International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Employing Machine Learning Models in the Prediction and Diagnosis of Chronic Kidney Disease

Anandkumar A. Sutariya[1], **Dr. Dushyantsinh B. Rathod** [2]

Submitted: 28/01/2024 **Revised:** 06/03/2024 **Accepted:** 22/03/2024

Abstract*:* Chronic Kidney Failure is the medical term for chronic kidney disease. It portrays the moderate disintegration of renal disappointment and how, assuming that constant kidney infection has advanced to a high level stage, a high volume of fluid and undesirable electrolytes may develop in the body. We may see less evidence of chronic renal disease in the early phases.The treatment for chronic kidney disease focuses on slowing down the process of kidney damage. Without a trace of dialysis or kidney migration, persistent renal sickness can advance to the last periods of kidney annihilation, which is inoperable. The focal point of this examination is on early discovery of constant obstructive pneumonia illness utilizing different AI techniques, which are K-Nearest Neighbour, Decision Tree and Bayesian Classifier.

Keywords: *Chronic Kidney Disease, K-Nearest Neighbor, Decision Tree, Bayesian Classifier, Machine Learning and AI techniques.*

1. Introduction

Kidney disease is not a kidding medical problem that makes a ton of terrible things end up peopling, particularly in low-and centre pay countries, where a great many individuals bite the dust consistently inferable from an absence of therapy. The lethality of each persistent sickness is corresponding to the stage it has reached with-out being restored. Diabetic patients are turning out to be more normal, as are hyper-tension, coronary illness, diabetes, and a family background of renal disappointment. On the off chance that disease goes unrecognized and untreated, it can prompt hyper-tension and, in the most dire outcome imaginable, renal disappointment. We focus on CKD, which can help patients in an assortment of ways whenever identified early and suitably. It works on the possibilities of a fruitful treatment while additionally extending the patient's life. The objective of this paper is to utilize a few AI calculations to early analyse constant obstructive aspiratory infection.

Classification is a type of data analysis in which models defining relevant data classes are extracted.

- K-Nearest Neighbour (K-NN): It is a simple strategy that employs classification and regression algorithms. Whether k-NN is used for classification or regression, the output of this method varies.
- Decision Tree (DT): An information base choice tree is a tree-like design of in- formation. It is utilized in tasks exploration and AI to pursue choices that will prompt a significant end, as well as in information mining to characterize information and recover information.
- Bayesian Classifier: It is based on the probability theorem and can be used to perform medical diagnosis in a rational manner, especially in automated medical diagnosis decision support systems. It can deal with an unlimited number of independent variables, both continuous and categorical.

The paper is divided into several sections. The research works relevant to this study are discussed in Section 2. The research objectives were outlined in Section 3. The findings of all approaches are shown in Section 4. Section 5 brings the research to a close, while Section 6 discusses the research's future directions.

2. Literature survey

AI calculations have for quite some time been utilized to settle difficulties in the clinical field. This has been endeavored by various scientists. For the order and expectation of the patient's disease status, different procedures and techniques have been utilized. They fostered a choice emotionally supportive network that utilizes characterization calculations to analyze and foresee ongoing renal disappointment.

J. Snegha et al. [1] tried two information mining strategies: irregular backwoods (RF) and back spread calculation. Various models are fabricated utilizing a CKD dataset acquired from www.kaggle.com, and the exhibition of these techniques is contrasted all together with figure out which is the best in foreseeing constant kidney illness. As indicated by the aftereffects of their investigations, the back engendering calculation has a conviction of 98.40 percent, contrasted with RF's 88.7 per-cent.

R. Gupta et al. [2] utilized three AI techniques: choice tree (DT), irregular woods (RF), and strategic relapse (LR). These models are assembled utilizing a CKD dataset acquired from the UCI AI store, and their exhibition is assessed to decide the best classifier for anticipating persistent kidney sickness. As per the preliminary information, the LR classifier has the best exactness of 99.24 percent and the most noteworthy review of 100%. Moreover, subsequent to preparing the dataset, DT has the greatest accuracy

1 PhD Scholar, Computer Engineering, Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar, Gujarat, India
ORCID ID: 0000-0001-9046-2919
2 Professor & HoD, Computer Engineering, Ahmedabad Institute of Technology, Ahmedabad, Gujarat, India
ORCID ID: 0000-0001-9155-660X
** Corresponding Author Email: asutariya82@email.com*

of 100%.

In order to diagnose chronic renal illness, S. Vashisth et al. [3], used multi-layer perceptron, support vector machine (SVM), and naïve bayes classifiers. He used a dataset taken from Apollo hospitals across India, and the testing results suggest that the multi-layer perceptron had the best accuracy of all, at 92.5 percent.

J. R. Lambert et al. [4] proposed a Correlation-based highlight choice - consecutive least streamlining (CFS-SMO) and Ranker-successive least advancement (Ranker-SMO) for relative examination of numerical and ostensible traits of constant kidney infection with Special clinical dataset for kidney disease, between 2 years to 83 years old, was acknowledged and distributed at the UCI AI archive. The CFS-SMO beats other Ranker-SMOs concerning results. With CFS-SMO, ostensible traits are utilized to further develop order.

P. Arulanthu and E. Perumal [5] has gathered information from Apollo Hospitals and given it to the UCI information vault, where it will be utilized as a preparation dataset for driving order calculations. JRip, Naïve Bayes, Sequential Minimal Optimization, and Instance Based Learner are the four calculations. While contrasting the presentation of the JRip and the exhibition of the other calculation, the presentation outline plainly shows that the JRip performs better (for example precision of 98.8%).

3. Proposed Architecture and Dataset

3.1. Proposed Architecture

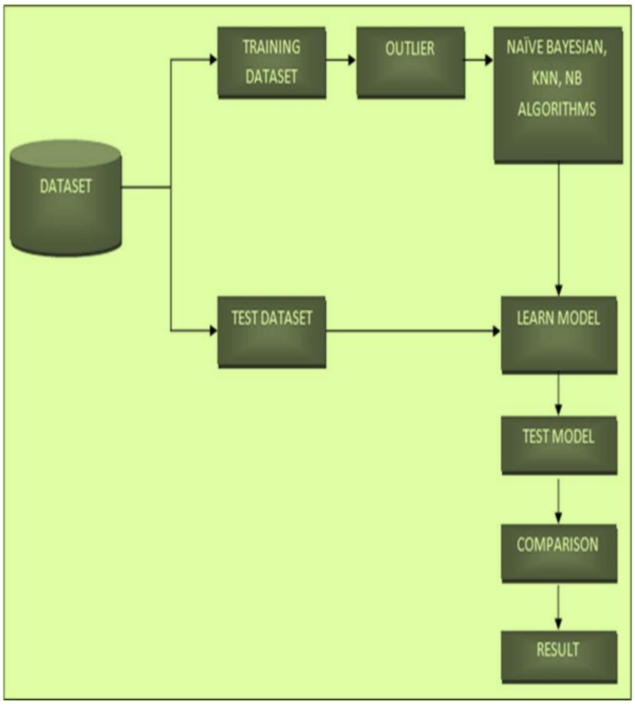
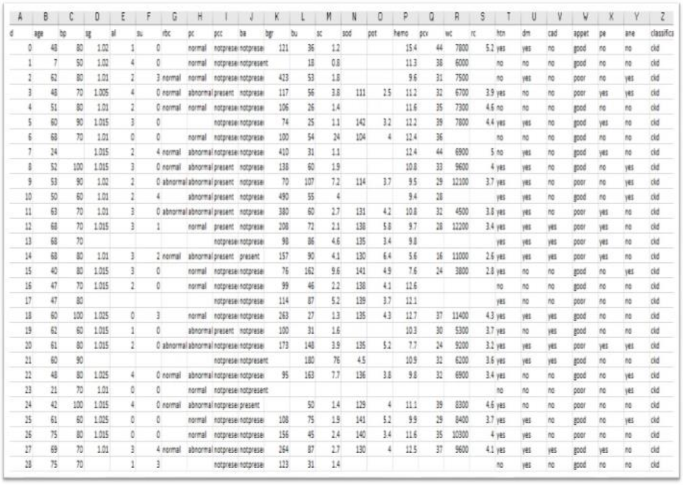


Fig. 1. Proposed Architecture.

1. Data Preparation:
    - Dataset: Start with the complete dataset.
    - Splitting the Data: Divide the dataset into two subsets:
        - Training Dataset: Used for training the model.
        - Test Dataset: Used for testing the model's performance.
2. Outlier Detection:
    - Outlier Handling: Check the training dataset for outliers and handle them accordingly (e.g., remove or adjust outliers).
3. Algorithm Selection:
    - Choose one or more algorithms for training. The options provided include:
        - Naïve Bayesian (NB): A probabilistic classifier based on Bayes' Theorem.
        - K-Nearest Neighbors (KNN): A non-parametric method that classifies a sample based on the majority vote of its neighbors.
        - Naïve Bayes (NB): A different variant or implementation of the Naïve Bayes classifier.
4. Model Learning:
    - Learn Model: Train the selected algorithm(s) on the training dataset.
5. Model Testing:
    - Test Model: Evaluate the model's performance using the test dataset.
6. Performance Comparison:
    - Comparison: Compare the performance of the trained models. This step might involve evaluating metrics like accuracy.

3.2. Dataset



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
d	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classifica
0	48	80	1.02	1	0		normal	notprese	notprese	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	7	50	1.02	4	0		normal	notprese	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
2	62	80	1.01	2	3	normal	normal	notprese	notprese	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
3	48	70	1.005	4	0	normal	abnormal	present	notprese	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	51	80	1.01	2	0	normal	normal	notprese	notprese	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
5	60	90	1.015	3	0			notprese	notprese	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	68	70	1.01	0	0		normal	notprese	notprese	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
7	24		1.015	2	4	normal	abnormal	notprese	notprese	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
8	52	100	1.015	3	0	normal	abnormal	present	notprese	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
9	53	90	1.02	2	0	abnormal	abnormal	present	notprese	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	50	60	1.01	2	4		abnormal	present	notprese	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
11	63	70	1.01	3	0	abnormal	abnormal	present	notprese	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	68	70	1.015	3	1		normal	present	notprese	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	68	70						notprese	notprese	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
15	40	80	1.015	3	0		normal	notprese	notprese	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
16	47	70	1.015	2	0		normal	notprese	notprese	99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
17	47	80						notprese	notprese	114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
18	60	100	1.025	0	3		normal	notprese	notprese	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
19	62	60	1.015	1	0		abnormal	present	notprese	100	31	1.6			10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
20	61	80	1.015	2	0	abnormal	abnormal	notprese	notprese	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
21	60	90						notprese	notpresent		180	76	4.5		10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
22	48	80	1.025	4	0	normal	abnormal	notprese	notprese	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
23	21	70	1.01	0	0		normal	notprese	notpresent										no	no	no	poor	no	yes	ckd
24	42	100	1.015	4	0	normal	abnormal	notprese	present		50	1.4	129	4	11.1	39	8300	4.6	yes	no	no	poor	no	no	ckd
25	61	60	1.025	0	0		normal	notprese	notprese	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes	no	good	no	yes	ckd
26	75	80	1.015	0	0		normal	notprese	notprese	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes	no	poor	no	no	ckd
27	69	70	1.01	3	4	normal	abnormal	notprese	notprese	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes	yes	good	yes	no	ckd
28	75	70		1	3			notprese	notprese	123	31	1.4							no	yes	no	good	no	no	ckd

Fig. 2. Dataset.

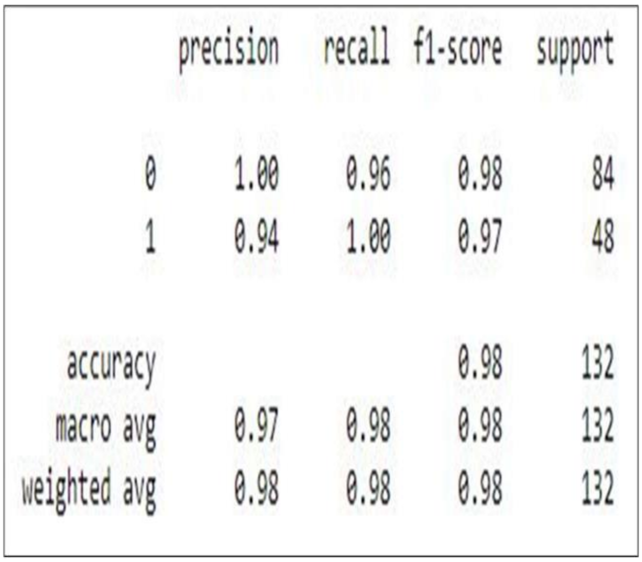
The dataset shown in Fig. 2 is taken over 2-month period in India. It has 400 rows with 25 features like red blood cells, pedal edema, sugar, etc. The aim is to classify whether a patient has chronic kidney disease or not. The classification is based on a attribute named 'classification' which is either 'ckd'(chronic kidney disease) or 'notckd. I've performed cleaning of the dataset which includes mapping the text to numbers and some other changes. After the cleaning I've done some EDA (Exploratory Data Analysis) and then I've divided the dataset into training and testing and applied the models on them. It is observed that the classification results are not much satisfying initially. So, instead of dropping the rows with Nan values I've used the lambda function to replace them with mode for each column. After that I've divided the dataset again into training and testing sets and applied models on them. This time the results are better and we see that the decision tree is the best performers with an accuracy of 97%. The performance of the classification is measured by printing confusion matrix, classification report and accuracy.

The essential objective of this study is to exhibit the worth of information mining in surveying way of life related messes. It is tried to audit the writing piece whose ex-amination action is centered on the two specialists and patients. The fundamental focuses (disease, technique, discoveries, precision) of the different review studies will be featured, as well as the use of instruments or strategies. At last, the objective is to recognize what locales request additional consideration from information mining and AI devices.

The examination objectives are to sort material as per conduct informatics and group information to dissect information designs. By surveying demonstrative data with regulated and solo AI calculations, we desire to work on the symptomatic execution of present indicative methodologies for infection forecast. To survey the pro-posed approach's presentation utilizing models, for example, accuracy, review, F-measure, and precision with characterization rate. On various datasets, analyze the exhibition of various classifiers and bunching calculations.

4. Results

Fig. 3 shows the precision, recall, F1, and support scores for the proposed model in the classification report for the algorithm decision tree.



```
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        84
           1       0.94      1.00      0.97        48

    accuracy                           0.98       132
   macro avg       0.97      0.98      0.98       132
weighted avg       0.98      0.98      0.98       132
```

Fig. 3. Classification Report of Decision Tree (Precision, Recall, F1-score)

Accuracy suggests a classifier's ability to do whatever it takes not to name a negative event as certain. Not entirely settled as the extent of authentic up-sides of how many certifiable up-sides and deceiving up-sides for each class. It's generally called positive assumption precision.

The constraint of a classifier to observe every one of the specific cases is known as review. Still up in the air as the extent of certifiable empowering focuses on how many veritable up- sides and misdirecting negatives for each class. It decides the degree of really perceived upsides.

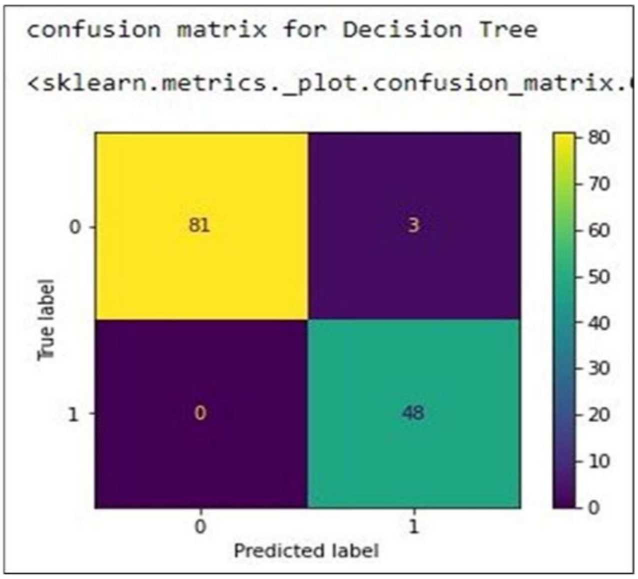
The F1 score is a weighted consonant mean of precision and survey, with 1.0 being the most imperative and 0.0 being the least. They are less careful than accuracy assessments since exactness and survey are considered along with the calculation. To dissect classifier models, utilize the weighted ordinary of F1 rather than overall accuracy as a rule.

The amount of genuine occasions of the class in the given dataset is known as help. Support doesn't shift between models; rather, it dissects the evaluation connection.

As addressed in the confusing organization of the decision tree in Fig. 4, there are four methods for managing to choose if the gauges are correct or not.

- True Negative (TN): the case unendingly was projected to be negative.
- True Positive (TP): the case was positive and should be positive.
- False Negative (FN): the case was positive, but the outcome was projected to be negative.
- False Positive (FP): the case was negative, at this point, it was expected to be positive.

The confusion matrix reveals not just a predictive model's performance, but also which classes are successfully predicted, which are incorrectly forecasted, and what types of errors are being made.



```
confusion matrix for Decision Tree
<sklearn.metrics._plot.confusion_matrix.
```

Fig. 4. Decision Tree Confusion Matrix

At the point when we utilize the term precision, we typically suggest exactness. The quantity of right expectations partitioned by the all-out number of information tests is the proportion. We considered the precision of each of the three calculations in this estimation (for example k- NN, Naïve Bayes and DT) are displayed in Fig. 5.

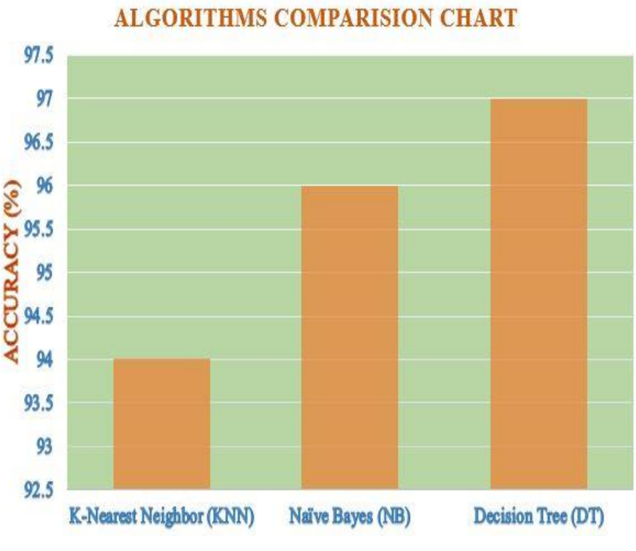


Fig. 5. Accuracy Chart of Machine learning algorithms

In the wake of ascertaining the presentation of proposed models and looking at them all, the best classifier to anticipate Chronic Obstructive Pulmonary Disease was picked. As per the exploratory information, the Decision Tree strategy has the greatest exactness of 97%, contrasted with 96% and 94 percent for the Nave Bayes and k-NN calculations, individually. The outcomes are displayed in Table 1.

Table 1. Comparison of algorithms

Algorithms	Accuracy (%)
K-Nearest Neighbour (KNN)	94
Naïve Bayes (NB)	96
Decision Tree (DT)	97

5. Conclusion

The reason for this review is to foster an original structure in light of bunching and order information digging methods for foreseeing and diagnosing these problems in the medical services region utilizing genomic data sets.

6. Future Direction

The recommended approach may also be applicable to other diseases classification challenges, including datasets of the same sort as those utilized in this study, according to a review of numerous literature papers. However, there is still much work to be done in terms of doing research on clustering, noise removal, and fuzzy rule-based illness diagnosis algorithms in order to fully harness their potential and utility. In the future, the datasets for disease categorization and prediction utilizing incremental machine learning algorithms will demand greater attention. As a result, it is necessary to test this method on additional datasets, particularly large datasets, in order to demonstrate its efficacy in terms of large data computing time. In addition, the study looks into how the proposed technology may be adapted to work with various types of medical datasets.

Author contributions

Anandkumar Sutariya: Role of Primary Author (a) Feasibility Study of Healthcare (b) Requirement Analysis from Health Stakeholder (c) Requirement Gathering from Medical Officers (d) Planning (e) Designing (f) Implementation/Coding [Python / Google Collab] (g) Testing/Validation

Dr. Dushyantsinh Rathod: Guidance about research problem, Documentations, Deadline Maintenance.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] J. Snegha, V. Tharani, S. D. Preetha, R. Charanya and S. Bhavani, "Chronic Kid-ney Disease Prediction Using Data Mining," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vel-lore, India, 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.482.

[2] R. Gupta, N. Koli, N. Mahor and N. Tejashri, "Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154147.

[3] S. Vashisth, I. Dhall and S. Saraswat, "Chronic Kidney Disease (CKD) Diagnosis using Multi-Layer Perceptron Classifier," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 346-350, doi: 10.1109/Confluence47617.2020.9058178.

[4] J. R. Lambert, P. Arulanthu and E. Perumal, "Identification of Nominal Attrib-utes for Intelligent Classification of Chronic Kidney Disease using Optimization Algorithm," 2020 International Conference on Communication and Signal Pro-cessing (ICCSP), Chennai, India, 2020, pp. 0119-0125, doi: 10.1109/ICCSP48568.2020.9182206.

[5] P. Arulanthu and E. Perumal, "Predicting the Chronic Kidney Disease using Var-ious Classifiers," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 70-75, doi: 10.1109/ICEECCOT46775.2019.9114653.

[6] Bai Q, Su C, Tang W, Li Y. "Machine learning to predict end stage kidney disease in chronic kidney disease". Sci Rep. 2022 May 19;12(1):8377. doi: 10.1038/s41598-022-12316-z. PMID: 35589908; PMCID: PMC9120106.

[7] Pal, S. "Prediction for chronic kidney disease by categorical and non_categorical attributes using different machine learning algorithms". Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-15188-1.

[8] Khalid H, Khan A, Zahid Khan M, Mehmood G, Shuaib Qureshi M. "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease". Comput Intell Neurosci. 2023 Mar 14;2023:9266889. doi: 10.1155/2023/9266889. PMID: 36959840; PMCID: PMC10030216.

[9] Ullah Z, Jamjoom M. "Early Detection and Diagnosis of Chronic Kidney Disease Based on Selected Predominant Features". J Healthc Eng. 2023 Jan 30;2023:3553216. doi: 10.1155/2023/3553216. PMID: 36756136; PMCID: PMC9902122.

[10] Islam MA, Majumder MZH, Hussein MA. "Chronic kidney disease prediction based on machine learning algorithms". J Pathol Inform. 2023 Jan 12;14:100189. doi: 10.1016/j.jpi.2023.100189. PMID: 36714452; PMCID: PMC9874070.

[11] Debal, D.A., Sitote, T.M. "Chronic kidney disease prediction using machine learning techniques". J Big Data 9, 109 (2022). https://doi.org/10.1186/s40537-022-00657-5

[12] Modhugu, V.R. and Ponnusamy, S. 2024. "Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree". Asian Journal of Research in Computer Science. 17, 6 (May 2024), 188–201. DOI:https://doi.org/10.9734/ajrcos/2024/v17i6467.

[13] A. E. Topcu, E. Elbasi and Y. I. Alzoubi, "Machine Learning-Based Analysis and Prediction of Liver Cirrhosis," 2024 47th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2024, pp. 191-194, doi: 10.1109/TSP63128.2024.10605929.

[14] Mandakini Priyadarshani Behera, Archana Sarangi, Debahuti Mishra, Shubhendu Kumar Sarangi, "A Hybrid Machine Learning algorithm

for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine", Procedia Computer Science, Volume 218, 2023, Pages 818-827, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2023.01.062.

[15] Md, A.Q.; Kulkarni, S.; Joshua, C.J.; Vaichole, T.; Mohan, S.; Iwendi, C. "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease". Biomedicines 2023, 11, 581. https://doi.org/10.3390/biomedicines11020581

[16] Ruhul Amin, Rubia Yasmin, Sabba Ruhi, Md Habibur Rahman, Md Shamim Reza,"Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms", Informatics in Medicine Unlocked, Volume 36, 2023, 101155, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2022.101155.