IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 5, May 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

# Efficient detection of Congestive Heart Failure Detection through Machine Learning

**Ajit Gedam[1], Sarthak Baravkar[2] , Shreyash Undre[3] , Atharv Salunke[4] , Sahil Shaikh[5]**

Lecturer, Dept. of Computer Engineering, A.I.S.S.M.S. Polytechnic Pune, Maharashtra, India[1]

Student, Dept. of Computer Engineering, A.I.S.S.M.S. Polytechnic Pune, Maharashtra, India[2, 3, 4, 5]

**ABSTRACT:** The diminution in the global death rate has resulted in an upsurge in the number of elderly people. These people are quite elderly and are more susceptible toward certain illnesses than their younger colleagues. This has become one of the most serious medical issues that has put a pressure on the health-care system in recent decades. This is a troubling phenomenon that has resulted in an increase in heart failure-related deaths. Due to the fundamental complexity of these diseases, which necessitates comprehensive testing and diagnostics, heart failure is particularly difficult to identify. The delay in receiving the findings causes a delay in delivering early treatment for patients, which is critical in the case of heart failure, and delaying to do so might end in death. Therefore, this research paperdeploys machine learning techniques on a dataset consisting ofECG parameters such as QR intervals, T wave axis, etc. to perform predictions for congestive heart failure. The proposed methodology deploys K-Means clustering for the clustering and Linear Regression for regression analysis along with Artificial Neural networks and a novel amalgamation of Decision tree and Fuzzy techniques. The Extensive experimental evaluations have been performed which have been effective in getting a highly satisfactory accuracy of 79.10 percent for congestive heart failure prediction.

**KEYWORDS**: Congestive Heart Disease, K-means clustering, Artificial Neural Network, Fuzzy Decision Tree.

## I. INTRODUCTION

Heart disease has been one of the leading contributors of mortality in recent decades. This has gotten a lot of press throughout the globe. The surge in the number of older persons might be related to a dramatic increase in the proportion of people suffering from heart disease. As individuals become older, their risks of having a heart attack or suffering from heart failure increase considerably. Cardiovascular disease and heart failure are more common in the elderly since they are feeble and have a weaker blood circulation system. As a consequence, the overwhelming majority of patients with heart disease are seniors who require extra care and protection.

As a result, a patient's heart failure status must be predicted in order to ensure that the patients are receiving prompt therapy. This is because cardiac arrest is a particularly time-sensitive sickness, and in such a circumstance, every second counts. As a response, it is critical that a prediction strategy can foresee the real-world heart attack condition so that it may better prepare for it, perhaps saves a life of a patient.

Forecasting cardiac failure requires a large quantity of data that includes a range of criteria as well as knowledge on the patient's vital stats as well as other heart-related information. The data comprises a large number of characteristics, and the reliability of the forecast decreases as the variety of attributes increases. Because it can produce incredibly exact forecasts, the Machine Learning model is a useful option. This paradigm identifies the likelihood as well as other indicators that may be used to accurately assess the patient's status and create reasonable results.

The heart attack failure predictions offers doctors and experts a significant amount of time that would otherwise have been spent reviewing many studies and the patients' prior conditions, which takes a significant amount of time. Additionally, there are several traits and disorders that cause heart failure, and these qualities may have a substantial influence on the results. A wide range of traits adds to the complexity of the forecast, imposing even more load on the patient's care provider. Due to the rapid growth in complexity and the rising strain on medical professionals, the development of autonomous prediction via the use of Machine Learning is a very possible and practical resolution to the issue.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for future enhancement.

## II. LITERATURE REVIEW

M. Wang et al. [1] present a UCO data processing algorithm that combines three methods: under-sampling, clustering, and oversampling. The algorithm can deal with skewed data from stroke patients. From the original data, eight medical indicators related to heart attacks were chosen. The prediction of heart attacks by distinct machine learning models is compared. Experiment results on the MIMIC-III database of stroke patients' datasets show that RF is the finest technique for predicting the possibility of a heart attack. Precision is 70.05 %, while accuracy is 70.29 %. Using the UCO (120) +RF method described in this study, the authors use clinical monitoring data from the ICU to predict whether stroke patients will have a heart attack.

An examination of various machine learning techniques used in the prediction of heart disease is being carried out by D. Swain et. al. They compared the classification accuracy of all discussed techniques when dimensionality reduction and output levels were taken into account. Therefore, it can be concluded that dimensionality reduction is an excellent technique for any type of classification problem, as it reduces the number of input features, which contributes significantly to disease occurrence [2]. However, the problem with dimensionality reduction is that if it is done excessively, the model may become overfit. The problem with an overfitted model is that it becomes overly specific to a specific data set and is unable to function properly when a different set of data is fed into the prediction.

S. Sarmah presents a machine learning technique called a DLMNN-based HD prediction system for forecasting the patient's heart state. The proposed technique imagines patients suffering from HD using Hungarian HD as well as healthcare sensors. Classification algorithms are utilized to categorize patient data to detect HD. The prediction process is divided into two stages: training and testing [3]. During the training phase, the classifier is trained using data from the benchmark dataset. Actual patient data is used during the testing phase to find out the presence of the ailment. The experiment's outcomes are compared to planned and existing approaches. Three comparisons have been carried out for the indicated approaches utilized in DC, data transfer, and disease prediction. The planned MHA used in DC achieves the highest CR values in the shortest amount of time. PDH-AES technology, which is utilized in secure data transfer, yields the best results, providing the highest level of security (95.87 percent) and the fastest time for encryption and decryption. Finally, the DLMNN for sickness prediction, in conjunction with the f-measure, provides the best level of sensitivity, accuracy, and specificity in disease prediction.

X. Chu et al. analyze AIS data with a big data processing architecture based on Spark, Hadoop, and Mesos, to process a massive amount of AIS data and information [4]. Using K-means cluster analysis, different ship trajectory points are classified, and the analysis efficiency of the big data platform established in this research and the computer without a big data platform is compared and evaluated. The test findings show that the platform can greatly enhance the efficiency of AIS data analysis. In an environment where the number of AIS data is expanding, the platform proposed in this study can improve data processing efficiency, perform effective AIS data cluster analysis, and lay the basis for deeper AIS data analysis.

R. Ge et al. present a novel multi-label neural network approach (ML-NN) that combines neural networks with multi-label learning technology. Because the neural network is a multiple-input multiple-output system, it is ideal for multi-label learning problems [5]. The model is highly irresponsible and can search for optimal answers at breakneck speed. In this research, the knowledge foundation of multi-source data is unclear, as are the reasoning principles. Self-learning, fault tolerance, and associative memory are all functions of the neural network. It applies no limits to the input or residual distributions and can learn and build nonlinear intricate relationships from this chronic disease dataset. The activation function based on the cross-entropy loss function and the backward propagation technique has two distinct advantages: it captures the non-linear correlation between the inputs. In the meantime, a thorough comparative analysis demonstrated that our solution beat many traditional multi-label learning methods. Experiments demonstrated that the proposed technique might enhance prediction accuracy and operational efficiency, supporting clinicians in the accurate diagnosis and treatment of chronic disease patients.

G. R. Vasquez-Morales et al. developed a neural network technology that can predict the likelihood of developing the chronic renal disease with 95% accuracy. Given that the most comparable research achieves an accuracy of 89.7 %, this is an impressive result. This result was obtained by training a neuronal network with five layers: an input layer with

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

7,492 neurons representing the model's variables or characteristics, three hidden layers with 500, 100, and 50 neurons, respectively, and an output layer with a single neuron representing the class of the binary classification problem. A ReLU activation function was utilized in the hidden layers to discover the probability, and a sigmoid function was employed in the output layer to discover the probability [6]. Adam was employed as the model's training algorithm, and it demonstrated a faster convergence rate than other more standard techniques such as gradient descent and stochastic gradient descent (SGD). To reduce the impact of overfitting in the network, regularization using the dropout approach in conjunction with early halting was used. In contrast to the neural network, other machine learning models, such as random forest and vector support machines (SVM), were trained, yielding accuracy ratings of 92 % and 61 %, respectively.

In the suggested framework, S. Y. Yashfi et al. analyzed UCI datasets and real-time datasets to identify CKD in patients. They worked with missing data, trained it, and built a Random Forest and ANN model. The Python programming language was used to create these two algorithms. The Random Forest approach has an accuracy of 97.12 % and the ANN algorithm has an accuracy of 94.5 %, both of which are pretty good. They will be able to predict the danger of CKD at an early stage using the technique they have developed. It is challenging to obtain real-time datasets [7]. Several difficulties developed during the real-time data collection process. Following data collection, another issue addressed was the processing of missing values. The concept is the driving force behind progress and achievement. They also have other suggestions that will make their intended job more efficient. People with low to high incomes now own cell phones. So, as part of their planned endeavor, they hope to develop a mobile application. Therefore, users will be able to utilize the application, making it uncomplicated for them to diagnose CKD.

## III. PROPOSED METHODOLOGY

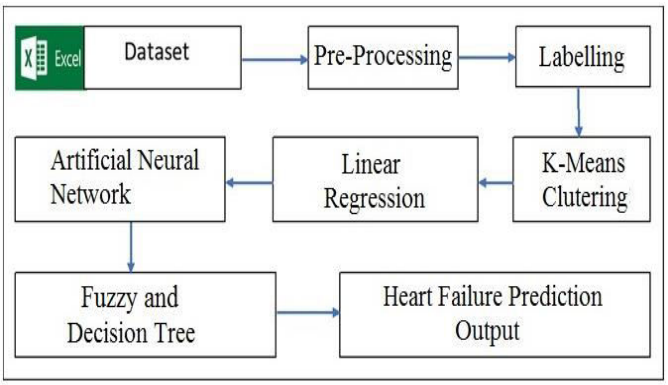



Figure 1: Proposed model System Overview

The suggested strategy for detecting congestive heart failure through machine learning methodologies is shown in Figure 1. The proposed framework concept entails conducting a sequence of processes in order to obtain heart failure prediction outcomes, which are detailed in the steps below.

*Step1: Dataset collection, preprocessing and Labeling*- The proposed approach requires an extensive dataset that is downloaded from the URL – https://figshare.com/s/3ea30ba1ac39af387358 in a csv format.

This data is then transformed to a Microsoft excel spreadsheet and sent into the system. This spreadsheet is loaded into the double dimension list after the information is supplied into the system. However a fraction of the dataset characteristics are crucial in the diagnosis of heart failure situations such asQT, QRS, PR, and RR. As a result, just those attributes are gathered in a distinct list after the preprocessing stage, which is called to as the preprocessed list.

*Step 2: K-Means Clustering*- The preprocessed and labelled list is provided as an input to the step of the procedure for the purpose of clustering. The clustering in this approach is performed using the k-means clustering technique for grouping together similar data. The k-means approach utilizes the distances between the data to form a semantic group based on the closest data points. This process segregates the data effectively which can be utilized for the prediction of heart failure accurately. The k-means technique is performed in a step-by-step manner as given below.

*Distance Evaluation* –The distance between the data points need to be evaluated to achieve the differences between their semantic closeness. The distance in this implementation of key means clustering utilizes Euclidean

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

distance. Attributes in the preprocessed and labelled list are utilized for achieving the distances with all the other rows in the list. These distances are added at the end of the respective row which are then utilized to achieve the row distance stored as $R_D$. All of these row distances for all the rows are then utilized for extracting the average row distance stored as $A_{VG}$. Equation utilized for equality and distance is given below in equation 1.

$$ED = \sqrt{\sum(ATi - ATj)^2} \_\_\_\_\_ (1)$$

Where,
ED=Euclidian Distance
$A_{Ti}$=Attribute at index i
$A_{Tj}$= Attribute at index j

*Centroid Estimation*-The row distances achieved previously in the previous step are subjected to the bubble sort technique for sorting the list into ascending order of the row distances. From the sorted list the data points are collected using random indices. The respective row distance and their Euclidean distance of the selected data points are the utilized for achieving the boundary values.

The boundaries are calculated through the addition and subtraction of the average row distance from the row distance of the selected data point. A double dimensional list is then formed to store the generated boundaries which will be used for cluster formation subsequently.

*Cluster Formation*-The boundary values achieved in the previous step are utilized to generate the clusters. This boundary values are employed to perform segregation based on the selected attributes. The resultant clusters are then provided to the next step for regression analysis using linear regression.

*Step 3: Linear Regression* –The linear regression is performed on the attributes that have been selected previously, QT, QRS, PR, and RR. These attributes are utilized for the creation of a frequent itemset. This frequent itemset it is generated through the indices of the selected attributes by multiplying them and storing them in a list. The linear regression is estimated for each of the attributes with respect to the other attributes in the list.

An array is created for each of the clusters and the selected attributes in the form of an array called x[] and y[] reliant on the indices of the frequent item list. Linear regression is performed on these arrays and the values of m and b are evaluated as depicted in the equation 6. The intercept value is there measure through the deployment of equation 3 and 4 and the entire process is described in the algorithm 1.

$$Y = M\,x + B \_\_\_\_\_ (2)$$

$$M = \frac{N\sum(xy) - \sum x \sum y}{N\sum(x^2) - (\sum x)^2} \_\_\_\_\_ (3)$$

$$B = \frac{\sum y - M\sum x}{N} \_\_\_\_\_ (4)$$

Where:
x = how far up (Array of Attribute 1)
y = how far along (Array of Attribute 2)
M = Slope or Gradient (how steep the line is)
B = the Y Intercept (where the line crosses the Y axis)
N= Size of the array
Y=Intercept value

The intercept values are collected in a list for each of the attributes of all the clusters and sort them in an ascending order. This is done to extract the lowest off the intercept values for each of the attributes and the clusters. This is useful in achieving the distance between the achieved values and the standard values generated in the first step of this regression section. The distance extracted from this process is directly proportional to the risk of heart failure.

Respective weights to each of the clusters lowest attributes with the lowest distance are assigned. Then based on these rates the list is sorted in the ascending order of the weights. The higher the weights higher is the data regarding

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

heart failure for the particular cluster. ANN input list is created by selecting k number of clusters on the cluster list which is sorted in the descending order.

The sorted cluster indices help to select top K number of clusters, which are then segregated in a single list to call it as an ANN input list to feed to the next step of ANN.

*Step 4: Neural Network* –This step of the procedure deploys the artificial neural network that takes the ANN input list to this module generated previously. The process of anal assesses the output layer and hidden layer values through instance learning for the heart failure prediction. The estimation of the output and hidden layer requires the assignment of 16 random weights including two bias values that are required to balance the random weights.

The target values are assigned as the initial values in the attributes and the tanh activation function is deployed to achieve the output and hidden layer values through the equations 5 and 6 given below.

$$T=\left(\sum_{k=0}^{n} AT * W\right) + B \quad \text{____(5)}$$

$$H_{LV}= 2\left(\frac{1}{(1+\exp(-T))}\ X\ 2T\right) - 1 \quad \text{_____ (6)}$$

Where,
n- Number of attributes
$A_T$- Attribute Values
W- Random Weight
B- Bias Weight
$H_{LV}$ – Hidden Layer Value

The target values and the achieved output layer values are aggregated together to form the neural network probability list. The achieved neural network problem it is provided as an input to the next step for the segregation using fuzzy decision tree.

*Step 5: Fuzzy Decision Tree* –This is the final stage of the presented methodology which is utilized for the prediction of the probability of heart failure. The network probability list generated in the previous step is utilized to extract the largest and the smallest probability values of each of the attributes such as QT, QRS, PR, and RR. The achieved smallest and largest values are then utilized for the generation of fuzzy crisp values such as Very Low, Low, Medium, High and Very High. The algorithm 2 below depicts the weight assignment of the fuzzy crisp values through the utilization of decision tree.

ALGORITHM 2: Fuzzy crisp weight assignment through Decision Tree

//Input : Fuzzy Crisp set $FC_S$
//Input : Neuron Probability list $NP_L$
//Output: Fuzzy Weight List $FW_L$
fuzzyWeightEstimation($FC_S$, $NP_L$)
1: **Start**
2: $FW_L$ =∅
3: ***for*** i=0 to size of $FC_S$
4: TMP = $FC_{S\,[i]}$
5: MIN=$TMP_{[0]}$
6: MAX=$TMP_{[1]}$
7: ***for*** j=0 to size of $NP_L$
8: $R_L$= $NP_{L\,[i]}$
9: VAL=$R_{L[INDEX]}$
10: ***if*** (VAL >=MIN AND VAL<=MAX)
11:$R_{L[i+1]}$=i
12: ***end if***
13: $FW_L$= $FW_L$+$R_L$
14: ***end for***
15: ***end for***

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

16: return $FW_L$
17: **Stop**

The fuzzy crisp ranges are then utilized for the assignment of weights to the rows in the neural network probability list. The assign weights are then added together for sorting the list in descending order. This sorted list is utilized for extracting the top 20 entries as the most accurate heart failure prediction data.

## IV. RESULT AND DISCUSIONS

The presented approach for the prediction of congestive heart failure is developed on a laptop running windows operating system. This machine is facilitated with an Intel core i5 processor paired with 6 GB of Ram. An additional 500 GB of Hard drive storage is also present. The python programming language is the language of choice for the purpose of development of this system. The Spyder IDE is being utilized for the purpose of coding this methodology. The MySQL database server fulfills the database responsibilities.

The data set utilized for giving as an input to the system is extracted from the URL. https://figshare.com/s/3ea30ba1ac39af387358.

There are a number of attributes in the data set which is mainly populated with ECG readings along with attributes such as person ID, ECG Date, ECG dept., ECG source, RR (time taken between each beat), PR(the ventricle delay),QRS(ventricle depolarization), QT(ventricular depolarization in total) and other parameters are like P_wave_axis, T_wave_axisQtc.

The prediction of outcomes generated by the proposed methodology should be effectively scrutinized for the purpose of analyzing its accuracy. The accuracy of the prediction generated by the presented methodology has been analyzed through the use of parameters such as true positive, true negative, and false positive.

True positive is the number of accurately identified positive outcomes. True negative is the analysis of the outcomes wherein the model precisely identifies the negative outcome. False positive is the number of predictions that are performed incorrectly for the positive outcome. Whereas false negative are the imprecise predictions of the negative class which are non-existent in our model.

The equations 7 and 8 given below are utilized for the evaluation of the accuracy.

$$\text{Accuracy} = \frac{\text{NTP+NTN}}{\text{NTP+NTN+NFP+NFN}} \quad (7)$$

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions NTP}}{\text{Total Number of Predicitons (NTP+NFP)}} \quad (8)$$

The equation number 8 is a modification of equation 7 wherein the false negatives are removed due to the fact that it is non-existent in our proposed methodology. The proposed methodology is subjected to extensive evaluation and the outcomes are converted into respective readings of the parameters discussed above such as true positive, true negative, false positive, and false negative which are tabulated in the table 1 with its graphical representation in figure 2.

Table 1: Performance of the System reading for the accuracy

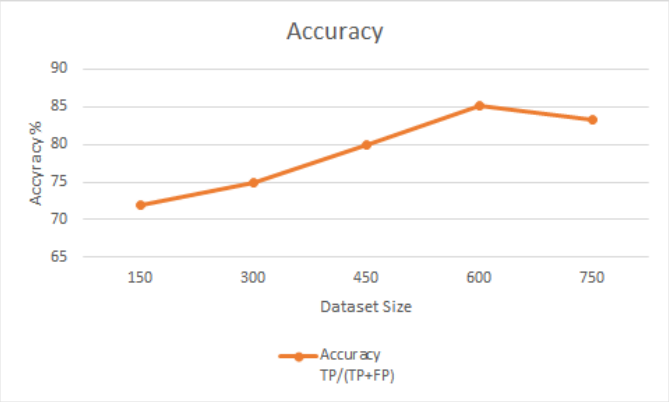
| Dataset Size | Total Outcome | True Positive | True Negative | False Positive | False Negative | Accuracy TP/(TP+FP) |
|---|---|---|---|---|---|---|
| 150 | 24 | 18 | 134 | 7 | 0 | 72 |
| 300 | 23 | 18 | 284 | 6 | 0 | 75 |
| 450 | 24 | 20 | 432 | 5 | 0 | 80 |
| 600 | 26 | 23 | 579 | 4 | 0 | 85.18518519 |
| 750 | 23 | 20 | 732 | 4 | 0 | 83.33333333 |

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 5, May 2022||

| DOI:10.15680/IJIRSET.2022.1105136|

Figure 2: Accuracy Plot of the proposed model for Heart Failure Prediction




The outcomes in the figure 2 and the table one given above depict that the presented approach in this research paper achieves an accuracy of 79.10 percent which is a highly satisfactory result.

## V. CONCLUSION AND FUTURE SCOPE

The prediction of congestive heart failure has been performed by the methodology presented in this research paper. The methodology executes through the initiation of the model by supply a data set containing ECG values of the patients as an input. The attributes required for performing the prediction are effectively extracted from the data set and are subsequently preprocessed and labelled for the purpose of improving the execution accuracy.The preprocessing labelled data is utilized for the clustering purpose which is achieved through k-means clustering technique. The achieve clusters are then provided to linear regression for the purpose of achieving regression list. This regression list is further provided as an input to the artificial neural networks that use this list for the hidden layer and output layer evaluation which results in precise predictions. The generated predictions need to be evaluated and segregated which is done by a novel technique that amalgamates decision tree and Fuzzy classification. Extensive experimental evaluations have been performed which have been effective in getting and accuracy of 79.10 percent.

For the purpose of future research directions this technique can be improved and implemented in real time for faster congestive heart failure predictions.

## REFERENCES

[1] M. Wang, X. Yao and Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," in IEEE Access, vol. 9, pp. 25394-25404, 2021, DOI: 10.1109/ACCESS.2021.3057693.
[2] D. Swain, S. K. Pani, and D. Swain, "A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, pp. 1-6, DOI: 10.1109/ICACAT.2018.8933603.
[3] S. S. Sarmah, "An Efficient IoT-Based Patient Monitoring and Heart Disease Prediction System Using Deep Learning Modified Neural Network," in IEEE Access, vol. 8, pp. 135784-135797, 2020, DOI: 10.1109/ACCESS.2020.3007561.
[4] X. Chu, J. Lei, X. Liu, and Z. Wang, "KMEANS Algorithm Clustering for Massive AIS Data Based on the Spark Platform," 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), 2020, pp. 36-39, DOI: 10.1109/CRC51253.2020.9253451.
[5] R. Ge, R. Zhang and P. Wang, "Prediction of Chronic Diseases with Multi-Label Neural Network," in IEEE Access, vol. 8, pp. 138210-138216, 2020, DOI: 10.1109/ACCESS.2020.3011374.
[6] G. R. Vasquez-Morales, S. M. Martínez-Monterrubio, P. Moreno-Ger, and J. A. Recio-García, "Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning," in IEEE Access, vol. 7, pp. 152900-152910, 2019, DOI: 10.1109/ACCESS.2019.2948430.
[7] S. Y. Yashfi et al., "Risk Prediction of Chronic Kidney Disease Using Machine Learning Algorithms," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5, DOI: 10.1109/ICCCNT49239.2020.9225548.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details