IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Survey Study on Chronic Kidney Disease using Machine Learning

[1]Rachna Sable, [2]Ajaykumar Eklare, [3]Ankita Kulkarni, [4]Seema Jadhav, [5]Rutuja Das

[1]Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Dept. of Comp. & IT,
[1]G. H. Raisoni Institute of Engineering & Technology, Pune, India

Abstract: Chronic kidney infection is a major and growing problem in Developing countries. It is one of the most well-known health concerns, with an increasing demand for early detection in order to provide Prosperous and eternal care. Chronic Kidney Disorder can affect one out of every five men and one for every four women worldwide in between both the ages of 65 and 74. (CKD). (CKD) affects 10% of the world population, and millions of people die each year owing to a lack of inexpensive treatment options. Many factors contribute to the progressive decline of kidney function over time, including haemoglobin, blood pressure, diabetes, and obesity. The severity of chronic renal failure is divided into several phases. To minimize further deterioration, a better diagnosis of chronic renal disease is needed. We're looking at alternative models that can detect the presence of CKD based on specific input characteristics.

Index Terms **- Chronic Kidney Disease, UCI.**

INTRODUCTION

chronic kidney disease is a very harmful disease which will lead to kidney failure if not detected at an early stage. Kidney disease is such a disease which will grow slowly and if this disease is not predicted right on time, it can cause a very serious damage. Millions of people around the world who are suffering from kidney disease suffer from serious damage because of very slow prediction of this disease. The prediction time is slow because it grows very slowly with unpredictable symptoms. The need of quick prediction of this disease is necessary as it could save millions of lives. Machine Learning is a methodology which can help in predicting the CKD. A huge dataset with all the important parameters of diagnosing a patient is needed to predict the disease accurately. Then pre-processing the data will help us to deal with the missing values. Later on, important attributes are considered and rest are ignored during the prediction. This step is called Feature Selection. It removes the dimensionality of the dataset or in other words, it only considers a meaningful attribute which can be an important parameter in predicting CKD. A high dimensional dataset could reduce the accuracy due to unwanted attributes involved in it. A better accuracy is expected when it comes to health-related problems. Many researchers have worked on making a better prediction of this disease. This paper will help in understanding on analysing the past work done the researchers over this topic.

The paper is organized in different sections. The next section discusses research works related to these papers. Later the next section elaborates generalized architecture diagram. In the further section the study describes the analysis of various research papers and the final section concludes the Survey.

GENERALIZED ARCHITECTURE

Below is the generalized architecture of the model used in various research papers. The architecture consists of a dataset, Pre-Processing, Feature Selection, Classification methods with or without feature selection, Evaluation performance of Algorithm and Prediction CKD or not CKD.

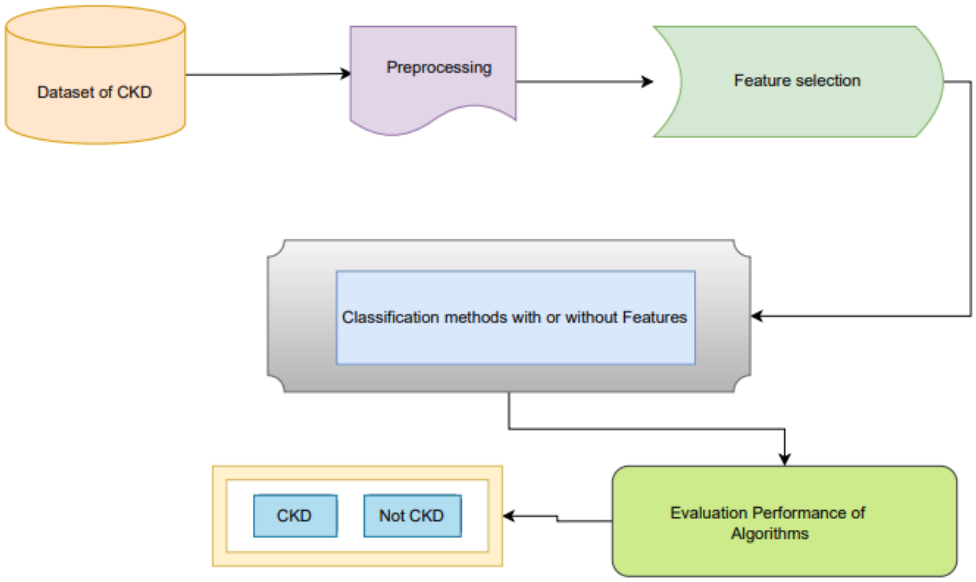


Fig.1 General Architecture Diagram

The dataset used in various research papers is taken from UCI repository having 25 attributes of 400 patients. The pre-processing of the data is next step which includes handling the missing values and converting nominal values to numerical values. Later on, Feature selection is implemented. Feature selection is something which will help reduce the dimensionality of the dataset. It means that it will reduce the size of dataset by considering only impactful attributes. By using feature selection most of the time the accuracy of the model increases. After feature selection, different types of algorithms are implemented.

LITERATURE SURVEY

When looking at the existing research, it's clear that forecasting CKD has become a top priority for scientists. Looking at the available evidence, it's evident that scientists have made anticipating CKD a high focus. Looking at the facts, it's clear that scientists have placed a high priority on predicting CKD. Under WFS and Boruta, the research paper revealed the best accuracy of 99 percent in ANN. [1]. The impact of employing clinical variables to categorise patients with chronic renal disease using the support vector machines method is examined. Clinical history, physical examinations, and laboratory testing are used to create the chronic kidney disease dataset. [2]. This paper aims to test the ability of machine learning algorithms for the prediction of chronic kidney disease using the smallest subset of features. Several statistical tests have been done to remove redundant features such as the ANOVA test, the Pearson's correlation, and the Cramer's V test. Logistic regression, support vector machines, random forest, and gradient boosting algorithms have been trained and tested using 10-fold cross-validation. This research achieves an accuracy of 99.1 according to F1-measure from Gradient Boosting classifier [3]. K-Means Clustering, DB-Scan, I-Forest, and Autoencoder are five unsupervised methods used in the work. As well as combining multiple them with other feature selection algorithms. Patterns can be extracted from unlabelled CKD-related clinical data using unsupervised learning. Patients can be correctly classified as "CKD" or "Non-CKD" using the derived patterns. K-means performs well with less features. SHAP is one of the feature selection strategies utilized in this paper. It has a 99.9% accuracy rate. It has a 99.2 % f1-score and a 99% accuracy. The strategy proposed can assist clinicians in managing several patients and as well as providing CKD diagnoses more quickly. As an extension of this work, the five phases of chronic kidney disease can be diagnosed in a similar way [4]. Use experiential analysis of ML techniques to classify the kidney patient dataset as CKD or NOT CKD in this study. It mainly focuses on empirical comparisons of seven machine learning algorithms. J48 Decision Tree, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Naive Bayes Tree (NBTree), Logistic Regression (LR), Naive Bayes (NB), and Composite Hypercube on Iterated Random Projection are the seven ML techniques examined (CHIRP). The main focus of this empirical analysis is also to suggest CHIRP for CKD prediction, which has not been used for this purpose in earlier studies. Overall, the results indicate that CHIRP is effective at reducing mistake rates and increasing accuracy. It also has an overall accuracy of 99.75 %. [5]. The framework results shows that better predictions are possible in the early stages of CKD. The accuracy of prediction is used to assess the models' performance. In comparison to Decision trees and Support Vector Machines, the research found that the Random Forest Classifier model better predicts CKD. In this work, a confusion matrix was used to calculate accuracy. The random forest classifier has an overall accuracy of 99.16 %. [6]. Classifier algorithms used in this paper are Artificial neural networks, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2, and random tree. some feature selection techniques are applied on the dataset. The results for each classifier algorithm were calculated using (I) full features, (ii) correlation-based feature selection, and (iii) Wrapper technique feature selection (Iv) least absolute shrinkage and selection operator regression (v) synthetic minority over-sampling approach with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority over-sampling technique with full features [7]. The paper presents a machine learning framework for diagnosing CKD. The CKD data set was obtained from the University of California, Irvine (UCI) machine learning repository, which contains a large number of missing values. To fill up the missing values, KNN imputation was employed, which selects many full samples with the most similar measurements and processes the missing data for each incomplete sample. Patients may miss some measurements for various reasons, hence some missing valued are seen in real-life medical data. After effectively filling out the incomplete data set, six machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbour, naive Bayes classifier and feed forward neural network) were used to establish models. By analysing the misjudgements generated by the established models, we proposed an integrated model that combines logistic regression and random forest by using perceptron. We hypothesised

that this technology may be used to diagnose diseases using more complex clinical data. [8]. The main focus in this paper is on the classification techniques, that is, tree-based decision tree, random forest, and logistic regression has been analysed. Different measure has been used for comparison between algorithms for the dataset collected from standard UCI repository. In this paper pre-process the dataset and then used the filter method of feature selection that is univariate selection and correlation matrix along with feature importance to find best features from the dataset [9]. In this research paper, a Data Mining algorithm, Boruta analysis is performed to extrapolate the factors which can fortify the chances of a patient having CKD. This analysis covers statistic data along with historic and medical details. The dataset has been obtained from UCI source which contains data of 400 samples from the southern part of India with their ages ranging between 2-90 years [10]. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11-machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. The research also considers the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction [11]. In this paper, four ensemble algorithms are used to diagnose the patient with Chronic Kidney Disease at the earlier stages. The machine learning models are evaluated based on seven performance metrics including Accuracy, Sensitivity, Specificity, F1-Score, and Mathew Correlation Coefficient. Based on the evaluation the AdaBoost and Random Forest performed the best in terms of accuracy, precision, Sensitivity compared to Gradient Boosting and Bagging. The AdaBoost and Random Forest also showed the Mathew Correlation Coefficient and Area Under the curve scores of 100%. The machine learning model proposed in this paper will provide an efficient way to prevent Chronic Kidney diseases by enabling the medical practitioners to diagnose the disease at an early stage [12].

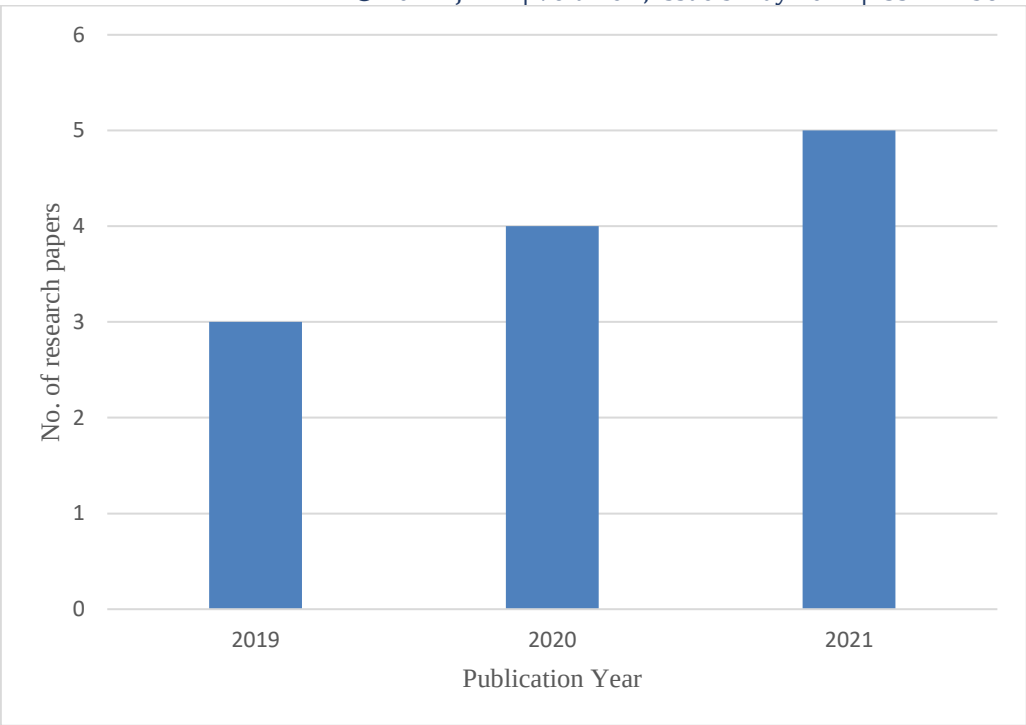
COMPARATIVE ANALYSIS

The below table represents the analysis of 12 research papers. The research papers were taken in between year 2019 to 2021. Various algorithms and deep learning models and feature selection were used. The highest accuracy of different research papers is listed below. The highest accuracy obtained from the survey paper is 100%.

Ref. no.	Year	Algorithm/Deep Learning models	Feature Selection	Highest Accuracy
1	2021	ANN, LSTM, GRU, Bi-Directional GRU, Bi-Directional LSTM, Multi-Layer Perceptron and simple RNN	WFS, CFS, RFE, Lasso and Boruta	97.00%
2	2021	Random Forest, Gradient Boosting, Extreme Gradient Boosting, Logistic Regression, Support Vector Machine	SHAP	99.50%
3	2019	Random Forest, Gradient Boosting, Support Vector Machine, Logistic Regression	Pearson's correlation, Cramer's V, and ANOVA	99.00%
4	2021	Logistic Regression, Random Forest, K means clustering, DB-Scan, Isolation Forest and Autoencoder	Pearson, Chi-2, RFE, SHAP	99.00%
5	2020	NB, LR, SVM, NBTree, J48, CHIRP, MLP	WFS	99.75%
6	2019	Decision tree, SVM, Random Forest	WFS	99.16%
7	2021	C5.0, LR, LSVM, CHAID, KNN, RT, ANN	WFS, CFS, Lasso, Wrapper Forward	98.86%
8	2020	LR, RF, SVM, KNN, NB, FFNN	WFS, CFS	99.75%
9	2020	DT, RF, LR	ANOVA, Pearson's correlation	99.24%
10	2019	Random Forest	WFS, Boruta	100%
11	2020	DT, RF, XGB, Extra Trees, Ada Boost, KNN, SVC Linear, LR, SVC RBF, NB, CNN	Correlation Matrix	100%
12	2021	Bagging, AdaBoost, Gradient Boosting, Random Forest	WFS	100

Table 1: Analysis chart based on prediction and accuracy

This study includes the survey on the research papers between the year 2019 to 2021. As per the survey, 3 research papers are from 2019, 4 are from 2020 and 5 are from 2021.



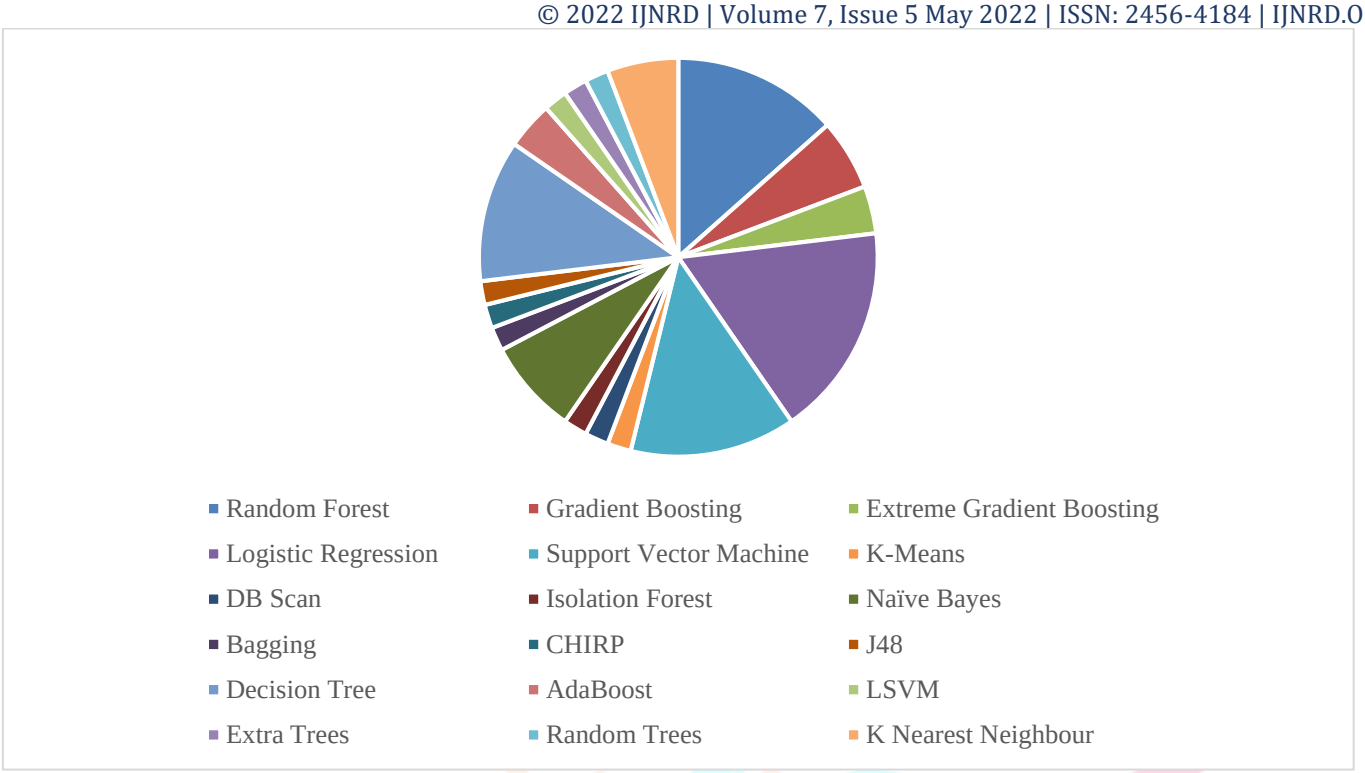
Graph 1: Analysis on the basis of publication year

Below table represents different feature selection algorithms implemented in different research papers. According to the survey research, mostly used feature selection is WFS.

Feature Selection	**Number of Research Paper**
WFS	[1,5,6,7,8,12]
CFS	[1,7,8]
RFE	[1,4]
Lasso	[1]
Boruta	[1,7]
SHAP	[2,4]
Pearson's Correlation	[3,4,9,10]
Cramer's V	[3]
ANOVA	[3,9,10]
Variance Thresholding	[10]
Chi-2	[4]
Wrapper Forward	[7]
Correlation Matrix	[11]

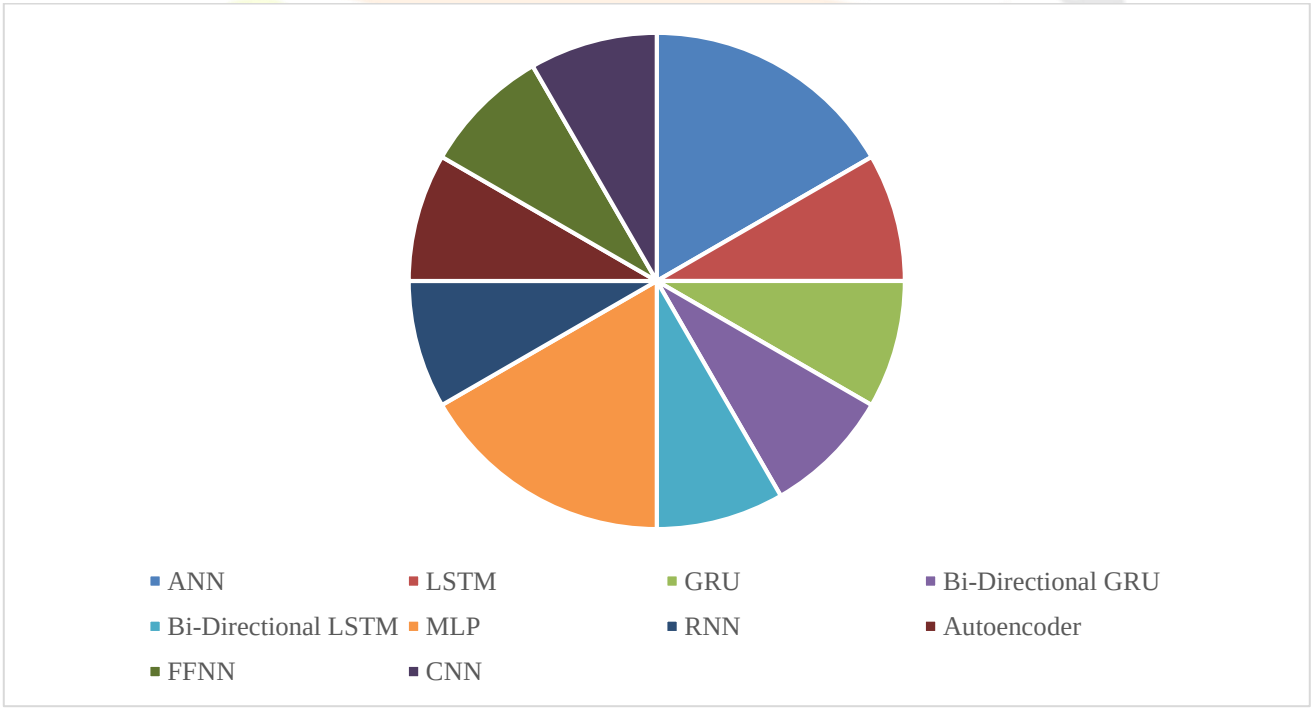
Table 2: Analysis on the basis of Feature Selection

The pie chart below shows the implementation of different algorithms in different research papers. Logistic Regression, Random Forest, Support Vector Machine and Decision Tree were the most used algorithm according to the survey.



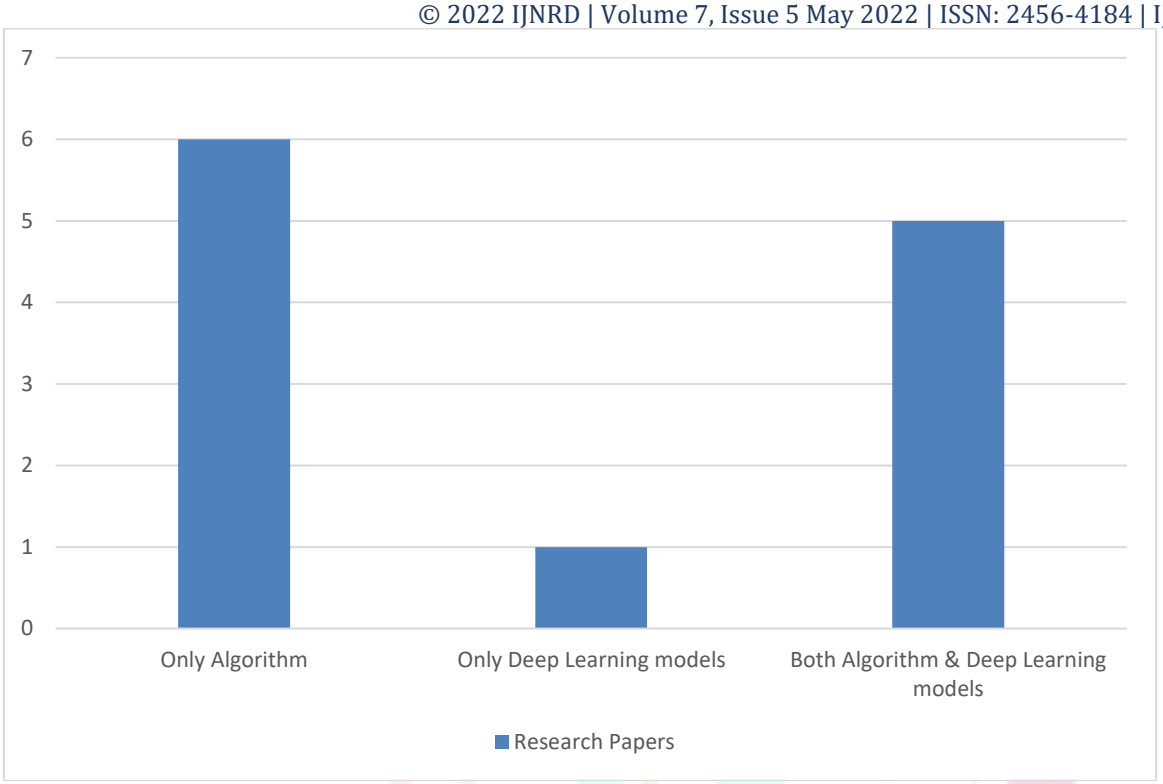
Pie Chart 1: Analysis on the basis of algorithms used

The below pie chart shows the use of 10 feature selection that have been implemented in the studied research papers.As per the survey, ANN and MLP are the most used feature selection.



Pie Chart 2: Analysis on the basis of Deep Learning models used

From the below graph, it can be noticed that 6 research papers have used only algorithms, 1 have used only deep learning models and 5 have used both.



Graph 2: Analysis on the basis of Algorithm and Deep Learning models used:

CONCLUSION

The focus of this research was to analyse results from various methods employed in various studies using a standard dataset from the UCI repository. A survey of various machine learning techniques for detecting chronic kidney disease has been provided in this paper.

Algorithmic classifiers can predict the severity of chronic kidney disease. Using 12 research publications, the goal of this survey is to classify existing strategies in terms of publication year, methodologies employed, dataset used, and performance indicators. SVM, LR, RF, ANN, DT, GB, and other algorithms are commonly used to predict chronic kidney disease.

A total of 24 traits and one label variable are used in this investigation. In the future, these categorization systems could be modified to include symptoms as an input, allowing them to anticipate disease based on unusual case records. In the medical industry, we may improve these strategies even more by incorporating a query model into the programme, which allows the doctor and the app's administrator to communicate with one another. Various strategies can be used to improve results using different feature selection methods.

REFERENCES

[1] Shamima Akter, Ahsan Habib, MD. Ashiqul Islam, MD. Sagar Hossen, Wasik Ahmmed Fahim, Puza Rani Sarkar and Manik Ahmed," Comprehensive Performance Assessment of Deep Learning Models in Early Prediction and Risk Identification of Chronic Kidney Disease" 2021 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2021, doi: 10.1109/ACCESS.2021.312949.

[2] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-4, doi: 10.1109/ICAICT.2018.8747140.

[3] Marwa Almasoud, Tomas E Ward, "Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors", international Journal of Advanced Computer Science and Applications (IJACSA), Volume 10 Issue 8, 2019, doi:10.14569/IJACSA.2019.0100813

[4] Linta Antony, Sami Azam, Eva Ignatious, Ryana Quadir, Abhijith Reddy Beeravolu, Mirjam Jonkman and Friso De Boer, "A Comprehensive Unsupervised Framework for Chronic Kidney Disease Prediction", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), doi: 10.1109/ICCCNT49239.2020.9225548.

[5] Bilal Khan, Rashid Naseem, Fazal Muhammad, Ghulam Abbas, Sunghwan Kim," An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy", doi: 10.1109/ACCESS.2020.2981689

[6] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, " Chronic Kidney Disease Prediction using Machine Learning Models ", International Journal of Engineering and Advanced Technology (IJEAT), doi: 10.35940/ijeat.A2213.109119.

[7] Prediction of Chronic Kidney Disease - A Machine Learning Perspective:= P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763

[8] A Machine Learning Methodology for Diagnosing Chronic Kidney Disease:- J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," in IEEE Access, vol. 8, pp. 20991-21002, 2020, doi: 10.1109/ACCESS.2019.2963053

[9] Performance Analysis of Machine Learning Classifier for Predicting Chronic Kidney Disease:- 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), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154147

[10] Early Detection and Prevention of Chronic Kidney Disease: - M. Desai, "Early Detection and Prevention of Chronic Kidney Disease," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2019, pp. 1-5, doi: 10.1109/ICCUBEA47591.2019.9128424

[11] I. U. Ekanayake and D. Herath, "Chronic Kidney Disease Prediction Using Machine Learning Methods," 2020 Moratuwa Engineering Research Conference (MERcon), 2020, pp. 260-265, doi: 10.1109/MERCon50084.2020.9185249

[12] Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm: Nikhila, "Chronic Kidney Disease Prediction using Machine Learning Ensemble Algorithm," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 476-480, doi: 10.1109/ICCCIS51004.2021.9397144 **[9]** S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, " Chronic Kidney Disease Prediction using Machine Learning Models ", International Journal of Engineering and Advanced Technology (IJEAT), doi: 10.35940/ijeat.A2213.109119.