerebral infarction; N05: unspecified nephritic syndrome; K29: gastritis and duodenitis; J18: pneumonia, organism unspecified; I10: essential (primary) hypertension; E79: disorders of purine and pyrimidine metabolism; D64: other anaemias; N04: nephrotic syndrome; N02: recurrent and persistent haematuria; N28: other disorders of kidney and ureter, not elsewhere classified; N20: calculus of kidney and ureter; M10: Gout.

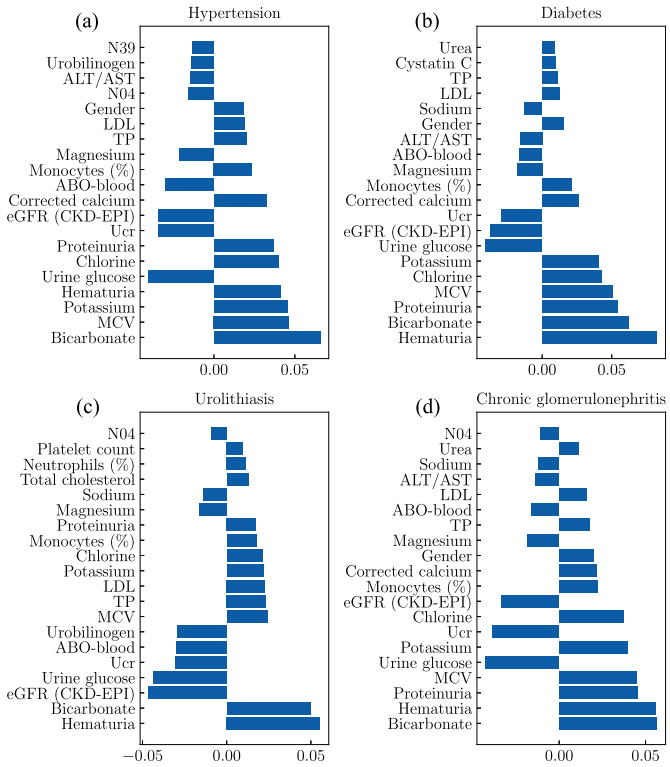

Fig. 4. Top 20 important features identified by DNN with DeepLIFT for the individuals with different etiologies: hypertension (a), diabetes (b), urolithiasis (c) and chronic glomerulonephritis (d).

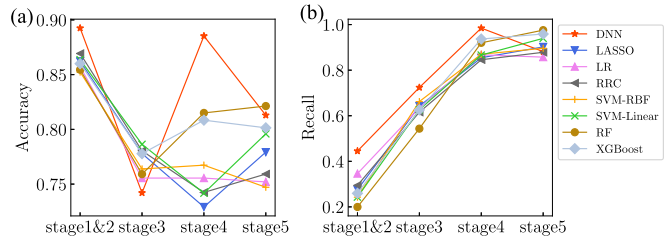

Fig. 5. The accuracy (a) and recall (b) of all the models for people with CKD at different stages. The stages for people with CKD are determined by their first eGFR record.

D. Performance for People At Different Stages

We divided all subjects into four groups according to their first eGFR record to test the predictive ability of all models for people at different stages of CKD. The prediction accuracy and recall of all the models in this study are reported in Fig. 5. It is noted that all the machine learning models achieved similar performances except for people at stage 4. These machine learning models achieved the lowest and highest accuracy for people suffering from stage 3 or stage 4, and stages 1&2, respectively, and the recall increased from stages 1&2 to 5. In contrast, the DNN model generally outperformed other machine learning models in view of both accuracy and recall, except for slighter lower accuracy for people at stage 3 and lower recall for people in stage 5. More specifically, for people at stage 4, the DNN model gets the highest accuracy (0.8853) and the highest recall (0.9846) at the same time, indicating that the DNN model not only accurately predicts whether these people might progress to ESRD, but it also identifies these people more comprehensively.

IV. Discussion

Hematuria, potassium and proteinuria were screened as important independent risk predictors for the progression of CKD patients based on the machine learning model and deep learning model in this study. More importantly, hematuria was the most important risk factor for the progression of DN and urolithiasis in this study, which is inconsistent with clinicians' inherent knowledge. Even though the deep learning model used in our study is not specifically designed, this model still achieves superior performance compared with other machine-learning-based models (Table II and Figs. 2 and 5). The reasons behind its superior performance are the complex non-linear relationships between the input features and output predictions. We also observed that in the machine-learning-based models, the non-linear models better describe the ESRD prediction task, such as the better performance achieved by SVM-RBF compared with SVM-Linear. Considered together, these findings indicate that the relationship between the features of CKD patients and their impacts on ESRD cannot be described by simple linear equations, which were commonly used in previous studies [38].

LASSO, widely adopted in many medical studies [39], [40], [41], [42], has not shown reliable predictive power and intelligence in this study. The critical features identified by LASSO include more comorbidities while ignoring widely-used markers in clinical treatment, such as eGFR [43]. Previous studies have shown that eGFR and the reduction of eGFR are solid markers of CKD progression [43]. In contrast, the critical features identified by DNN-DeepLIFT, RF, and XGBoost are more consistent with clinical knowledge. For example, they all identify eGFR as a critical feature of the progression of CKD. Furthermore, RF and XGBoost also identify serum creatine and cystatin C (both are within top three critical features), which are used to compute the value of eGFR, and show a high correlation with ESRD [44]. ACR is a commonly-used marker for the clinical evaluation of CKD progress and guiding treatment [45], which is also identified by RF and XGBoost. This further provides confidence in these models.

In previous studies, deep learning technologies are usually assumed to be black-box models [46], lacking interpretability even though achieving outstanding performance. We addressed this challenge by introducing novel attribution algorithms, such as Integrated Gradients and DeepLIFT. Three of four attribution algorithms achieve similar patterns of feature contribution weights, which is more consistent with clinical studies [37], except for GradientSHAP. In addition to common features such as eGFR and proteinuria, DNN-DeepLIFT also found some markers that were less reported in clinical studies compared with eGFR, such as hematuria, and considered hematuria to be the most important predictor of the progression of DN and kidney stones, and it is also an important predictor of primary glomerular disease. However, eGFR and proteinuria are clinically recognized as the most important independent risk factors

for CKD progression, and hematuria is often ignored [47], [48], [49]. The reason for the differences may be that previous studies did not distinguish the etiology of CKD when evaluating the prognosis of CKD. Glomerular disease, one of the pathological types of CKD, can cause visible or invisible red blood cells in the urine (hematuria) [50]. DN is the main microvascular complication of diabetes, which can easily cause glomerulosclerosis and damage the glomerular filtration barrier, which may increase the risk of erythrocyte leakage in the glomerulus [51]. In addition, bicarbonate and potassium are the most critical features for individuals with CKD caused by hypertension. Hypertension not only leads to decreased nephron mass, but also increases sodium retention and extracellular volume expansion [52], which may be the reason why electrolyte levels were found to be important factors affecting the progression of CKD due to hypertension. Our study may provide clinicians with a different perspective, that is hematuria may need to be considered as a more important factor in the progression of CKD in patients with diabetes and kidney stones, but this also requires larger cohort studies and external validation.

Potassium and Ucr are identified as critical indicators screened by the three models in our study. Previous studies have shown that low urine potassium excretion is related to CKD progression [53]. The impaired function of the "sodium-potassium pump" in the renal tubules leads to decreased urinary potassium excretion. Wilson et al. found that the appearance of low Ucr is an important risk marker for the adverse consequences of CKD, which is also consistent with our research findings [37].

Timely and accurate prediction of whether the individuals may progress to ESRD within a given duration is critical to determine the most appropriate treatment plan. Thus, for people in an early stage of CKD (stage 1-3), we should identify those who will progress to ESRD as soon as possible, to achieve early intervention and early treatment. Meanwhile, for people in the more advanced stages (stages 4 and 5), it is important to recommend development of good lifestyle and eating habits and to start using drugs such as angiotensin-converting enzyme inhibitors to slow the progression of CKD, in order to avoid premature initiation of hemodialysis treatment and kidney transplantation [54]. Among all the models used in this study, we found that people with stage 3 CKD were the most difficult for accurate prediction (Fig. 5(a)), while all the models achieved the best performance for people with stage 4 disease. This phenomenon may be due to the strong compensatory ability of the kidneys [55]. Some people suffering from stage 3 CKD have no significant abnormalities in view of some clinical indicators. However, when entering CKD stage 4, compensatory mechanisms are overcome, and multiple test indicators showed significant changes. The DNN model continues to demonstrate better power compared with other machine-learning-based models, except for people in stage 5. We took a careful look at the detailed predictions of these models and found that all the machine-learning-based models predicted that all people in stage 5 would progress to ESRD within three years, achieving a lower accuracy but a higher recall (nearly 1.0), which conflicts with real life. In contrast, the DNN model has been better trained and appeared to make more accurate predictions for each individual (a much higher accuracy of 0.90). The relatively lower recall value is possibly due to the limited number of people at stage 5 CKD in the test dataset (n = 34).

V. Conclusion

This study concluded that the DNN model has better performance in predicting the progression of CKD patients to ESRD compared with other machine learning-based models. Furthermore, it provides a potentially more important and different perspective for clinicians' understanding of CKD. That is, hematuria may be an important predictor of the progression of DN and urolithiasis. We compared the DNN model with other machine learning-based models and found that the DNN model performed the best in all CKD stages.

References

[1] L. Zhang et al., "Prevalence of chronic kidney disease in China: A cross-sectional survey," *Lancet*, vol. 379, no. 9818, pp. 815–822, 2012.
[2] M. C. Thomas, M. E. Cooper, and P. Zimmet, "Changing epidemiology of type 2 diabetes mellitus and associated chronic kidney disease," *Nature Rev. Nephrol.*, vol. 12, no. 2, pp. 73–81, 2016.
[3] K. Kalantar-Zadeh, T. H. Jafar, D. Nitsch, B. L. Neuen, and V. Perkovic, "Chronic kidney disease," *Lancet*, vol. 398, no. 10302, pp. 786–802, 2021.
[4] T. H. Hostetter, "Progression of renal disease and renal hypertrophy," *Annu. Rev. Physiol.*, vol. 57, no. 1, pp. 263–278, 1995.
[5] J. Perl and J. M. Bargman, "Peritoneal dialysis: From bench to bedside and bedside to bench," *Amer. J. Physiol.-Renal Physiol.*, vol. 311, no. 5, pp. F999–F1004, 2016.
[6] M. Taal and B. Brenner, "Predicting initiation and progression of chronic kidney disease: Developing renal risk scores," *Kidney Int.*, vol. 70, no. 10, pp. 1694–1705, 2006.
[7] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, 2017.
[8] A. S. Levey et al., "Change in albuminuria and GFR as end points for clinical trials in early stages of CKD: A scientific workshop sponsored by the national kidney foundation in collaboration with the us food and drug administration and European medicines agency," *Amer. J. Kidney Dis.*, vol. 75, no. 1, pp. 84–104, 2020.
[9] A. S. Levey et al., "GFR decline as an end point for clinical trials in CKD: A scientific workshop sponsored by the national kidney foundation and the us food and drug administration," *Amer. J. Kidney Dis.*, vol. 64, no. 6, pp. 821–835, 2014.
[10] W. F. Keane et al., "Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: The RENAAL study," *Clin. J. Amer. Soc. Nephrol.*, vol. 1, no. 4, pp. 761–767, 2006.
[11] D. N. Koye et al., "Risk of progression of nonalbuminuric CKD to end-stage kidney disease in people with diabetes: The CRIC (chronic renal insufficiency cohort) study," *Amer. J. Kidney Dis.*, vol. 72, no. 5, pp. 653–661, 2018.
[12] N. Tangri et al., "A predictive model for progression of chronic kidney disease to kidney failure," *Jama*, vol. 305, no. 15, pp. 1553–1559, 2011.
[13] J. Tang, Q. Su, B. Su, S. Fong, W. Cao, and X. Gong, "Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition," *Comput. Methods Prog. Biomed.*, vol. 197, 2020, Art. no. 105622.
[14] E. Kalafi, N. Nor, N. Taib, M. Ganggayah, C. Town, and S. Dhillon, "Machine learning and deep learning approaches in breast cancer survival prediction using clinical data," *Folia Biologica*, vol. 65, no. 5/6, pp. 212–220, 2019.
[15] R. Brehar et al., "Comparison of deep-learning and conventional machine-learning methods for the automatic recognition of the hepatocellular carcinoma areas from ultrasound images," *Sensors*, vol. 20, no. 11, 2020, Art. no. 3085.
[16] K. Zhang et al., "Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 533–545, 2021.
[17] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[18] T. Ferguson et al., "Development and external validation of a machine learning model for progression of CKD," *Kidney Int. Rep.*, vol. 7, no. 8, pp. 1772–1781, 2022.
[19] F. P. Schena et al., "Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with immunoglobulin a nephropathy," *Kidney Int.*, vol. 99, no. 5, pp. 1179–1188, 2021.
[20] A. S. Levey et al., "A new equation to estimate glomerular filtration rate," *Ann. Intern. Med.*, vol. 150, no. 9, pp. 604–612, 2009.
[21] C.-Y. Hsu, E. Vittinghoff, F. Lin, and M. G. Shlipak, "The incidence of end-stage renal disease is increasing faster than the prevalence of chronic renal insufficiency," *Ann. Intern. Med.*, vol. 141, no. 2, pp. 95–101, 2004.
[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.
[26] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[28] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, vol. 26. New York, NY, USA: Springer, 2004.
[29] S. Menard, *Applied Logistic Regression Analysis*, no. 106. London, U.K.: SAGE, 2002.
[30] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[32] E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research," *J. Data Anal. Inf. Process.*, vol. 7, no. 4, pp. 190–207, 2019.
[33] S. Devika, L. Jeyaseelan, and G. Sebastian, "Analysis of sparse data in logistic regression in medical research: A newer approach," *J. Postgraduate Med.*, vol. 62, no. 1, 2016, Art. no. 26.
[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[35] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[36] T. Chen and C. Guestrin, "XGboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.
[37] F. P. Wilson et al., "Urinary creatinine excretion, bioelectrical impedance analysis, and clinical outcomes in patients with CKD: The CRIC study," *Clin. J. Amer. Soc. Nephrol.*, vol. 9, no. 12, pp. 2095–2103, 2014.
[38] K. Matsuo et al., "Survival outcome prediction in cervical cancer: Cox models vs deep-learning model," *Amer. J. Obstet. Gynecol.*, vol. 220, no. 4, pp. 381.e1–381.e14, 2019.
[39] Y.-Q. Huang et al., "Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer," *J. Clin. Oncol.*, vol. 34, no. 18, pp. 2157–2164, 2016.
[40] J. Y. Kim et al., "Incorporating diffusion-and perfusion-weighted MRI into a radiomics model improves diagnostic performance for pseudoprogression in glioblastoma patients," *Neuro- Oncol.*, vol. 21, no. 3, pp. 404–414, 2019.
[41] G.-W. Ji et al., "A radiomics approach to predict lymph node metastasis and clinical outcome of intrahepatic cholangiocarcinoma," *Eur. Radiol.*, vol. 29, no. 7, pp. 3725–3735, 2019.
[42] H. U. Zacharias et al., "A predictive model for progression of CKD to kidney failure based on routine laboratory tests," *Amer. J. Kidney Dis.*, vol. 79, no. 2, pp. 217–230, 2022.
[43] M. Raman, R. J. Middleton, P. A. Kalra, and D. Green, "Estimating renal function in old people: An in-depth review," *Int. Urol. Nephrol.*, vol. 49, no. 11, pp. 1979–1988, 2017.
[44] L. A. Inker et al., "Estimating glomerular filtration rate from serum creatinine and cystatin C," *New England J. Med.*, vol. 367, no. 1, pp. 20–29, 2012.
[45] J. Coresh et al., "Change in albuminuria and subsequent risk of end-stage kidney disease: An individual participant-level consortium meta-analysis of observational studies," *Lancet Diabetes Endocrinol.*, vol. 7, no. 2, pp. 115–127, 2019.
[46] C. Rudin and J. Radin, "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition," *Harvard Data Sci. Rev.*, vol. 1, pp. 10–1162, 2019.
[47] S. Matsuo et al., "Clinical guides for immunoglobulin a (IGA) nephropathy in Japan, third version," *Jpn J. Nephrol*, vol. 53, no. 2, pp. 123–35, 2011.
[48] Y. Yuzawa et al., "Evidence-based clinical practice guidelines for IGA nephropathy 2014," *Clin. Exp. Nephrol.*, vol. 20, no. 4, pp. 511–535, 2016.
[49] N. Tangri et al., "Risk prediction models for patients with chronic kidney disease: A systematic review," *Ann. Intern. Med.*, vol. 158, no. 8, pp. 596–603, 2013.
[50] M. A. Perazella, "The urine sediment as a biomarker of kidney disease," *Amer. J. Kidney Dis.*, vol. 66, no. 5, pp. 748–755, 2015.
[51] H.-J. Anders, T. B. Huber, B. Isermann, and M. Schiffer, "CKD in diabetes: Diabetic kidney disease versus nondiabetic kidney disease," *Nature Rev. Nephrol.*, vol. 14, no. 6, pp. 361–377, 2018.
[52] E. Ku, B. J. Lee, J. Wei, and M. R. Weir, "Hypertension in CKD: Core curriculum 2019," *Amer. J. Kidney Dis.*, vol. 74, no. 1, pp. 120–131, 2019.
[53] J. He et al., "Urinary sodium and potassium excretion and CKD progression," *J. Amer. Soc. Nephrol.*, vol. 27, no. 4, pp. 1202–1212, 2016.
[54] B. Lerner, S. Desrochers, and N. Tangri, "Risk prediction models in CKD," *Seminars Nephrol.*, vol. 37, no. 2, pp. 144–150, 2017.
[55] V. Soi and J. Yee, "Sodium homeostasis in chronic kidney disease," *Adv. Chronic Kidney Dis.*, vol. 24, no. 5, pp. 325–331, 2017.