

Human Disease Prediction Using Machine Learning Classification Algorithms

Dillip Narayan Sahu¹, Satheesh Kumar²

¹Research Scholar, Department of Computer Science and Application, OPJS University, Churu, Rajasthan, India ²Professor, Department of Computer Science and Engineering, Chaitanya (Deemed to be University), Telangana, India

ABSTRACT

The importance of healthcare system is growing and the pandemic has proved that the healthcare management is an important part of an individual's life. Most medical cases requires proper diagnoses in a prior consultation, and it is very important to get an accurate prediction of the disease. In this paper, we have taken some most common and very serious human diseases as an input in our dataset, after preprocessing, data cleaning, and using different machine learning classification algorithms such as Random Forest, Naïve Bayes, Ada Boost, Bagging Stacking, we found that the Machine learning classification algorithms can be a better option for the early prediction with more accuracy for the human diseases.

Keywords: Algorithm, Classifier, Human Diseases, Machine Learning, Prediction

1. Introduction

Machine learning tools are widely used in all fields of science and medicine and are responsible for revolutionizing businesses everywhere. Healthcare systems, on the other hand, are very slow to adopt these advances and are far behind [1][2]. Machine learning is often useful in the treatment of chronic diseases, namely kidney disease, heart diseases, diabetes etc[3][4]. In fact, machine learning is already being used to predict diabetes risk supported by genomic data, supported by EHR data to diagnose diabetes to predict risk of complications. The introduction of machine learning methods can significantly increase the detection and early treatment of diabetes complications in patients [5][6].

The medical disease prediction application can also be used. basic knowledge about the disease and can tell us if we should seek immediate medical attention for temporary relief or at least start with home remedies. Combining machine learning with an API for user interaction provides an opportunity to facilitate interaction with users by using a machine learning model to make more accurate predictions.[7]

Chronic renal disorder (CKD) is a major burden on the healthcare system because of its increasing prevalence, high risk of progression to end-stage renal disease, and poor morbidity and mortality prognosis. it's rapidly becoming a global health crisis. Unhealthy dietary habits and insufficient water consumption are significant contributors to the present disease. Without kidneys, an individual can only live for 18 days on average, requiring kidney transplantation and dialysis. it's critical to have reliable techniques at predicting CKD in its early stages[8][9]. Machine learning (ML) techniques are excellent in predicting CKD. the present study offers a methodology for predicting CKD status using clinical data, which includes data preprocessing, a way for managing missing values, data aggregation, and have extraction. variety of physiological variables, also as ML techniques such as logistic regression (LR), decision tree (DT) classification, and -nearest neighbor (KNN), were utilized in this work to train three distinct models for reliable prediction. The LR classification method was found to be the foremost accurate in this role, with an accuracy of about 97 percent during this study. The dataset that was utilized in the creation of the technique was the CKD dataset, which was made available to the general public. Compared to prior



research, the accuracy rate of the models employed during this study is considerably greater, implying that they're more trustworthy than the models used in previous studies as well. an outsized number of model comparisons have shown their resilience, and therefore the scheme may be inferred from the study's results[10].

Diabetes is one of the serious diseases and many people suffer from this disease. Aging, obesity, lack of exercise, hereditary diabetes, lifestyle habits, unbalanced diet, hypertension, etc., can cause diabetes. People with diabetes are at high risk for diseases such as heart disease, kidney disease, stroke, eye disease, and nerve damage. Current practice in hospitals is to collect the information necessary for diagnosing diabetes through various tests, and based on that, the diagnosis and appropriate treatment are made. Big data analytics plays an important role in the healthcare industry. The healthcare industry has a huge database. Big data analytics can be used to explore massive data sets, find hidden information and patterns, discover knowledge from data, and predict corresponding outcomes. Existing methods are not very accurate in classification and prediction. In this article, we proposed a diabetes prediction model to better classify diabetes. It contains few external factors that contribute to diabetes besides the usual factors such as glucose, BMI, age and insulin. The new dataset has better classification accuracy compared to the existing dataset. Additionally, a diabetes prediction pipeline model was applied to improve classification accuracy.

2. Experiments and Observations

We have used different dataset for different diseases which are in csv as well as arff format, for the analysis, we have taken Weka machine learning tool and also for the purpose of preprocessing, cleaning, classification and accuracy acceptance comparative analysis purpose.

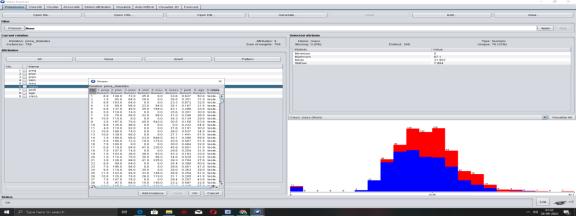


Fig.1 Preprocess of Diabetes Dataset having 10 Attributes

Open file	Open URL.	Open Di	a	0en	arate	Undo	Edit	Save
None								Apply
nt relation					Selected attribute			
lation: kidney_disease ances: 400				Attributes: 26 Sum of weights: 400	Name: hemo Missing: 52 (13%)	Distinct 115		Type: Numeric Inique: 28 (7%)
85					Statistic Minimum		Value 3.1	
All	None	Invert	Pat	ttern	Maximum Mean BldDev		17.8 12.525 2.913	
Name	G Viewer				unitativ	×	2.013	
1 id 2 age	Relation: kidney_disease							
3 bp 4 sg	No. 1:1d 2:age 3:bp 4:sg 5:a	al 6:su 7:rbc 8:pc 9:pcc 10:b etc Numeric Nominal Nominal Nominal Nomin	a 11:bgr 12:bu 13	I: sc 14: sod 15: pot 1	hema 17:pcv 18:wc 19:rc 20:htm	21:1		
5 🔲 al	1 0.0 48.0 00.0 1.02 1	1.0 0.0 not notp notp.	. 121.0 36.0	1.2	15.4 44.0 780 5.2 yes	Yes A		
6 su 7 fpc	2 1.0 7.0 50.0 1.02 4 3 2.0 52.0 80.0 1.01 2	4.0 0.0 nor notp notp. 2.0 3.0 nor not notp notp.		0.6	11.3 30.0 600 no 9.6 31.0 750 no	no Yes		
9 🔲 pc	4 3.0 48.0 70.0 1.005 4	4.0 0.0 nor abn pres notp.	117.0 56.0	3.8 111.0 2.5	11.2 32.0 670 3.9 yes	ne		
9 pcc 10 ba	6 5.0 60.0 90.0 1.015 3	2.0 0.0 nor not notp notp. 3.0 0.0 notp notp.	. 74.0 25.0	1.4	12.2 39.0 780 4.4 yes	no yes		
11 bgr 12 bu		0.0 0.0 nor notp notp. 2.0 4.0 nor abn notp notp.		24.0 104.0 4.0	12.4 36.0 no 12.4 44.0 630 5.0 no	no		
13 🔜 sc	9 8.0 52.0 100.0 1.015 3	3.0 0.0 nor abn pres notp.	. 138.0 60.0	1.0	10.8 33.0 960 4.0 yes	yes		
14 sod 15 pot	10 9.0 53.0 90.0 1.02 2 11 10.0 50.0 60.0 1.01 2	2.0 0.0 abn., abn., pres., notp. 2.0 4.0 abn., pres., notp.		7.2 114.0 3.7 4.0	9.5 29.0 121 3.7 yes 9.4 28.0 yes	988 995		
16 hemo	12 11.0 63.0 70.0 1.01 3	3.0 0.0 abn abn pres notp.		2.7 131.0 4.2	10.0 32.0 450 3.0 yes	yes		 Visua
17 pcv 18 wc	13 12.0 68.0 70.0 1.015 1 14 13.0 68.0 70.0	3.0 1.0 nor pres notp. notp notp.		2.1 138.0 5.8 4.6 135.0 3.4	9.7 20.0 122 3.4 yes 9.8 ves	yes ves		
19 rc 20 http	15 14.0 68.0 80.0 1.01 1 16 15.0 40.0 80.0 1.015 1	3.0 2.0 nor abn pres pres 3.0 0.0 nor notp notp.		4.1 130.0 6.4 9.6 141.0 4.9	5.6 16.0 110 2.8 yes 7.6 24.0 380 2.8 yes	yes no		82
21 🔲 dm	17 16.0 47.0 70.0 1.015 2	2.0 0.0 nor notp notp.		2.2 130.0 4.1	12.6 no	no	50	
22 cad 23 appot	18 17.0 47.0 80.0 19 18.0 60.0 100.0 1.025 5	notp notp. 0.0 3.0 not notp notp.	. 114.0 97.0	5.2 139.0 3.7 1.3 135.0 4.3	12.1 yes 12.7 37.0 114 4.3 yes	no		
24 pe	20 19.0 62.0 60.0 1.015 1	1.0 0.0 abn pres notp.	100.0 31.0	1.6	10.3 30.0 530 3.7 yes	no	47 49	
25 ene 26 classification	21 20.0 61.0 80.0 1.015 2 22 21.0 60.0 90.0	2.0 0.0 abn abn notp notp.	. 173.0 149.0	3.9 135.0 5.2 76.0 4.5	7.7 24.0 920 3.2 yes 10.9 32.0 620 3.6 yes	yes yes		
To C construction	23 22.0 48.0 80.0 1.025 4	4.0 0.0 nor abn notp notp.	95.0 163.0	7.7 136.0 3.8	9.8 32.0 690 3.4 yes	no		
	25 24.0 42.0 100.0 1.015 4	0.0 0.0 nor notp notp. 4.0 0.0 nor abn notp pres	50.0	1.4 129.0 4.0	na 11.1 39.0 830 4.6 yes	no 30		
		0.0 0.0 nor notp notp. 0.0 0.0 nor notp notp.		1.9 141.0 5.2 2.4 140.0 3.4	9.9 29.0 040 3.7 yes 11.6 35.0 103 4.0 yes	yes v25		20
	28 27.0 69.0 70.0 1.01 2	3.0 4.0 nor abn notp notp.	264.0 97.0	2.7 130.0 4.0	12.5 37.0 960 4.1 yes	yes .		
	2 200 760 700	10 00 solis	400.0 04.0		,	7		
					dd instance Undo OK C	ancel		
	Re	mave						
					5.1		10.49	
								Log 🚙

Fig.2 Preprocess of Kidney Dataset having 26 Attributes



International Journal of Engineering Technology and Management Sciences Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022 DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

Open file	Open URL	Open DB	Gen	Undo Undo			Edit		Save		
None										Apply St	
nt relation				Selected attribute							
Relation: Indian_liver_patient stances: 583			Attributes: 11 Sum of weights: 583	Name: Age Missing: 0 (0%)	Di	itinct 72		Type: Nume Unique: 10 (29			
utes				Statistic			alue				
				Minimum Maximum		4	1 10				
All	None	Inwert	Pattern	Mean		1	14.746				
Name				StdDev		1	6.19				
1 Age											
2 🗌 Gender		Q Viewer								×	
3 🔲 Total_Bilirubin		-								×	
4 Direct_Bilirubin 5 Alkaline_Phosphotase		Relation: Indian_liver_patient									
6 Alamine_Aminotransferase		No. 1: Age 2: Gender 3: Total_Bill Numeric Neminal Numeric		Ajkaline_Phosphotase 6: Numeric	Alamine_Aminotransferase 7: Aspartate					Uneric	
7 📃 Aspartate_Aminotransferase		1 65.0 Female	0.7 0.1	187.0	16.0	18.0	6.8	3.3	0.9	1.0 🔺	
8 Total_Protiens 9 Albumin			10.9 5.5 7.3 4.1	699.0	64.0	100.0	7.5 7.0	3.2	0.74	1.0	
10 Albumin_and_Globulin_Ratio		3 62.0 Male 4 58.0 Male	7.3 4.1 1.0 0.4	490.0 182.0	60.0 14.0	68.0 20.0	7.0	3.3 3.4	0.89	1.0	
11 📃 Dataset		5 72.0 Male	3.9 2.0	195.0	27.0	59.0	7.3	2.4	0.4	1.0	
		6 46.0 Male	1.8 0.7	208.0	19.0	14.0	7.6	4.4	1.3	1.0	
		7 26.0 Female 8 29.0 Female	0.9 0.2	154.0 202.0	16.0	12.0	7.0	3.5 3.6	1.0	1.0	
		8 29.0 Female 9 17.0 Male	0.9 0.3	202.0	14.0	11.0	6.7 7.4	3.6	1.1		
		10 55.0 Male	0.7 0.2	290.0	53.0	58.0	6.8	3.4	1.0	1.0	
		11 57.0 Male	0.6 0.1	210.0	51.0	59.0	5.9	2.7	0.8	1.0	
		12 72.0 Male	2.7 1.3 0.9 0.3	260.0 310.0	31.0 61.0	56.0 58.0	7.4	3.0	0.6	1.0	
		13 64.0 Male 14 74.0 Female	0.9 0.3	310.0	61.0	58.0	7.0	3.4 4.1	0.9	2.0	
		15 61.0 Male	0.7 0.2	145.0	53.0	41.0	5.8	2.7	0.87	1.0	
		16 25.0 Male	0.6 0.1	183.0	91.0	53.0	5.5	2.3	0.7	2.0	
		17 38.0 Male	1.8 0.8 1.6 0.5	342.0	168.0	441.0	7.6	4.4	1.3	1.0	
		18 33.0 Male 19 40.0 Female	1.6 0.5 0.9 0.3	165.0 293.0	15.0 232.0	23.0 245.0	7.3	3.5 3.1	0.92	2.0	
		20 40.0 Female	0.9 0.3	293.0	232.0	245.0	6.8	3.1	0.8	1.0	
		21 51.0 Male	2.2 1.0	610.0	17.0	28.0	7.3	2.6	0.55	1.0	
		22 51.0 Male	2.9 1.3	482.0	22.0	34.0	7.0	2.4	0.5	1.0	
		23 62.0 Male 24 40.0 Male	6.8 3.0 1.9 1.0	542.0 231.0	116.0	66.0 55.0	6.4 4.3	3.1 1.6	0.9	1.0	
		25 63.0 Male	0.9 0.2	194.0	52.0	45.0	6.0	3.9	1.85	2.0	
		26 34.0 Male	4.1 2.0	289.0	875.0	731.0	5.0	2.7	1.1	1.0	
		27 34.0 Male	4.1 2.0	289.0	875.0	731.0	5.0	2.7	1.1	1.0	
		28 34.0 Male 29 20.0 Male	6.2 3.0 1.1 0.5	240.0 128.0	1680.0 20.0	850.0 30.0	7.2	4.0 1.9	1.2	1.0 2.0	
		30 84.0 Female	0.7 0.2	188.0	13.0	21.0	6.0	3.2	1.1	2.0	
	Remave	31 57.0 Male	4.0 1.9	190.0	45.0	111.0	5.2	1.5	0.4	1.0	
		32 52.0 Male	0.9 0.2	158.0	35.0	44.0	4.9	2.9	1.4	1.0	
		22 67.0 Mola	4.0 0.2	107.0	10.0	22.0	6.2				
								Add instance	Undo OK		
										LUG	

Fig.3 Preprocess of Liver Dataset having 11 Attributes

Algorithm taken- UltraBoost

Classifier Output-=== Run information ===
Scheme: weka.classifiers.meta.UltraBoost -S 1 -B "weka.classifiers.meta.FilteredClassifier -F
$\$ we kalcular sine the sine that the sine the sine that the sine the s
weka.classifiers.bayes.NaiveBayes" -B "weka.classifiers.meta.FilteredClassifier -F
\"weka.filters.unsupervised.attribute.RemoveType -V -T numeric\" -S 1 -W
weka.classifiers.functions.LogisticR 1.0E-8 -M -1 -num-decimal-places 4"
Relation: pima_diabetes
Instances: 768
Attributes: 9
preg plas pres skin insu mass
pedi age class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
UltraBoost
Base classifiers
FilteredClassifier using weka.classifiers.bayes.NaiveBayes on data filtered through
weka.filters.unsupervised.attribute.RemoveType -V -T nominal
Filtered Header
@relation pima_diabetes-weka.filters.unsupervised.attribute.RemoveType-V-Tnominal
@attribute class {tested_negative,tested_positive}
@data
Classifier Model
Naive Bayes Classifier
Class
Attribute tested_positive
(0.65) (0.35)
FilteredClassifier using webs classifiers functions Logistic -R 1.0E-8 -M -1 -num-decimal-places A

FilteredClassifier using weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4 on data filtered through weka.filters.unsupervised.attribute.RemoveType -V -T numeric



Website: ijetms.in Issue: 6 Volume No.6 October - November - 2022 DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

Filtered Header Classifier Model Logistic Regression with ridge parameter of 1.0E-8 Coefficients...

Variable	Class tested_negative	
preg	-0.1234	
plas	-0.0351	
pres	0.0131	
skin	-0.0004	
insu	0.0012	
mass	-0.0901	
pedi	-0.9771	
age	-0.0159	
Intercept	8.2873	

Odds Ratios...

	Class	
Variable	tested_negative	
preg	 0.8839	
plas	0.9656	
pres	1.0132	
skin	0.9996	
insu	1.0012	
mass	0.9138	
pedi	0.3764	
age	0.9842	

Time taken to build model: 0.03 seconds === Stratified cross-validation ===== Summary === **Correctly Classified Instances** 590 76.8229 % Incorrectly Classified Instances 23.1771 % 178 Kappa statistic 0.4458 Mean absolute error 0.3795 Root mean squared error 0.4124 Relative absolute error 83.4983 % Root relative squared error 86.5164 % **Total Number of Instances** 768 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class 0.914 0.504 0.772 0.914 0.837 0.892 0.465 0.832 tested_negative 0.496 0.086 0.756 0.496 0.599 0.465 0.832 0.715 tested_positive Weighted Avg. 0.768 0.358 0.766 0.768 0.754 0.465 0.832 0.830 === Confusion Matrix === a b <-- classified as $457 \ 43 \mid a = tested_negative$ 135 133 | b = tested_positive



International Journal of Engineering Technology and Management Sciences Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022 DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

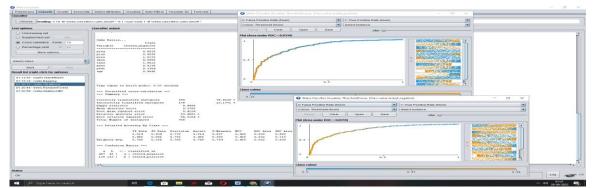


Fig.4 UltraBoost Classifier with Visualize curve (for Diabetes disease)

Algorithm taken Classifier Outpu			•	e)					
Scheme:		weka.classifi		AdaRoostM	[1_P_1	00 -S 1	-I	10	-W
weka.classifiers.			1015.1110ta. <i>1</i>	MaDOOstivi		00 -5 1	-1	10	- • •
	ney_disease	notump							
Instances: 400	•								
Attributes: 26									
id	age	bp	sg	al	su	rbc	pc		
pcc	ba	bgr	bu	SC	sod	pot	1		
hemo	pcv	wc	rc	htn	dm	cad			
appet	pe	ane							
classific									
Test mode: 10	-fold cross-v	alidation							
=== Classifier n	nodel (full tra	aining set) =	==						
AdaBoostM1: N	lo boosting p	ossible, one	classifier	used!					
Decision Stump									
Classifications									
id <= 249.5 : ck									
id > 249.5 : note	kd								
id is missing : cl	sd								
Class distributio	ns								
id <= 249.5									
ckd notckd									
1.0 0.0									
id > 249.5									
ckd notckd									
0.0 1.0									
id is missing									
ckd notckd									
0.625 0.375									
Time taken to bu									
=== Stratified ci			-						
Correctly Classi									
Incorrectly Class	sified Instanc		0.5	5 %					
Kappa statistic		0.9893							
Mean absolute e		0.005	~ -						
Root mean squar		0.070							
Relative absolut	e error	1.0663	%						

International Journal of Engineering Technology and Management Sciences



Website: ijetms.in Issue: 6 Volume No.6 October - November – 2022 DOI:10.46647/ijetms.2022.v06i06.002 ISSN: 2581-4621

Root relat Total Nur === Deta Weighted === Conf a b < 249 1	mber c iled A TP R 0.990 0.993 Avg. fusion clas a = c	of Ins accur ate 6 0 3 0 0.9 Mat sifie kd	stance acy I FP R .007 .004 995 rix =	es By Cla ate P 0.99 0.99 0.00	ass == recisi 96 93	on R 0.990 0.993	ecall 5 0.9 3 0.9	996 993		89 89	ICC 0.995 0.995).989	ROC 0.99 0.98 0.995	95 c 19 n	PRC A kd otckd 993	Area Class
Classifier Choose of Test optimus Ose Pani Supplied @ Cross-val	Community Olivitier A EducedTages Implet EducedTages Implet EducedTages Implet EducedTages Implet I		- 249, 5 1 ctcl - 249,				L 0.5 - 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2001 Turindia to the the the the the the the the	er, (Chur value (Ad) Open Br Ak Churdhy Violada na Postoka Atat (A Postoka Atat (A) Chur value (Ad) Chur va	e. 5	n Sa	a Paska Ada Jako			
	Fig							th Vi						lisease)	
							ith al	The second secon	Alternie	butes inter-	s (for]	Kidney	/ disea	ase)	
nametrik Marine Provinski (1990) Marine Mari	et dame							x ! +	Pattscaling Code Code	108	»	Liver			100 Fame)

Fig.7 Visualize curve with all the attributes (for Liver disease)

3. Discussion

We have taken 5 different machine learning algorithms and with the help of different experimental observations using the machine learning tool, it is found clearly that the ML is giving good accuracy rate to analyze, detection and prediction for the different human diseases. In the above observations, it is seen that, machine learning algorithms are no doubt an excellent method to predict different human diseases at an early stage. It is found that the accuracy level is acceptable and so will be efficient for the medical sciences.



4. Conclusion

In the study of the above real time medical dataset analysis, experimentation and observation in different parameters of algorithms, it is found that the accuracy level using the machine learning classification model AdaBoostM1 is much satisfactory, having good accuracy rate of 99.5% (for kidney disease) and so will be a good option in the field of medical health care sector to opt or to predict early prediction with proper diagnosis of different human diseases.

References

1. Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B. & Yang, C. W. (2013). Chronic kidney disease: global dimension and perspectives. The Lancet, 382(9888), 260-272.

2. Ali, S., Dave, N., Virani, S. S., & Navaneethan, S. D. (2019). Primary and secondary prevention of cardiovascular disease in patients with chronic kidney disease. Current Atherosclerosis Reports, 21(9), 1-9.

3. Levey, A. S., Coresh, J., Bolton, K., Culleton, B., Harvey, K. S., Ikizler, T. A. & Briggs, J. (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. American Journal of Kidney Diseases, 39(2 SUPPL. 1), i-ii+.

4. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

5. Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 291-296). IEEE.

6. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

7. Noble, W. S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12), 1565-1567.

8. Pregibon, D. (1981). Logistic regression diagnostics. Annals of Statistics, 9(4), 705-724.

9. Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. International Journal of Advances in Engineering & Technology, 7(1), 242.

10. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2), 1-12.

11. Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. International Journal of Engineering Research and Technology, 4(7), 206–210.