International Journal of Artificial Intelligence and Machine Learning
Volume 11 • Issue 2 • July-December 2021

Intelligent Prediction Techniques for Chronic Kidney Disease Data Analysis

Shanmugarajeshwari V., Kalasalingam Academy of Research and Education, Krishnankoil, India
Ilayaraja M., Kalasalingam Academy of Research and Education, Krishnankoil, India

ABSTRACT

Information is stored in various domains like finance, banking, hospital, education, etc. Nowadays, data stored in medical databases are growing rapidly. The proposed approach entails three parts comparable to preprocessing, attribute selection, and classification C5.0 algorithms. This work aims to design a machine-based diagnostic approach using various techniques. These algorithms improve the efficiency of mining risk factors of chronic kidney diseases, but there are also have some shortcomings. To overcome these issues and improve an effectual clinical decision support system exhausting classification methods over a large volume of the dataset for making better decisions and predictions, this paper presents grouping classification assembly through consuming the C5.0 algorithm, pointing towards assembling time to acquire great accuracy to identify an early diagnosis of chronic kidney disease patients with risk level by analyzing the chronic kidney disease dataset.

KEYWORDS

Artificial Intelligence, Chronic Kidney, Data Mining, Deep Learning Techniques, Intelligent Decision Support System, Machine Learning Techniques

INTRODUCTION

In data mining is an analyzing or discovering good knowledge to develop the meaningful collection of data from a huge amount of data using the knowledge. The health specifying care is the solicitation of information using machine learning algorithms. To developing also exploring healthcare data records analytical surroundings are using various methods to superior raise the value of health-related problem to prediction.

Health-care record data is mostly gorgeous derived from a worldwide diversity of foundations such as sensor devices, images, text in the system of automated electrical archives. In this miscellaneous in the collection of data and depiction method clues to several trials in together the handling process and analysis of the original data. World wide assortment in the methods is essential to evaluate dissimilar forms of records (Reddy & Aggarwal, 2015).

The kidneys' operations are to pass through a filter of the blood. It eliminates unwanted blood to regulate the stability of electrolytes and fluid. It strains blood, they create urine, which two bean-shaped structure of the kidney. Every one kidney surrounds a million things of unit so-called nephrons (Urinary Incontinence, n.d.).

DOI: 10.4018/IJAIML.20210701.oa2

Factors of Chronic Kidney Disease

The following are some of the factors which lead to chronic kidney disease, the main cause is diabetes and others are hypertension, smoke, fatness, heart illness, family record, alcohol, and age problem.

Symptoms

Some of the warning sign is listed down, that could be variations to urinary function, plasma in the urine, bulge & pain, severe tiredness and weakness.

Types: Acute and Chronic

- Acute Prerenal Kidney Failure -Suddenly decreases blood flow.
- Acute Intrinsic Kidney Failure -Straight injury to the kidneys foundations unexpected damage in kidney.
- Chronic Prerenal Kidney Failure – Gradually decreases blood flow.
- Chronic Intrinsic Kidney Failure - Direct damage to the kidneys cause a gradual loss in kidney function (Bala & Kumar, 2014).

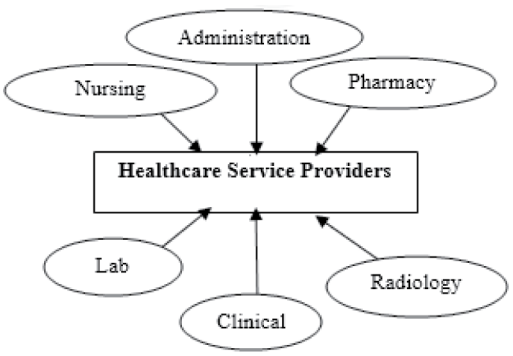
Chronic Kidney Disease (CKD) is a worldwide health crisis. In 2019, the World Health Organization agree to fifty-eight million deaths and 35 million recognized to chronic kidney disease. The world level 850 million people now predicted to have kidney diseases from many causes, chronic kidney disease causes at least 2.4 million deaths world wide-reaching per year sixth fastest-growing cause of disease and death. Dialysis is a fashion of life for many patients pain with kidney sicknesses in India. The medical record of Government of TamilNadu, India, Every one year 2.2 Lakh fresh patients affected by final point renal disease or end-stage renal disease. According to the Global Burden of Disease (GBD) learning, kidney disease was hierarchical 27th 1990 but rose to 18th in 2010 and 9th in 2019. Motivations on the development and use of machine learning algorithms for classical methods using other machine learning approaches to achieve high accuracy.

Figure 1 represents the various factors are affecting the patient data are evaluated with healthcare data analytics.

RELATED WORKS

A Literature survey refers to a critical summary. Literature reviews contextualize research about a topic. A literature review is an evaluative report of studies found in the literature related to a selected area. The review should describe, summarize, evaluate and clarify this literature. It determines the

Figure 1. Affecting factors of the Healthcare Data Analytics



context of the research around a subject area. It is appraising explosion preparations start in the literature associated with a particular area. The journal should designate, review, estimate from the survey (Feature Selection, n.d.). It is a full theory base aimed at the study and benefits the researcher to the environment of the investigation. The analyses have been done on various topics of an outline. The root of the prevailing information, everybody to building advanced knowledge and thought for advance study perseverance (Queens U, n.d.) (Table 1).

FINDINGS

From this review, it is concrete that healthcare decision support clinical performance can be assessed by smearing, machine learning techniques can be valued by various algorithms. In this survey, our research work ordered as three parts. The best algorithm is deep learning to deal with huge datasets, using an R programming language is used. This research work presents an algorithm on the classification structure by various artificial intelligence and machine learning algorithms that have resulted in good accuracy. In the future, the proposed research work has been successfully implemented in R with the Graphical User Interface (GUI) environment.

Overview of the Model

The first objective is an early diagnosis of Chronic Kidney Disease (CKD) patients with risk levels by analyzing chronic kidney disease dataset. This objective plays a valuable role in current research since many patients suffer from this disease around the world (Figures 2-3).

Phase 1: Preprocess
- Dataset Depiction
- Cataloging

The second objective is the power of the feature selection using machine learning methods to detect the patients with the risk level of chronic kidney diseases while affected by particular symptoms of a particular disease (Figures 4-9).

Phase 2: Feature Extraction
- oneR
- Random Forest
- Relief
- Symmetrical Uncertainty
- Chi_Square
- Information_Gain
- Gain_Ratio

Attribute Selection

It is likewise a feature selection (HKU, n.d.). Now, choosing to apply features and neglect the inappropriate attributes. These methods were applied to the preprocessed data set which has 4050 samples, concentrated on picking out all attributes. The greatest attribute variety technique is to gain ratio feature selection that has been functional to the preprocessed records (Celik et al., 2014).

Gain Ratio Feature Selection

It is attributed variety technique to gain ratio with the attribute method. The feature selection method is used to extract the relevant features and discard the irrelevant features. This method applied for chronic kidney disease datasets (Celik et al., 2014):

Table 1. Literature Review for Analysis of Chronic Kidney Disease

Author	Problem	Software	Techniques	Accuracy Performance (%)
Dowluru, et al. (2012)	Analysis of Kidney stone	WEKA Tool	Naive Bayes Classification	0.99
			Logistic Regression	1.00
			J48 Algorithm	0.97
			Random Forest	0.98
		ORANGE Tool	Naive Bayes	0.79
			K-NN	0.7377
			Classification tree	0.9352
			C4.5	0.9352
			SVM	0.9198
			Random Forest	0.9352
Lakshmi, et al. (2014)	Problem with Kidney dialysis	TANAGRA	ANN	93.852
			Decision Tree(C5)	78.4455
			Logical Regression	74.7438
Eyck, et al. (2012)	A.K.I	matlab	Gaussian-aROC	0.758
			Gussian- RMSER	0.408
Zadeh, et al. (2013)	Early AVF Failure	WEKA Tool	W-Simple Cart	85.11
			WJ48	80.85
Abeer Y. Al- Hyari, et al. (2012)	Enduring Kidney disease	WEKA Tool	Decision tree	--
Song et al. (2012)	Renal failure Hemodialysis	WEKA Tool	Decision tree	60-80
Sriraam, et al. (2006)	Treatment of Kidney Dialysis	WEKA Tool	Association Rule	97.7
Jain, et al. (2014)	Nephrotic syndrome(total protein)	TANAGRA Tool	C4.5	11 (error rate)
Jose, et al. (2012)	Analysis of Kidney Image	MATLAB	Association Rule	92
			Navie Bayes	
Kumar, et al. (2012)	Kidney Stone treatment and symptoms	WEKA	ANN – MLP	0.9613
			ANN – LVQ	0.8459
			ANN – RBF	0.8732
Leung, et al. (2013)	Danger forms in diabetic kidney disease	MATLAB	SVM	0.91
			PLS	0.83
			FFNN	0.85
			RPART	0.87
			Random Forest	0.91
			Naïve Bayes	0.86
			C5.0	0.90
Bala, et al. (2014)	Review of Kidney Disease Prediction	-	-	-
Vijayarani, et al. (2015a)	Kidney Disease Prediction	MATLAB	SVM	76.32
			ANN	87.70

continued on following page

Table 1. Continued

Author	Problem	Software	Techniques	Accuracy Performance (%)
Vijayarani, et al. (2015b)	Arrangement Processes for Kidney Ailment Forecast	MATLAB	SVM	76.32
			Naïve Bayes	70.96
Sinha, et al. (2015)	Learning of Lasting Kidney Virus Expectation	MATLAB	SVM	0.7375
			KNN	0.7875
Jena, et al. (2015)	Guess of Chronic-Kidney-Disease	WEKA	Naïve Bayes	95
			Multilayer Perceptron	99.75
			SVM	62
			J48	99
			Conjuctive Rule	94.75
			Decision Table	99
Ramya, et al. (2016)	Judgment of Chronic Kidney Disease	R	BP	80.4
			RBF	85.3
			Random Forest (RF)	78.6
Celik, et al. (2014)	Diagnosis besides Estimate of Fleeting Kidney Disease	WEKA	SVM	97.06
			Decision Tree	96.12
Norouzi, et al. (2016)	Renal Miscarriage Problem in CKD	MATLAB	ANFIS	95
Sharma, et al. (2016)	Chronic Kidney Disease Diagnosis	MATLAB	Decision Tree	98.60
			SVM	90.50
			ANN - MLFFNN	88.50
			KNN	88.88
			Discriminant Analysis	90.80
Kumar, M., (2016)	Prediction of Chronic Kidney Disease	MATLAB	RF	95.67
			SMO	90
			Naïve Bayes	87.64
			RBF	83.78
			MLPC	89
			SLG	87
Chatterjee, et al. (2017)	Chronic Kidney Disease Classification	MATLAB	MLP-FFN	96.33
			PSO-NN	98.5
			NN-MCS	99.6
Subhashini, et al. (2017)	Performance analysis of Chronic Kidney Disease	MATLAB	ANN	93
			KNN	96.76
			S_V_M	87
			Naïve _Bayes	88.9
			Decision _Tree	86
			Fuzzy_Ambiguous_ Classifier	90

continued on following page

Table 1. Continued

Author	Problem	Software	Techniques	Accuracy Performance (%)
Alasker, et al. (2017)	Detection of Kidney Disease	WEKA	ANN	99.5 100
			Naïve Bayes	99.5 100
			Decision Table	97.619
			J48	98.4127
			One R	99.2063
			KNN	97.619
Mahdavi-mazdeh, et al. (2018)	Predict chronic kidney disease progression	MATLAB	ANFIS	98
Lakshmanaprabu, et al. (2019)	Medical decision support system	MATLAB	DNN	98.25
			PSO	99.25
Pasadana, et al. (2019)	Chronic Kidney Disease Prediction	MATLAB	Decision _Stump	92
			Hoeffding _Tree	95.75
			J_48	99
			CTC	97
			J48graft	98.75
			LMT	98
			NB _Tree	98.5
			Random _Forest	100
			Random _Tree	95.5
			REP Tree	96.75
			Simple Cart	97.5
Shetty, et al. (2019)	CKD Prediction	Pycharm	SVM	90.09
Ahmad, M., et al. (2017)	Chronic condition of kidney disease	R	SVM	98.34
Alassaf, R.A., et al. (2018)	Preemptive Diagnosis of Chronic Kidney Disease	Weka and Python	ANN	98
			SVM	98
			Naïve Bayes	98
			K-NN	93.9
Aljaaf, A.J., et al. (2018)	Initial Estimate of CKD	MATLAB	RPART	95.6
			SVM	95
			LOGR	98.1
			MLP	98.1
Arif-Ul-Islam and Ripon, S.H., Rule (2019)	Regulation Orientation of CKD	Weka	AdaBoost	99
			LogitBoost	99.75
Avci, E., et al. (2018)	Performance comparison of CKD	WEKA	NB	95
			k-star	97.75
			SVM	91.75
			J48	99

continued on following page

Table 1. Continued

Author	Problem	Software	Techniques	Accuracy Performance (%)
Banerjee, A., et al. (2019)	Food Recommendation of CKD	MATLAB	Random Forest	99.75
			SVM	98.25
			Naïve Bayes	95.5
Basarslan, M.S., and Kayaalp, F., (2019)	Detection of Chronic Kidney Disease	MATLAB	K-Nearest _Neighbor	97
			Navie _Bayes	96.5
			LR	97.56
			RF	99
Bhaskar, N., and Suchetha, M., (2019)	Automated Sensing of Chronic Kidney Disease	MATLAB	CNN-SVM	98.04
Lakshmanaprabu, S.K., et al.	Clinical decision support system	MATLAB	DNN	98.25
			PSO-DNN	99.25
Shankar, K., et al. (2018)	Finest attribute Selection	MATLAB	D_N_N	98
			C_N_N	90
			N_N	92
			B_P	80
			K_N_N	79
Zhang, H., et al. (2018)	Survival Prediction	Python	Classical MLPs	97.69
			LASSO preset MLPs	93.23
Dulhare, U.N., and Mohammad Ayesha, M., (2016)	Mining of CKD	Weka	Naïve Bayes	85
			Naïve Bayes with OneR	97.5
Devika R, et al. (2019)	Classify to CKD	MATLAB	Naive _Bayes	99.635
			K_N_N	87.78
			Random_Forest	99.844
Jain, D., and Singh, V., (2018)	Various level of CKD	MATLAB	SVM	99
			ANN	95
Lee, M.-C., Wu, S.-F. V., Hsieh, N.-C., & Tsai, J.-M. (2016)	Kidney_Self-esteemed progress	Meta- data-Analysis	-	-
Almansour, N.A., et al. (2019)	NN and SVM prediction CKD: Review study	WEKA	ANN	99.75
			SVM	97.75
Zhao, J., et al. (2019)	Expecting consequences of CKD using EMR	MATLAB	Random Forest	98.75
			Sequential Minimal Optimization,	97.75
			Naïve Bayes	98.25
			Radial Basis Function	95
			Multilayer Perceptron Classifier	90
			Simple Logistic	92

Figure 2. CKD Dataset Loading R

	ageg	sex	sbp	dbp	htn	smoking	alc	sug	rbc	pcell	pccellc	bac	bgr	blu	sercr	sdi	pota	hg	pcv	wbcc	rbcc
1	3	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	121	36.0	1.20	145	4.2	15.4	44	7800	5.2
2	2	2	110	70	1	0	0	0	normal	abnormal	notpresent	notpresent	128	36.0	1.20	141	4.4	15.4	44	7800	5.2
3	2	2	150	80	2	0	0	0	normal	normal	notpresent	notpresent	100	22.0	0.70	136	4.8	10.7	34	12300	5.2
4	3	2	130	80	1	0	0	0	normal	normal	notpresent	notpresent	99	23.0	0.60	138	4.4	12	34	12300	5.2
5	4	2	130	90	2	0	0	0	normal	normal	notpresent	notpresent	65	16.0	0.70	138	3.2	8.1	34	12300	5.2
6	4	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	83	25.0	0.60	138	3.2	11.8	36	12400	5.2
7	4	2	120	70	1	0	0	0	abnormal	abnormal	notpresent	present	94	67.0	1.00	135	4.9	9.9	30	16700	4.8
8	3	2	140	70	2	0	0	0	abnormal	normal	notpresent	notpresent	89	18.0	0.80	135	4.9	11.3	38	6000	4.8
9	4	2	120	70	1	0	0	0	normal	normal	notpresent	notpresent	78	27.0	0.90	135	4.9	12.3	41	6700	4.8
10	3	2	145	70	2	0	0	0	normal	normal	notpresent	notpresent	72	46.0	1.00	135	3.8	12.3	41	6700	4.8
11	2	2	130	80	1	0	0	0	normal	normal	notpresent	notpresent	80	66.0	2.50	142	3.6	12.2	38	6700	4.8
12	3	1	130	90	2	0	0	0	normal	normal	notpresent	notpresent	86	17.0	0.80	142	3.6	15	45	8600	4.8
13	2	1	100	90	2	1	1	0	abnormal	abnormal	present	notpresent	65	51.0	1.80	142	3.6	12.1	45	10300	4.8
14	2	1	140	70	2	0	0	0	normal	normal	notpresent	notpresent	100	26.0	0.60	137	4.4	15.8	49	6600	4.8
15	3	2	130	80	1	0	0	0	normal	normal	notpresent	notpresent	192	15.0	0.80	137	4.2	14.3	40	9500	5.4
16	4	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	86	15.0	0.60	138	4.0	11	33	7700	3.8
17	4	2	130	90	2	0	0	0	normal	normal	notpresent	notpresent	93	17.0	0.90	136	3.9	16.7	50	6200	5.2
18	3	2	130	90	2	0	0	0	normal	normal	notpresent	notpresent	92	32.0	2.10	141	4.2	13.9	52	7000	5.2
19	3	2	150	70	2	0	0	0	abnormal	normal	notpresent	notpresent	22	1.5	7.30	145	2.8	13.1	41	11200	5.2
20	3	2	110	90	2	0	0	0	normal	normal	notpresent	notpresent	114	50.0	1.00	135	4.9	14.2	51	7200	5.9
21	3	1	130	80	1	0	0	0	normal	normal	notpresent	notpresent	114	50.0	1.00	135	4.9	11.5	51	6900	5.9
22	4	2	130	70	1	0	0	0	normal	normal	notpresent	notpresent	107	23.0	0.70	141	4.2	14.4	44	6900	5.9
23	4	2	140	70	2	0	0	0	normal	normal	notpresent	notpresent	107	23.0	0.70	137	4.7	14	41	4500	5.5
24	1	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	123	44.0	1.00	135	3.8	14.6	44	5500	4.8
25	4	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	123	44.0	1.00	135	3.8	14.6	44	5500	4.8
26	2	2	150	70	2	0	0	0	normal	abnormal	present	present	107	40.0	1.70	125	3.5	8.3	23	12400	3.9
27	3	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	97	18.0	1.20	138	4.3	13.5	42	7900	6.4
28	4	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	70	36.0	1.00	150	4.6	17	52	9800	5.0
29	2	2	150	70	2	0	0	0	normal	normal	notpresent	notpresent	111	34.0	1.10	145	4.0	14.3	41	7200	5.0
30	1	2	125	82	1	0	0	0	normal	normal	notpresent	notpresent	99	46.0	1.20	142	4.0	17.7	46	4300	5.5

Figure 3. Finally, feature selection attributes only selected to R

RGui (64-bit) - [Data: data]
File

	age	sex	sysbp	diabp	hyptn	smoking	alc	ldl	ldlg	ldlgl	egfr	ckd	Class
1	60	Female	150	70	2	0	0	98	100-129	Optimal	68.8	0	Mild
2	58	Female	110	70	1	0	0	143	130-189	high	98.7	0	Normal
3	59	Female	150	80	2	0	0	160	130-189	high	64.1	0	Mild
4	64	Female	130	80	1	0	0	111	100-129	Optimal	47.9	1	Moderate
5	80	Female	130	90	2	0	0	80	100-129	Optimal	25.9	1	Severe
6	78	Female	150	70	2	0	0	70	100-129	Optimal	16.2	1	Severe
7	70	Female	120	70	1	0	0	95	100-129	Optimal	83.7	0	Mild
8	63	Female	140	70	2	0	0	91	100-129	Optimal	41.4	1	Moderate
9	87	Female	120	70	1	0	0	61	100-129	Optimal	29.7	1	Severe
10	67	Female	145	70	2	0	0	102	100-129	Optimal	74.3	0	Mild
11	58	Female	130	80	1	0	0	128	100-129	Optimal	99.7	0	Normal
12	63	Male	130	90	2	0	0	128	100-129	Optimal	87.1	0	Mild
13	51	Male	100	90	2	0	0	70	100-129	Optimal	103.4	0	Normal
14	56	Male	140	70	2	0	0	86	100-129	Optimal	27.1	1	Severe
15	63	Female	130	80	1	0	0	86	100-129	Optimal	76.4	0	Mild
16	72	Female	150	70	2	0	0	131	130-189	high	87.2	0	Mild
17	75	Female	130	90	2	1	1	85	100-129	Optimal	36.5	1	Moderate
18	67	Female	130	90	2	0	0	123	100-129	Optimal	91.7	0	Normal
19	65	Female	150	70	2	0	0	100	100-129	Optimal	19.7	1	Severe
20	63	Female	110	90	2	0	0	80	100-129	Optimal	93.8	0	Normal
21	63	Male	130	80	1	0	0	93	100-129	Optimal	75.2	0	Mild
22	86	Female	130	70	1	0	0	103	100-129	Optimal	35.3	1	Moderate
23	74	Female	140	70	2	0	0	104	100-129	Optimal	25.0	1	Severe
24	47	Female	150	70	2	0	0	94	100-129	Optimal	96.8	0	Normal
25	76	Female	150	70	2	0	0	104	100-129	Optimal	19.6	1	Severe
26	58	Female	150	70	2	0	0	105	100-129	Optimal	32.8	1	Severe
27	69	Female	150	70	2	0	0	193	130-189	high	71.1	0	Mild
28	78	Female	150	70	2	0	0	85	100-129	Optimal	34.2	1	Severe
29	56	Female	150	70	2	0	0	84	100-129	Optimal	99.6	0	Normal
30	43	Female	125	82	1	0	0	76	100-129	Optimal	99.5	0	Normal
31	56	Female	130	80	1	0	0	100	100-129	Optimal	104.6	0	Normal

$$\text{Entropy } (D_j) = -\sum_{j=1}^{m} pj \ log2\left(pj\right) \tag{1}$$

$$\text{Info Gain } (D, A) = \text{Entropy } (D_j) - \sum_{j=1}^{v} \frac{Dj}{D} * Entropy\left(Dj\right) \tag{2}$$

$$\text{Gain Ratio } (A) = \text{Entropy } (D) - \text{Information Gain }_A (D) \tag{3}$$

Figure 4. OneR

	attr_importance
ID	0.000000e+00
age	2.148148e-02
ageg	1.481481e-02
sex	5.551115e-17
occu3	5.551115e-17
edu2	5.551115e-17
sbp	0.000000e+00
dbp	2.469136e-03
htn	5.551115e-17
smoking	5.551115e-17
alc	5.551115e-17
dx	5.551115e-17
ldl	5.551115e-17
ldlg	5.551115e-17
uhc	5.551115e-17
egfr	2.535802e-01

Figure 5. Random

	attr_importance
ID	5.4184276
age	12.0455084
ageg	8.5505324
sex	0.3803352
occu3	2.3442413
edu2	-1.9016220
sbp	8.6162261
dbp	3.5450150
htn	7.3814191
smoking	-0.4258979
alc	-0.3421356
dx	6.7957406
ldl	4.9443505
ldlg	4.6541099
uhc	-0.7720223
egfr	422.1423430

Figure 6. Relief

	attr_importance
ID	2.600980e-02
age	2.897436e-02
ageg	7.666667e-02
sex	0.000000e+00
occu3	2.500000e-02
edu2	3.008081e-02
sbp	7.105263e-03
dbp	5.000000e-02
htn	0.000000e+00
smoking	0.000000e+00
alc	0.000000e+00
dx	0.000000e+00
ldl	6.321321e-03
ldlg	3.000000e-02
uhc	-5.551115e-18
egfr	2.209169e-01

Figure 7. Symmetrical Uncertainty

	attr_importance
ID	0.017636518
age	0.092612490
ageg	0.087739999
sex	0.000000000
occu3	0.029974558
edu2	0.000000000
sbp	0.024272977
dbp	0.008686779
htn	0.024815333
smoking	0.000000000
alc	0.000000000
dx	0.024815333
ldl	0.000000000
ldlg	0.000000000
uhc	0.000000000
egfr	1.000000000

Figure 8. Gain Ratio

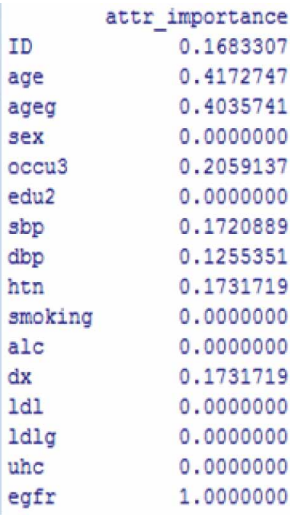
	attr_importance
ID	0.014043224
age	0.065389964
ageg	0.062190510
sex	0.000000000
occu3	0.024307074
edu2	0.000000000
sbp	0.022136135
dbp	0.006094753
htn	0.022700592
smoking	0.000000000
alc	0.000000000
dx	0.022700592
ldl	0.000000000
ldlg	0.000000000
uhc	0.000000000
egfr	1.000000000

$$\text{Information Gain}_{\text{A}}(\text{D}) = \sum_{j=1}^{v} \frac{Dj}{D} * Entropy\left(Dj\right) \tag{4}$$

This research work focuses on the gain ratio feature selection that comes under the filter method, which uses the measures entropy, information gain and gains ratio. Other feature selection methods are:

- Chi-square
- Random forest
- Relief
- OneR
- Symmetrical uncertainty

Figure 9. Chi-Square



	attr_importance
ID	0.1683307
age	0.4172747
ageg	0.4035741
sex	0.0000000
occu3	0.2059137
edu2	0.0000000
sbp	0.1720889
dbp	0.1255351
htn	0.1731719
smoking	0.0000000
alc	0.0000000
dx	0.1731719
ldl	0.0000000
ldlg	0.0000000
uhc	0.0000000
egfr	1.0000000

The third objective is the optimization of classical machine learning algorithms using another machine learning approaches to achieve high accuracy.

Part 3: Classification

- Decision Tree
- C4.5 Algorithm
- C5.0 Algorithm

Figure 10 represents the various parts are represented the patient data is evaluated with the healthcare data analytics block diagram of Chronic Kidney Disease (CKD).

DATA COLLECTION

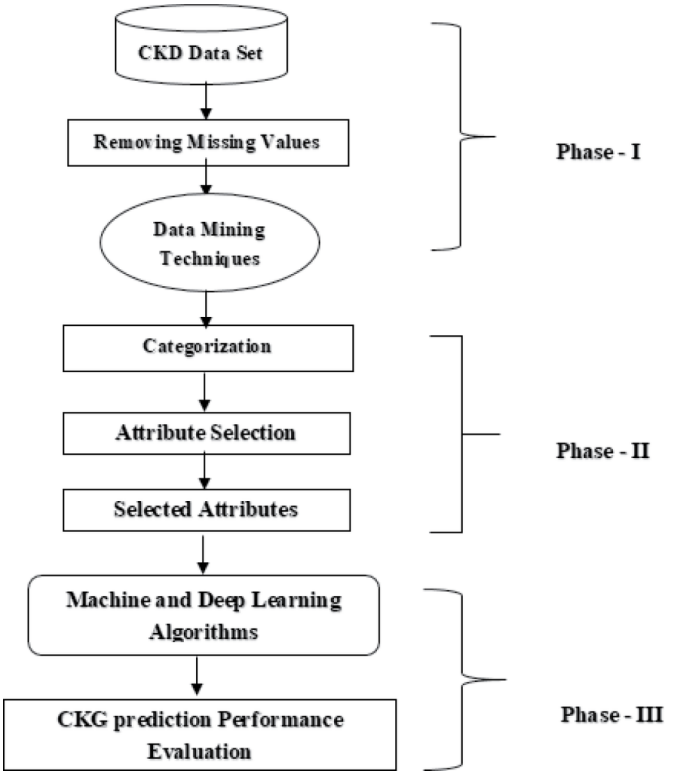
The chronic kidney data set files are composed of prediction is based on the given attributes. This dataset has thirty-two attributes that predict the CKD. It contains an attribute such as age, age group sex, (systolic and diastolic) blood_pressure, specific_gravity, albumin, sugar, red_blood_cells, plus_cell, pus_cell_clumps, bacteria, blood_glucose_random, blood_urea, serum_creatine, sodium, potassium, hemoglobin, packed_cell_volume, white_blood_cell_count, red_blood_cell count, hypertension, diabetes_mellitus, appetite, pedal_edema, Low_density lipoprotein, smoking_status, alcohol_drinking, anemia, Estimated_glomerular_filtration_rate, CKD Level and Class. Initially, data size are 4050 records and 33 attributes are preprocessing, attribute variety techniques, cataloging or classification algorithms toward spread over chronic kidney data using performance evaluation (Tables 2-3).

RESULTS

C5.0 Algorithm

It is an important classification algorithm for decision tree. These algorithms are handling continued values or categorical values. Feature selection is an essential phase to create the

Figure 10. Methodology Block Diagram of CKD



decision_tree. Associating to a Decision tree, C4.5 and C5.0 uppermost process. It was performed on pre_pruning. One of the decision tree classification algorithms. Entropy, information_gain and gain_ratio measures are considered and the classification model was developed. Applied on the data set to determine the unknown samples. Many classification methods are available to predict overall performance. That C5.0 is one of the best decision tree classification algorithms. It can handle continuous and categorical values. It can handle numeric attributes. Comparing with ID3, C4.5 and C5.0. C5.0 has the highest speed and pre-pruning. So, the proposed system carries out the prediction operation using the C5.0 classifier. Attribute selection is the fundamental step to construct a decision tree. Entropy, information gain, and gain ratio are used to process attribute selection. During attribute selection, C5.0 algorithm selects the root node of the decision tree (Figures 11-12).

Advantages:

- Accurate result
- Less memory space for the large data set
- Less time to build a model
- Increasing level support
- Highest speed
- Handle continuous value, categorical values and multi-value

Disadvantage:

- Empty branches and insignificant branches are allowed

Table 2. Attributes of Chronic Kidney Disease Dataset

S. No	Attribute Name	Attribute Type	Attribute Code	Possible Values
1.	Age	Numeric	age	E, VG, G, F, P
2.	Age Group	Numeric	ageg	E, VG, G, F, P
3.	Sex	Nominal	Sex	E, VG, G, F, P
4.	Systolic Blood Pressure	Numeric	sysbp	E, VG, G, F, P
5.	Diastolic Blood Pressure	Numeric	diabp	E, VG, G, F, P
6.	Specific Gravity	Numeric	sap	E, VG, G, F, P
7.	Albumin	Numeric	alb	E, VG, G, F, P
8.	Sugar	Numeric	sug	E, VG, G, F, P
9.	Red Blood Cell	Nominal	rbc	E, VG, G, F, P
10.	Pus Cell	Nominal	pcell	E, VG, G, F, P
11.	Pus Cell Clumps	Nominal	pcellc	E, VG, G, F, P
12.	Bacteria	Numeric	bac	E, VG, G, F, P
13.	Blood Glucose Random	Numeric	bgr	E, VG, G, F, P
14.	Blood Urea	Numeric	blu	E, VG, G, F, P
15.	Serum Creatine	Numeric	sercr	E, VG, G, F, P
16.	Sodium	Numeric	sdi	E, VG, G, F, P
17.	Potassium	Numeric	pota	E, VG, G, F, P
18.	Hemoglobin	Numeric	hg	E, VG, G, F, P
19.	Packed_Cell_Volume	Numeric	p_c_v	E, VG, G, F, P
20.	White_Blood_Cell_Count	Numeric	w_b_c_c	E, VG, G, F, P
21.	Red_Blood_Cell_Count	Numeric	r_b_c_c	E, VG, G, F, P
22.	Hypertension	Nominal	hyptn	E, VG, G, F, P
23.	Diabetes Mellitus	Numeric	diam	E, VG, G, F, P
24.	Appetite	Nominal	app	E, VG, G, F, P
25.	Pedal Edema	Nominal	peed	E, VG, G, F, P
26.	Low Density Lipoprotein	Numeric	ldl	E, VG, G, F, P
27.	smoking status	Numeric	smo	E, VG, G, F, P
28.	Alcohol Drinking	Numeric	alc	E, VG, G, F, P
29.	Anemia	Nominal	ane	E, VG, G, F, P
30.	Coronary Artery Disease	Nominal	Coad	E, VG, G, F, P
31.	Estimated Glomerular Filtration Rate	Numeric	egfr	E, VG, G, F, P
32.	CKD Level	Numeric or Nominal	ckd	E, VG, G, F, P
33.	Class	Numeric or Nominal	Class	E, VG, G, F, P

Table 3. Testing Performance for Chronic Kidney Disease Identification

Main Testing	Prediction	
All attributes measure level compare to Estimated Glomerular Filtration Rate value(egfr)	Excellent	Normal
	Very Good	Mild
	Good	Moderate
	Fair	Severe
	Poor or Failure	End-stage

Figure 11. C5.0 Algorithm

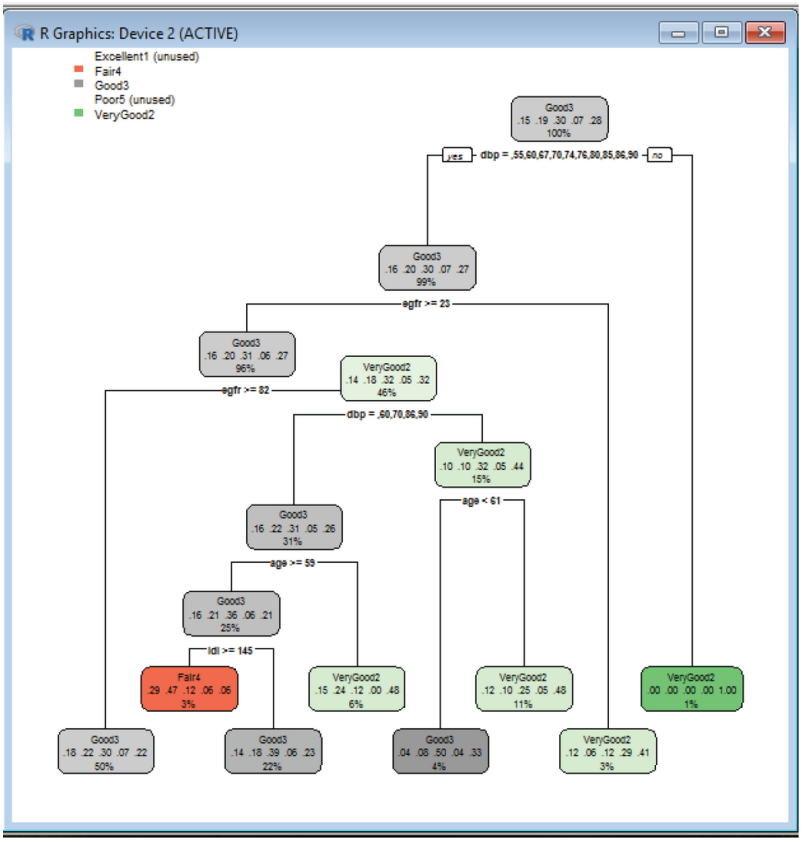


Figure 12. Pseudocode for C5.0 Classifier

Input: Chronic Kidney Data
Output: Selected Attributes
Step1: Read the data
Step2: Calculate entropy value
Step3: Compute information gain for each and every attribute
Step4: Compute gain ratio using entropy and information gain
Step5: Find the attribute with the highest gain ratio value
Step6: If there is no more attribute, the tree construction was completed. (Returns a leaf labeled with the most frequent class or the disjunction of all the classes).

C5.0_Classification

Numerous classification algorithms available to predict early stages are identified for CKD. It is the topmost algorithms. So, the proposed system carries out the prediction operation using the C5.0 classifier. In the proposed system, the C5.0 classifier classifies chronic kidney disease stages into predetermined classes such as normal (Excellent), mild (very good), moderate (Good), severe (Fair) and end-stage (Poor or Failure).

C5.0 algorithm applies the chronic kidney disease data to predicting all stages and identifies early stages for analyzing less time consuming and higher accuracy comparing other machine learning algorithms.

DECISION TREE

It signifies a test node. It is used to organize an order by beginning at the root other than a leaf node (Quinlan, 1986).

A decision tree is a supervised classification, which predicts both the classifier and regression models. Classification trees are mainly used to classify an object to a predetermined class based on the attributes. The tree contains no incoming edge is called as root, The node through one outward edge is named an internal_node, all other nodes stay notorious as leaf node which has no outgoing edge. Using the training sets the classifier model has been developed, the testing set was applied to the classification model to predict the previously unknown class (Figure 13).

C4.5 Algorithm

C4.5 is the basic classification algorithm for decision tree. It was developed by Quinlan. To uses a gain _ratio by way of a split the selection process. By calculating entropy and splitting information of an attribute. It is based on attributes selection for numeric and missing data values. Faster than the ID3 algorithm. It also cannot deal with missing values. A decision tree is built to scrutinizing a regular training examples class brands are identified. This selection is an identified model smeared to decide the property of unidentified models (Figures 14-15).

Advantages:

- accurate result
- less memory space for the large data set
- less time to build a model
- short searching time

Disadvantages:

- Empty branches and insignificant branches are allowed
- Overfitting is one of the most important problems in the C4.5 algorithm

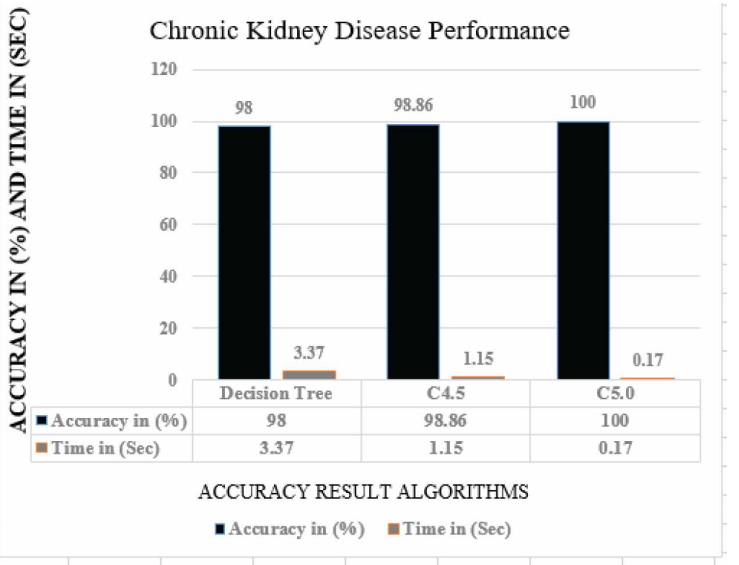
Figure 13. Pseudocode for Decision Tree

Input: Chronic kidney data with selected features
Output: Classified data for Decision Tree
Step1: Apply S to D to find a splitting criterion
Step2: If (t is not a leaf node)
Step3: Create children nodes of t
Step4: Partition D into children partitions
Step5: Repeat on each partition
Step6: End

Figure 14. Pseudocode for C4.5 Algorithm

Input : An Attribute – valued dataset *D*
Output: Classified CKD data for C4.5
Step1: Tree={}// Condition
Step 2: if *D* is "pure" or stopping criteria met then
Step 3: Terminate node level
Step 4: else if
Step 5: for all attribute *a*∈*D* do
Step 6: Compute information-theoretic criteria if we split on *a*
Step 7: End for
Step 8: a*best* = Best attribute according to above-computed criteria
Step 9: Tree = Create a decision node that tests a*best* in the root
Step 10: *Dv* = induced sub-datasetss from *D* based on a*best*
Step 11: For all *Dv* do
Step 12: Treev= C4.5(*Dv*)
Step 13: Attach Treev, to the corresponding branch of the tree
Step 14: End for
Step 15: Return Tree

Figure 15. Accuracy performance using various machine learning algorithms



CONCLUSION

In conclusion, chronic kidney disease considers approaching toward developing recommendations for machine learning techniques in healthcare has become a real-world emerging for obtaining accurate results of medical diagnosis, using the machine learning techniques involved the healthcare is evolving into a promising field for improving outcomes with reducing costs. Thus the system can improve the efficiency of mining risk factors of chronic kidney disease, but there are also have some shortcomings. To overcome these issues, improve an effectual clinical judgment care structure chronic kidney disease decision support system consuming classification algorithms over a large volume of the dataset for making better decisions and predictions. The gain ratio feature selection method is best and fewer time associates other selection methods. The information is verified by classification C5.0 algorithms. Then to predict chronic kidney disease using the C5.0 is high accuracy bring about and less time complexity in 100% cataloging accuracy.

REFERENCES

Abeer, Y., & Al-Hyari. (2012). *Chronic kidney disease prediction system using classifying data mining techniques.* Library of University of Jordan.

Ahmad, M., Tundjungsari, V., Widianti, D., Amalia, P., & Rachmawati, U. A. (2017). Diagnostic Decision Support System of Chronic Kidney Disease Using Support Vector Machine. *Second International Conference on Informatics and Computing (ICIC).* doi:10.1109/IAC.2017.8280576

Alasker, H., Alharkan, S., Alharkan, W., & Riza, L. S. (2017). Detection of Kidney Disease Using Various Intelligent Classifiers. *2017 3rd International Conference on Science in Information Technology (ICSITech),* 681-684. doi:10.1109/ICSITech.2017.8257199

Alassaf, R. A., Alsulaim, K. A., Alroomi, N. Y., Alsharif, N. S., Aljubeir, M. F., Olatunji, S. O., Alahmadi, A. Y., Imran, M., Alzahrani, R. A., & Alturayeif, N. S. (2018). Preemptive Diagnosis of Chronic Kidney Disease Using Machine Learning Techniques. *13 International Conferences on Innovations in Information Technology, 2018 IEEE*, 99-104.

Aljaaf, A. J., Al-Jumeily, D., & Hussein, M. (2018). Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics. *IEEE Congress on Evolutionary Computation (CEC).*

Almansour, N. A., Syed, H. S., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J., Alrashed, S., & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Computers in Biology and Medicine*, *109*, 101–111. doi:10.1016/j.compbiomed.2019.04.017 PMID:31054385

Arif-Ul-Islam, , & Ripon, , S.H. (2019). Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE).*

AVCI. (2018). *Performance Comparison of Some Classifiers on Chronic Kidney Disease Data*. IEEE.

Bala, S., & Kumar, K. (2014, July). A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique. *IJCSMC*, *3*(7), 960–967.

Banerjee, A., Noor, A., Siddiqua, N., & Uddin, M. N. (2019). Significance of Attribute Selection In The Classification of Chronic Renal Disease. *Second International Conference on Advanced Computational and Communication Paradigms (ICACCP).* doi:10.1109/ICACCP.2019.8882937

Basarslan, M. S., & Kayaalp, F. (2019). *Performance Analysis of Fuzzy Rough Set-Based and Correlation-Based Attribute Selection Methods on Detection of Chronic Kidney Disease with Various Classifiers*. IEEE. doi:10.1109/EBBT.2019.8741688

Bhaskar, N., & Suchetha, M. (2019). A Deep Learning-based System for Automated Sensing of Chronic Kidney Disease. *IEEE Sensors Letters, 3*(10).

Celik, E., Atalay, M., & Kondiloglu, A. (2014). The Diagnosis and Estimate of Chronic Kidney Disease Using the Machine Learning Methods. *International Journal of Intelligent Systems and Applications in Engineering, 4*, 27–31.

Chatterjee, S., Dzitac, S., Sen, S., Rohatinovici, N. C., Dey, N., Ashour, A. S., & Balas, V. E. (2017). Hybrid Modified Cuckoo Search-Neural Network in Chronic Kidney Disease Classification. *14th International Conference on Engineering of Modern Electric Systems (EMES).* doi:10.1109/EMES.2017.7980405

Devika, R., Avilala, S. V., & Subramaniyaswamy, V. (2019). Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest. *Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019).* doi:10.1109/ICCMC.2019.8819654

Dowluru, K., Rayavarapu, A.K., & Vadlapudi, V. (2012). Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis. *Open Access Scientific Reports, 1*(12).

Dulhare, U. N., & Mohammad Ayesha, M. (2016). Extraction of Action Rules for Chronic Kidney Disease using Naïve Bayes Classifier. *IEEE International Conference on Computational Intelligence and Computing Research.* doi:10.1109/ICCIC.2016.7919649

Eyck, V. J., Ramon, J., Guiza, F., Meyfroidt, G., Bruynooghe, M., & Berghe, V. G. (2012). *Data mining techniques for predicting acute kidney injury after elective cardiac surgery*. Springer.

Feature Selection. (n.d.). In *Wikipedia*. https://en.wikipedia.org/wiki/Feature_selection

HKU. (n.d.). https://ar.cetl.hku.hk/am_literature_reviews.htm

Jain, D., & Gautam, S. (2014). Predicting the Effect of Diabetes on Kidney using Classification in Tanagra. *International Journal of Computer Science and Mobile Computing, 3*(4).

Jain, D., & Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal, 19*, 179–189.

Jena, L., & Kamila, N. K. (2015). Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease. *International Journal of Emerging Research in Management &Technology, 4*(11).

Jose, J.S., Sivakami, R., UmaMaheswari, N., & Venkatesh, R. (2012). An Efficient Diagnosis of Kidney Images using Association Rules. *International Journal of Computer Technology and Electronics Engineering, 2*(2).

Kumar, K., & Abhishek, . (2012). Artificial Neural Networks for Diagnosis of Kidney Stones Disease. I.J. *Information Technology and Computer Science*, *7*, 20-25.

Kumar, M. (2016). Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. *International Journal of Computer Science and Mobile Computing, 5*(2), 24-33.

Lakshmanaprabu, S. K., Mohanty, S. N., Rani, S. S., Krishnamoorthy, S., Uthayakumar, J., & Shankar, K. (2019). Online clinical decision support system using optimal deep neural Networks. *Applied Soft Computing*, *81*, 105487. doi:10.1016/j.asoc.2019.105487

Lakshmi, K.R., Nagesh, Y., & VeeraKrishna, M. (2014). Performance comparison of three data mining techniques for predicting kidney disease survivability. *International Journal of Advances in Engineering & Technology*.

Lee, M.-C., Wu, S.-F. V., Hsieh, N.-C., & Tsai, J.-M. (2016). Self-Management Programs on eGFR, Depression, and Quality of Life among Patients with Chronic Kidney Disease: A Meta-Analysis. *Asian Nursing Research*, *10*(4), 255–262. doi:10.1016/j.anr.2016.04.002 PMID:28057311

Leung, R. K. K., Wang, Y., Ma, R. C. W., Luk, A. O. Y., Lam, V., Ng, M., So, W. Y., Tsui, S. K. W., & Chan, J. C. N. (2013). 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 Nephrology*, *14*(1), 162. doi:10.1186/1471-2369-14-162 PMID:23879411

Mahdavi-mazdeh, M., Yadollahpour, A., Nourozi, J., Mirbagheri, S. A., Macotela, F. R. T., & Simancas-Acevedo, E. (2018). Designing and implementing an ANFIS based medical decision support system to predict chronic kidney disease progression. *Frontiers in Physiology*. www.frontiersin.org

Norouzi, J., Yadollahpour, A., AhmadMirbagheri, S., Mazdeh, M., & Hosseini, S.A. (2016). Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System. *Computational and Mathematical Methods in Medicine*. 10.1155/2016/6080814

Pasadana, I. A., Hartama, D., Zarlis, M., & Sianipar, A. S. (2019). Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques. *The International Conference on Computer Science and Applied Mathematics*. doi:10.1088/1742-6596/1255/1/012024

Queens U. (n.d.). https://library.queensu.ca/webedu/grad/Purpose_of_the_Literature_Review.pdf

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. doi:10.1007/BF00116251

Ramya, S., & Radha, N. (2016). Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms. *International Journal of Innovative Research in Computer and Communication Engineering, 4*(1). DOI: .0401049 81210.15680/IJIRCCE.2016

Reddy, C. K., & Aggarwal, C. C. (2015). Healthcare Data Analytics. CRC Press Taylor & Francis Group. doi:10.1201/b18588

Shankar, K., Manickam, P., Devika, G., & Ilayaraja, M. (2018). Optimal Feature Selection for Chronic Kidney Disease Classification using Deep Learning Classifier. *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. doi:10.1109/ICCIC.2018.8782340

Sharma, S., Sharma, V., & Sharma, A. (2016). Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis. *International Journal of Modern Computer Science, 4*(3), 11-16.

Sheety, A. R., Ahmed, F. B., & Naik, V. M. (2019). CDK prediction using Data Mining Techniques as SVM and KNN with Pycharm. *International Research Journal of Engineering and Technology, 6*(5).

Sinha, P., & Sinha, P. (2015). Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM. *International Journal of Engineering Research & Technology, 4*(12). www.ijert.org

Song, X., Qiu, Z., & Jianwei. (2012). Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field. *International Journal of Advancements in Computing Technology, 4*(3).

Sriraam, N., Natashav, & Kaur, H. (2006). Data mining approaches for kidney dialysis treatment. *Journal of Mechanics in Medicine and Biology, 6*(2).

Subhashini, R., & Jeyakumar, M.K. (2017). Performance Analysis of Different Classification Techniques for the Prediction of Chronic Kidney Disease. *International Journal of Pharmacy & Technology, 9*(4).

Urinary Incontinence. (n.d.). In *WebMd*. https://www.webmd.com/urinary-incontinence-oab

Vijayarani, S., & Dhayanand, S. (2015a). Kidney Disease Prediction Using SVM And ANN Algorithms. *International Journal of Computing and Business Research, 6*(2).

Vijayarani, S., & Dhayanand, S. (2015b, August). Data Mining Classification Algorithms for Kidney Disease Prediction. *International Journal on Cybernetics & Informatics*, *4*(4). Advance online publication. doi:10.5121/ijci.2015.4402 13

Zadeh, M.K., Rezapour, M., & Sepehri, M.M. (2013). Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients. *International Journal of Hospital Research, 2*(1), 49-54.

Zhang, H., Hung, C., Chu, W. C., Chiu, P., & Tang, C. Y. (2018). Chronic Kidney Disease Survival Prediction with Artificial Neural Networks. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* doi:10.1109/BIBM.2018.8621294

Zhao, J., Gu, S., & McDermaid, A. (2019). Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Mathematical Biosciences*, *310*, 24–30. doi:10.1016/j.mbs.2019.02.001 PMID:30768948

V. Shanmugarajeshwari is a Research Scholar, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu, India.

M. Ilayaraja is Assistant Professor, Department of Computer Science and Information Technology, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu, India.