future internet

MDPI

Article

IgA Nephropathy Prediction in Children with Machine Learning Algorithms

Ping Zhang [1,2], Rongqin Wang [2,3,*] and Nianfeng Shi [3]

[1] School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471000, China; zping@haust.edu.cn
[2] School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China
[3] School of Computer and Information Engineering, Luoyang Institute of Science and Technology, Luoyang 471000, China; shinf@lit.edu.cn
* Correspondence: Wangrongqin0827@163.com; Tel.: +86-1563-900-7675

Received: 15 November 2020; Accepted: 14 December 2020; Published: 17 December 2020

Abstract: Immunoglobulin A nephropathy (IgAN) is the most common primary glomerular disease all over the world and it is a major cause of renal failure. IgAN prediction in children with machine learning algorithms has been rarely studied. We retrospectively analyzed the electronic medical records from the Nanjing Eastern War Zone Hospital, chose eXtreme Gradient Boosting (XGBoost), random forest (RF), CatBoost, support vector machines (SVM), k-nearest neighbor (KNN), and extreme learning machine (ELM) models in order to predict the probability that the patient would not reach or reach end-stage renal disease (ESRD) within five years, used the chi-square test to select the most relevant 16 features as the input of the model, and designed a decision-making system (DMS) of IgAN prediction in children that is based on XGBoost and Django framework. The receiver operating characteristic (ROC) curve was used in order to evaluate the performance of the models and XGBoost had the best performance by comparison. The AUC value, accuracy, precision, recall, and f1-score of XGBoost were 85.11%, 78.60%, 75.96%, 76.70%, and 76.33%, respectively. The XGBoost model is useful for physicians and pediatric patients in providing predictions regarding IgAN. As an advantage, a DMS can be designed based on the XGBoost model to assist a physician to effectively treat IgAN in children for preventing deterioration.

Keywords: IgA nephropathy; machine learning; XGBoost; prediction; decision-making system

1. Introduction

Immunoglobulin A nephropathy (IgAN) is the most common primary glomerular disease all over the world and it is a major cause of end-stage renal disease (ESRD). Additionally, there are similarities and differences of IgAN in adults and children. Clinical studies showed that the survival rate of IgAN patients over 18 years old was 70–80% within 10 years. Among them, 20–30% patients will progress to ESRD and need renal replacement therapy [1]. Wyatt et al. [2] found that the five-year survival rate of children with IgAN was 94–98%, and the 20-year survival rate was 70–89%. Although IgAN in children has relatively lower incidence and less progress in childhood as compared to adults, there is a continuous hazard of progression as children grow older [3]. The prognosis of children with IgAN varies from remission to progression to ESRD. Early detection and effective intervention are important in improving the outcome of IgAN, and the renal biopsy is still the cornerstone of the correct diagnosis of IgAN. However, most patients have entered ESRD at the time of diagnosis, since it is difficult for patients to accept invasive renal biopsy. It is difficult to determine a fixed treatment plan for IgAN in children because it is more troublesome to gather up the course of IgAN in children than in adults. Therefore, it is necessary to develop new predictive models for the progression of IgAN in children in

order to guide the selection of cases to be treated [4]. Machine learning has been successfully applied in many fields, providing a shield for diagnosis of IgAN for non-invasive diagnose in asymptomatic cases.

Traditional medical treatment methods rely entirely on doctors' diagnosis and the treatment of patients. In this way, it is difficult to distinguish between diseases with similar symptoms and discover the hidden diseases, leading to misdiagnosis, which may delay the patient's treatment or endanger the patient's life. With the explosive growth of electronic medical records, machine learning in medicine has attracted the attention of many scholars [5–7]. Ali et al. proposed intelligent smart healthcare monitoring systems that are based on machine learning approaches to extract useful features from the collected healthcare data of patients, reduce the dimensionality of the data, and improve the classification precision [8,9]. ESRD has the characteristics of high disability, high mortality, and high medical expenses. A lot of researches were invested in the early detection and prediction of kidney disease by machine learning algorithms in order to avoid the occurrence of ESRD, such as the prediction of the kidney disease [10,11], the formulation of the best course of treatment [12], the evaluation of the risk stratification [13], and so on.

The goal of treating IgAN is still to delay the progression of IgAN, since the exact treatment plan for IgAN has not yet been established. Researchers mainly focus on the prognosis [14–16], risk stratification [17–20], and ESRD prediction of IgAN [21,22], whose patients are over 18 years old. The researches on IgAN in children focused on pathological characteristics and statistics [23–25], but there are few on the prediction of IgAN in children based on machine learning. Machine learning algorithms are an effective method for solving the high-dimensional problem of medical data, and it plays an important role in smart medical care.

In this paper, a novel method is proposed for predicting the probability of children patients with IgAN reaching ESRD in five years. The main contributions of this paper are:

- A dataset about IgAN in children was created. And the chi-square test was used to extract the most useful features from the dataset.
- EXtreme Gradient Boosting (XGBoost) was adopted in order to predict whether IgAN disease in children patients would reach ESRD or not within five years using a new dataset instead of the traditional clinical pathology. A decision-making system that was based on the XGBoost algorithm was designed with the Django framework.
- Comparation of the performance of XGBoost with random forest (RF), CatBoost, support vector machines (SVM), k-nearest neighbor (KNN), and extreme learning machine (ELM) was conducted.

2. Materials and Methods

2.1. Dataset

The initial data include electronic medical records of 1167 patients aged 0–18 years old from 2003 to 2019, which came from the electronic medical records of the Nanjing Eastern War Zone Hospital. After the initial data cleaning and deletion of information with missing values, the dataset contained 1146 records. Each record contained 37 attributes, where the first 36 attributes are independent variables that correspond to the patient information, while the remaining attribute is the dependent variable of clinical interest. The 36 independent variables were collected according to five aspects, which are epidemiology, blood test, urine test, renal pathology, and treatment options. In detail, age, sex and hypertension are classified as epidemiology. Serum creatinine (Scr), cholesterol (CHOL), triglycerides (TG), albumin (ALB) complement C3, and glomerular filtration rate (eGFR) are blood test indicators. Urine tests include urine C3 (Ur_C3), α2-m, urine NAG enzyme (Ur_NAG), urine RBP (Ur_RBP), and uric acid (UA). ACEI_ARB, immunosuppression therapy, lipid lowering, and tonsillectomy are the treatment options. The remaining 19 attributes M, E, S, T, C, IgA, IgG, IgM, C3, C4, C1q, loop necrosis, focal segmental glomerular sclerosis (FSGS), glomerulosclerosis, arterial hyaline degeneration, crescent ratio, medullary interstitial fibrosis, thickening and stratification of elastic layer of interlobular artery, and vacuolar degeneration of arteriole smooth muscle cells are included in renal pathology. The dependent variable

was whether the patient would reach the end-stage renal disease within five years, which was expressed by ESRD.

These independent variable attributes are divided into continuous independent variables and categorical independent variables according to whether they are continuous values or not, as shown in Tables 1 and 2. Specifically, Table 1 shows the range, mean, and standard deviation for continuous independent variables and Table 2 displays the possible values, numeric value, as well as the number of records for categorical independent variables. Table 3 displays the possible values, numeric values, as well as the number of records for categorical dependent variables. The ESRD is distributed in 567 and 578 records, which correspond to the categories yes and no, respectively, which indicates that patients reach the final stage of IgAN or do not reach that within five years, respectively.

Table 1. Continuous independent variables.

Variable	Range	Mean	Standard Deviation
Age (years)	0–18	14	3.59
Scr (μmol/L)	16.00–1154.50	71.12	67.86
CHOL (mmol/L)	0.20–19.00	5.49	2.70
TG (mmol/L)	0.17–22.00	1.62	1.25
UA (μmol/L)	0.53–777.00	347.87	97.63
ALB (g/mL)	7.60–65.20	41.88	120.48
Complement C3 (g/L)	0.24–25.00	1.06	0.74
FSGS	0.00–61.50	4.20	8.75
Spherically sclerotic	0.00–84.20	3.75	9.21
Crescent ratio	0.00–77.80	5.25	9.50
eGFR ($mL{\cdot}min^{-1}{\cdot}(1.72\ m^2)^{-1}$)	4.90–141.75	111.77	21.11

Scr-Serum creatinine; CHOL-cholesterol; TG-triglycerides; UA-uric acid; ALB-albumin; FSGS-focal segmental glomerular sclerosis; eGFR-glomerular filtration rate.

Table 2. Categorical independent variables.

Variable	Possible Values	Numeric Value	Number of Records	Percentage (%)
Gender	M/F	1/0	776/369	67.78/32.22
Ur_C3	normal/abnormal	0/1	347/798	30.31/69.69
α2-m	normal/abnormal	0/1	214/931	18.69/81.31
Ur_NAG	normal/abnormal	0/1	909/236	79.39/20.61
Ur_RBP	normal/abnormal	0/1	164/981	14.32/85.68
M		0/1/2/3	283/562/267/33	24.72/49.08/23.32/2.88
E		0/1/2	756/386/3	66.03/33.71/0.26
S		0/1	728/417	63.58/36.42
T		0/1/2/3	845/251/42/7	73.80/21.92/3.67/0.61
C		0/1/2	648/444/53	56.59/38.78/4.63
IgA		0/1/2/3	16/31/1082/16	1.40/2.70/94.50/1.40
IgG		0/0.5/1/2	882/1/177/85	77.03/0.09/15.46/7.42
IgM		0/0.5/1/2	669/13/418/45	58.43/1.14/36.50/3.93
C3		0/0.5/1/2	207/9/173/756	18.08/0.79/15.11/66.02
C4		0/1	1133/12	98.95/1.05
C1q		0/1/2	1094/49/2	95.55/4.28/0.17
Hypertension	yes/no	1/0	322/823	28.12/71.88
ACEI_ARB		0/1	862/283	75.28/24.72
Immunosuppressive therapy		0/1	726/419	63.41/36.59
Lipid lowering		0/1	738/407	64.45/35.55
Tonsillectomy		0/1	1124/21	98.17/1.83
Loop necrosis		0/1/2/3	1039/91/7/8	90.74/7.95/0.61/0.70

Table 2. *Cont.*

Variable	Possible Values	Numeric Value	Number of Records	Percentage (%)
Arterial hyaline degeneration		0/1/5	1015/129/1	88.65/11.26/0.09
Medullary interstitial fibrosis		0/1/2/3	1090/39/11/5	95.20/3.41/0.96/0.43
Thickening and stratification of elastic layer of interlobular artery		0/1	1108/37	96.77/3.23
Vacuolar degeneration of arteriole smooth muscle cells		0/1	1085/60	94.76/5.24

Ur_C3-urine C3; Ur_NAG-urine NAG enzyme; Ur_RBP-urine RBP.

Table 3. Categorical dependent variable.

Variable	Possible Values	Numeric Value	Number of Records	Percentage (%)
ESRD	yes	1	567	49.52
	no	0	578	50.48

2.2. Feature Selection

We evaluated the importance and relevance of predictors with ESRD by the chi-square test for the purpose of identifying significant predictors of ESRD to be applied as inputs for the data mining methods. Feature analysis, as illustrated in Table 4, displays that the importance and relevance of all the predictors with ESRD. The *p*-value is called Pierce the correlation coefficient. A *p*-value of less than 0.05 means that there is significant difference. The score stands for the Chi-square statistics, which can be calculated according to Equation (1).

$$\chi^2 = \sum (A - T)^2 / T \quad (1)$$

where A represents the actual value and T represents the theoretical value.

Table 4. Chi-square analysis results of predictors.

Variable	Score	*p*-Value
Scr	827.885	0
FSGS	716.552	0
Crescent ratio	658.534	0
ALB	621.352	0
UA	461.223	0
Spherically sclerotic	321.243	0
CHOL	279.544	0
Ur_NAG	119.224	0
eGFR	163.170	0
TG	56.108	0
E	27.470	0
T	23.479	0
C	18.653	0
M	18.451	0
IgM	16.973	0
C3	13.189	0
Gender	0.203	0.652
Ur_C3	0.024	0.878
α2-m	7.233	0.007
Ur_RBP	5.823	0.016
Complement C3	0.003	0.953
S	5.760	0.016

Table 4. *Cont.*

Variable	Score	*p*-Value
IgA	0.025	0.875
IgG	0.816	0.366
C4	0.001	0.973
C1q	3.444	0.063
Hypertension	11.592	0.001
ACEI_ARB	0.211	0.646
Immunosuppressive therapy	0.232	0.630
Lipid lowering	9.725	0.002
Tonsillectomy	0.488	0.485
Loop necrosis	0.024	0.877
Arterial hyaline degeneration	0.012	0.911
Medullary interstitial fibrosis	5.655	0.017
Thickening and stratification of elastic layer of interlobular artery	6.373	0.012
Vacuolar degeneration of arteriole smooth muscle cells	0.490	0.484

The higher the score, the more important the attribute. Sixteen features, which are shown in the first 16 rows of Table 4, were selected based on *p*-value that is equal to 0 and score greater than 10 for data dimension reduction in this paper.

2.3. Model

In this section, XGBoost has been introduced as the best performing algorithm. For processing high-dimensional data, dimensionality reduction, feature extraction, etc., it has a higher accuracy than traditional algorithms. XGBoost is an improved gradient boosting algorithm [26]. The innovation of XGBoost lies in the optimization of the objective function with the second-order Taylor expansion. It merges multiple weak classifiers in order to evolve into a strong classifier, and the base classifier is a classification and regression tree (CART).

The objective function of XGBoost consists of a loss function and a regularization term, which are defined, as follows:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where f_k is the function expression of the k-th tree model and y_i and $\hat{y}_i$ are the true label and predicted value of the i-th sample x_i, respectively. XGBoost is an additive model, so the predicted value is the sum of the predicted values of each tree, i.e., $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$.

The sum of the complexity of K trees is used as a regularization term for preventing the model from over-fitting. Assuming that the tree model that is trained on the t-th iteration is f_t, then:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

Substitute Equation (3) into Equation (2) to obtain Equation (4).

$$Obj^{(t)} = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

Expand the loss function $\sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t)})$ to second-order Taylor

$$\begin{aligned} Obj^{(t)} &\cong \sum_{i=1}^{t}\left[l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))\right] = \\ &\sum_{\mathrm{i}=1}^{\mathrm{t}}\left[l(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i))\right] + \Omega(f_t) \end{aligned} \tag{5}$$

where g_i and h_i are the first-order partial derivative and second-order partial derivative of the loss function l with regard to $\hat{y}_i^{(t-1)}$, respectively, and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} {\omega_j}^2$.

Define the leaf node $I_j = \{i | q(x_i) = j\}$, and the objective function is finally reduced to

$$Obj^{(t)} = \sum_{j}^{T}\left[\left(\sum_{i \in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i\right){\omega_j}^2\right] + \gamma T \tag{6}$$

During the training process of the XGBoost model, when the t-th tree is established, the greedy strategy is adopted in order to split the tree nodes. Every time the tree node splits into two left and right leaf nodes, it will bring gain to the loss function, which is defined, as follows:

$$\begin{aligned} &Gain = Obj_{L+R} - (Obj_L + Obj_R) = \\ &\frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\left(\sum_{i \in I_L} h_i\right) + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\left(\sum_{i \in I_R} h_i\right) + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\left(\sum_{i \in I} h_i\right) + \lambda}\right] - \lambda \end{aligned} \tag{7}$$

If $Gain > 0$, then the result of this split is added to the model construction.

XGBoost provides three calculation methods for feature importance. The first way is gain, which refers to the average gain of the feature when it is used in trees. The second way is weight, which is the number of times that a feature is used to split the data across all trees. The last way is cover, which relates to the average coverage of the feature when it is used in trees. In this study, the gain method was mainly used for calculating feature importance.

2.4. Performance Evaluation

ROC curve and area under curve (AUC) were used in order to evaluate the pros and cons of a binary classifier (binary classifier) in our paper. The abscissa of the curve is the false positive rate (FPR) and the ordinate is the true positive rate (TPR).

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

where TP stands for True Positive, TN represents True Negative, FP symbolizes False Positive, and FN means False Negative.

Different (FPR, TPR) points can be obtained by adjusting the threshold value that is predicted by the model, and these points can be connected into a curve, which is the ROC curve. After the curve is drawn, qualitatively analyze the model that if you want, you need to calculate the AUC area.

AUC refers to the area under the ROC curve [27]. Calculating the AUC value only needs to integrate FPR on the horizontal axis of ROC. In a real scene, the ROC curve is generally above the line $y = x$, so the value of AUC is generally between 0.5 and 1. The larger the value of AUC, the better the performance of the model.

In addition, compare the recall rate and f_1 score of all the classifiers.

$$recall = TPR = \frac{TP}{TP + FN} \tag{10}$$

$$f_1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

The training set and test set were randomly selected with a ratio of 3 to 1, during model training. Training and testing applied the 10-fold cross validation method to separate the dataset into several partitions or fold, calculated the average of accuracy from all folds. In addition, the above-mentioned performances, such as ROC, AUC, recall, and f1 scores were used to evaluate all techniques.

3. System Implementation

XGBoost, which showed the best classification performance, was used in order to implement an online decision-making system. The core framework of the system is the Django framework, which was used to apply the machine learning models and build the web tool, and it made use of two programming languages, including Python (version 3.7.0) and HTML (version HTML 5). Django is an open source python web framework. Programmers can easily and quickly create high-quality, easy-to-maintain, database-driven applications with the framework. In addition, in the Django framework, it also contains many powerful third-party plug-ins, which makes Django highly extensible.

In the current implementation of the system, an HTML communicates with the Python service and formats the information that is shown to the user. The training model of the system that is shown in Figure 1 can be used for a single prediction. Additionally, Figure 1 shows a screenshot of the initial web page. When users enter this page, fill in the data of the initial web page according to the feature description in Section 2.1. The system backend predicts whether the patient will reach ESRD or not in five years based on the data that were submitted by the initial web page. Moreover, the web-based decision-making assistance system will obtain the probability of a patient reaching and not reaching ESRD within five years, which may help doctors to alert to some borderline patients. Figure 2 shows the prediction outcome of the decision-making system (DMS) that is based on XGBoost.

Index	Unit	Measured value	Reference range
Age	years old		0~18
Gender		Male	
Serum creatinine	μmol/L		16.00-1154.50
Cholesterol	mmol/L		0.20-19.00
Triglycerides	mmol/L		0.17-22.00
Uric acid	μmol/L		0.53-777.00
Ur_NAG		1	0/1
Albumin	g/mL		7.60-65.20
M			0/1/2/3
E			0/1/2
T			0/1/2/3
C			0/1/2
IgM			0/0.5/1/2
C3			0/0.5/1/2
Glomerulosclerosis			0.00-84.20
FSGS			0.00-61.50
Crescent ratio			0.00-77.80
eGFR			4.90-141.75

Submit

Figure 1. A screenshot of the initial web page of the implemented decision-making system (DMS).

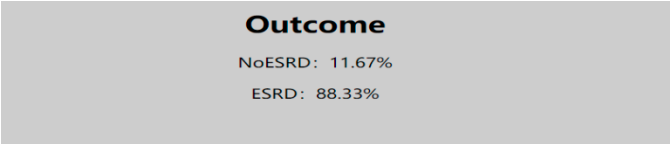

Figure 2. Outcome of the eXtreme Gradient Boosting (XGBoost)-based DMS.

4. Results

The best performing model was selected in order to predict whether children suffering from IgAN would reach the end-stage renal disease after five years among six kinds of machine learning algorithms of XGBoost, RF, CatBoost, KNN, SVM, and ELM.

Table 5 shows the accuracy, precision, recall, f1-score, and AUC values of XGBoost, RF, CatBoost, KNN, SVM, and ELM models. It can be concluded from the table that all of the performance indicators of XGBoost are the best, AUC, that accuracy, precision, recall, and f1-score are 85.11%, 78.60%, 75.96%, 76.70%, and 76.33%, respectively. Table 6 illustrates the importance scores of XGBoost on 16 variables that were selected from Table 4. The AUC indicators of RF and XGBoost are almost equal, but the accuracy, precision, recall, and f1-score of the RF model are all smaller than those of the XGBoost model.

Table 5. Results for Immunoglobulin A nephropathy (IgAN) prediction.

Algorithm	Accuracy	Precision	Recall	F1_Score	AUC
XGBoost	0.7860	0.7596	0.7670	0.7633	0.8511
RF	0.7642	0.7426	0.7282	0.7353	0.8507
CatBoost	0.7642	0.7379	0.7379	0.7379	0.8454
KNN	0.7555	0.7327	0.7184	0.7255	0.8090
SVM	0.7642	0.7333	0.7476	0.7404	0.8272
ELM	0.7598	0.7264	0.7476	0.7368	0.8174

Table 6. The corresponding variable importance score using XGBoost.

Variables	Importance Score
Ur_NAG	0.191591
ALB	0.175141
CHOL	0.104368
Crescent ratio	0.085010
FSGS	0.067488
Scr	0.062304
TG	0.050906
Spherically sclerotic	0.034839
UA	0.034522
M	0.033158
IgM	0.032431
T	0.031299
C3	0.025893
C	0.025279
eGFR	0.024145
E	0.021627

Figure 3 depicts the ROC curve of six machine learning models. From the figure, it can be seen that the ROC curves of the XGBoost and RF models are close at the top; the ROC curve of the KNN model is at the bottom, which means that the XGBoost and RF models have the best performance and the KNN model has the worst performance according to the ROC curve. The ROC curves of the CatBoost, SVM, and ELM models are between that of the XGBoost and KNN models.

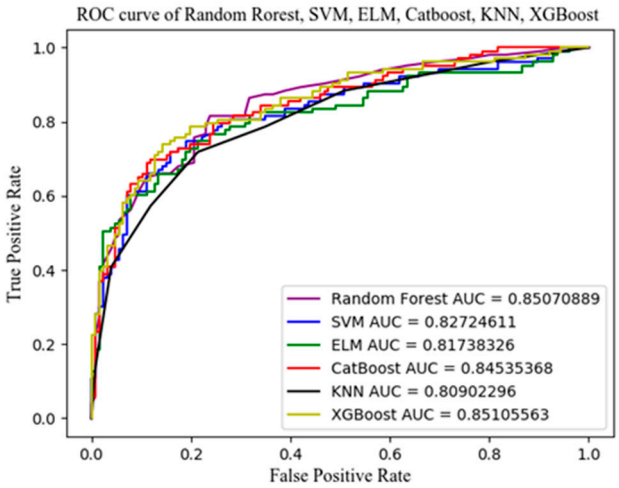

Figure 3. Receiver operating characteristic (ROC) curve of models.

5. Discussion

In the paper, the use of machine learning algorithms for predicting IgAN in children was researched. At first, 16 features (serum creatinine, focal segmental glomerular sclerosis, crescent ratio, albumin, uric acid, glomerulosclerosis, cholesterol, urine NAG enzyme, eGFR, triglycerides, E, T, C, M, IgM, and C3) were chosen as the input of the classifiers by the chi-square test for dimension reduction. Subsequently, the XGBoost, RF, CatBoost, KNN, SVM, and ELM models were applied to predict IgAN in children. Finally, a decision-making system was build based on the best performing model and Django framework. The results that are shown in Table 5 and Figure 3 indicate that the XGBoost model can provide better performance when compared to other models for the medical application that was considered in this study. The AUC value, accuracy, precision, recall, and f1-score of XGBoost were 85.11%, 78.60%, 75.96%, 76.70%, and 76.33%, respectively. While the AUC value, accuracy, precision, recall and f1-score of RF (76.42%, 74.26%, 72.82%, 73.53%, 85.07%), CatBoost (76.42%, 73.79%, 73.79%, 73.79%, 84.54%), KNN (75.55%, 73.27%, 71.84%, 72.55%, 80.90%), SVM (76.42%, 73.33%, 74.76%, 74.04%, 82.72%), and ELM (75.98%, 72.64%, 74.76%, 73.68%, 81.74%) models are all lower than those of XGBoost. Here, we can highlight an advantage of XGBoost, because we not only need interpretable models to assist clinical decision-making, but also help clinicians to discover hidden factors that affect the disease. The XGBoost algorithm has a regularization term to prevent overfitting. Moreover, XGBoost can specify the default direction of the branch for missing values or specified values, which can greatly improve the efficiency of the algorithm. More importantly, the XGBoost model has high generalization performance and it can clearly output the important scores of each attribute, namely it is interpretable, which is required by clinical medicine.

Despite the potential of this research, there are several limitations. First, the collected dataset is relatively small, and it cannot fully cover kidney cases in children, which leads to inaccurate predictions of special kidney disease cases in children. Therefore, the dataset needs to be increased and feature processing needs to be more refined by applying data mining techniques for predicting IgAN in children. Second, the prediction system is suitable for children with IgAN in the age range of 0–18 years old, but not for adults with IgAN.

In the future, we will devote to improve the accuracy of the model, perfect the system, add database to the system, expand the training dataset, and complement the system with an error correction function.

Author Contributions: Conceptualization, P.Z. and N.S.; methodology, R.W. and N.S.; software, R.W. and N.S.; validation, R.W. and N.S.; formal analysis, P.Z. and R.W.; investigation, P.Z. and N.S.; resources, N.S.; data curation, R.W.; writing—original draft preparation, R.W.; writing—review and editing, P.Z. and R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 11401172, Key Scientific Research Project Plan of Colleges and Universities in Henan Province, grant number 20A520012, and the Pediatric Medical Innovation Team of Jiangsu Province, grant number CXTDA2017022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Damico, G. The Commonest Glomerulonephritis in the World: IgA Nephropathy. *QJM Int. J. Med.* **1987**, *64*, 709–727.
2. Wyatt, R.J.; Kritchevsky, S.B.; Woodford, S.Y.; Miller, P.M.; Roy, S., III; Holland, N.H.; Jackson, E.; Bishof, N.A. IgA nephropathy: Long-term prognosis for pediatric patients. *J. Pediatr* **1995**, *127*, 913–919. [CrossRef]
3. Coppo, R.; Robert, T. IgA nephropathy in children and in adults: Two separate entities or the same disease? *J. Nephrol.* **2020**, *33*, 1219–1229. [CrossRef] [PubMed]
4. Coppo, R. Treatment of IgA nephropathy in children: A land without KDIGO guidance. *Pediatr. Nephrol.* **2020**, *2020*, 1–6. [CrossRef]
5. Jonsson, A. Deep Reinforcement Learning in Medicine. *Kidney Dis.* **2019**, *5*, 18–22. [CrossRef]
6. Toback, F.G. Regeneration after acute tubular necrosis. *Kidney Int.* **1992**, *41*, 226–246. [CrossRef]
7. Agar, J.W.; Webb, G.I. Application of machine learning to a renal biopsy database. *Nephrol. Dialys. Transplant.* **1992**, *7*, 472–478. [CrossRef]
8. Ali, F.; El-Sappagh, S.; Islam, S.M.R.; Ali, A.; Attique, M.; Imran, M.; Kwak, K.-S. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Futur. Gener. Comput. Syst.* **2020**, *114*, 23–43. [CrossRef]
9. Ali, F.; El-Sappagh, S.H.A.; Islam, S.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K.-S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf. Fusion* **2020**, *63*, 208–222. [CrossRef]
10. Leung, R.K.; Wang, Y.; Ma, R.C.; Luk, A.O.; Lam, V.; Ng, M.; So, W.Y.; Tsui, S.K.; Chan, J.C. Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: A prospective case–control cohort analysis. *BMC Nephrol.* **2013**, *14*, 162. [CrossRef]
11. Hamedan, F.; Orooji, A.; Sanadgol, H.; Sheikhtaheri, A. Clinical Decision Support System to Predict Chronic Kidney Disease: A Fuzzy Expert System Approach. *Int. J. Med. Info.* **2020**, *138*, 104134. [CrossRef] [PubMed]
12. Gupta, R.; Sharma, A.; Singh, S.; Dinda, A.K. Rule-based decision support system in the biopsy diagnosis of glomerular diseases. *J. Clin. Pathol.* **2011**, *64*, 862. [CrossRef] [PubMed]
13. Takayuki, K.; Masaki, O.; Akira, K.; Michiharu, K.; Kyoichi, H.; Jun, K.; Masaki, M.; Ryosuke, Y.; Atsushi, S. Risk Prediction of Diabetic Nephropathy via Interpretable Feature Extraction from EHR Using Convolutional Autoencoder. *Stud. Health Technol. Info.* **2018**, *247*, 106–110.
14. Lemley, K.V.; Lafayette, R.A.; Derby, G.C.; Blouch, K.L.; Anderson, L.; Efron, B.; Myers, B.D. Prediction of early progression in recently diagnosed IgA nephropathy. *Nephrol. Dialys. Transplant.* **2007**, *23*, 213–222. [CrossRef] [PubMed]
15. Noh, J.; Punithan, D.; Lee, H.; Lee, J.; Kim, Y.; Kim, D.; McKay, R.B. Machine Learning Models and Statistical Measures for Predicting the Progression of IgA Nephropathy. *Int. J. Soft. Eng. Know. Eng.* **2015**, *25*, 829–849. [CrossRef]
16. Ducher, M.; Kalbacher, E.; François, C.; de Vilaine, J.F.; McGregor, B.; Fouque, D.; Fauvel, J.P. Comparison of a Bayesian Network with a Logistic Regression Model to Forecast IgA Nephropathy. *BioMed Res. Int.* **2013**, *2013*, 686150. [CrossRef]
17. Barbour, S.J.; Reich, H.N. Risk Stratification of Patients with IgA Nephropathy. *Am. J. Kidney Dis.* **2012**, *59*, 865–873. [CrossRef]
18. Barbour, S.J.; John, F. Predicting the future in immunoglobulin A nephropathy: A new international risk prediction tool. *Nephrol. Dialys. Transplant.* **2020**, *35*, 379–382. [CrossRef]
19. Barbour, S.J.; Coppo, R.; Zhang, H.; Liu, Z.-H.; Suzuki, Y.; Matsuzaki, K.; Katafuchi, R.; Er, L.; Espino-Hernandez, G.; Kim, J.; et al. Evaluating a New International Risk-Prediction Tool in IgA Nephropathy. *JAMA Intern. Med.* **2019**, *179*, 942–952. [CrossRef]
20. Chen, T.; Li, X.; Li, Y.; Xia, E.; Qin, Y.; Liang, S.; Xu, F.; Liang, D.; Zeng, C.; Liu, Z. Prediction and Risk Stratification of Kidney Outcomes in IgA Nephropathy. *Am. J. Kidney Dis.* **2019**, *74*, 300–309. [CrossRef]

21. Liu, Y.; Zhang, Y.; Liu, D.; Tan, X.; Tang, X.; Zhang, F.; Xia, M.; Chen, G.; He, L.; Zhou, L.; et al. Prediction of ESRD in IgA Nephropathy Patients from an Asian Cohort: A Random Forest Model. *Kidney Blood Press. Res.* **2018**, *43*, 1852–1864. [CrossRef] [PubMed]
22. Han, X.; Zheng, X.; Wang, Y.; Sun, X.; Xiao, Y.; Tang, Y.; Qin, W. Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients. *Annal. Translat. Med.* **2019**, *7*, 234. [CrossRef] [PubMed]
23. Bertram, J.F.; Goldstein, S.L.; Pape, L.; Schaefer, F.; Shroff, R.C.; Warady, B.A. Kidney disease in children: Latest advances and remaining challenges. *Nat. Rev. Nephrol.* **2016**, *12*, 182–191. [CrossRef] [PubMed]
24. Coppo, R.; Troyanov, S.; Camilla, R.; Hogg, R.J.; Cattran, D.C.; Terence Cook, H.; Feehally, J.; Roberts, I.S.D.; Amore, A.; Alpers, C.E.; et al. The Oxford IgA nephropathy clinicopathological classification is valid for children as well as adults. *Kidney Int.* **2010**, *77*, 921–927. [CrossRef] [PubMed]
25. Shima, Y.; Nakanishi, K.; Sato, M.; Hama, T.; Mukaiyama, H.; Togawa, H.; Tanaka, R.; Nozu, K.; Sako, M.; Iijima, K.; et al. IgA nephropathy with presentation of nephrotic syndrome at onset in children. *Pediatr. Nephrol.* **2017**, *32*, 457–465. [CrossRef] [PubMed]
26. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754. [CrossRef]
27. Hanley, J.A.; Mcneil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]