IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125125

LIVER DISORDER DIAGNOSIS USING MACHINE LEARNING - A COMPARATIVE STUDY

Anusuya.R[1], Dr. S. Roselin Mary, Ph.D[2]

Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[1]

Head of Department, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[2]

Abstract: It is critical to diagnose liver illness early on in order to receive the best therapy possible. Machine learning algorithms is growing rapidly such as SVM, K-mean clustering, KNN, Random Forest, Logistic regression, and others. The input is usually numerical data of various factors, and the output findings are obtained in real-time, predicting whether or not the patient has a liver problem. In this project, used a variety of supervised machine-learning methods before deciding which one is best for the model. Existing systems rely on classical deep learning models, which are inefficient and imprecise. They aren't precise enough. This proposing model is to use classification algorithms to identify liver patients from healthy individuals. Here, we choose the algorithm in this module that serves as the best fit. The dataset is taken from the Kaggle dataset. The advantages of the proposed model are that it shows high accuracy, is fast processing and is highly scalable. With the effective use of the presented model, practitioners can make intelligent clinical decisions.

Keywords: Bivariate analysis, correlating columns, MSE Loss Function, Removing Null values, Replacing Non-acceptable zero values, Univariate analysis

I. INTRODUCTION

The liver, which resembles a football and is located on the right side of our stomach, is the most visible inner organ of the human body. It is involved in the removal of hazardous compounds, the synthesis of proteins, the generation of bile, the manufacture of albumin, the metabolic activity of bilirubin, carbohydrates, and fat, and the production of many chemicals necessary for the proper digestion of our food. It is the only visceral organ in vertebrates that can regenerate through a specific procedure. If at least 25% of the tissue survives in the human body, the rebuilding process takes 8 to 15 days to complete without any performance degradation. The liver is one of the most complicated components of the body in terms of functionality. As a result, the liver should be preserved. As a result, maintaining the liver's health is critical, as malfunctioning livers can lead to fatal disorders like fascioliasis, cirrhosis, hepatitis, fatty liver disease, liver cancer, liver failure, Gilbert's syndrome, ascites, hemochromatosis, and more.

Liver diseases can be caused by a variety of factors, including

1. Inheritance from family members,
2. Infection with viruses and parasites,
3. Excessive alcohol consumption,
4. Obesity and flabbiness,
5. Diabetes type 2,
6. Tattoo design,
7. Injecting narcotics and blood with shared needles, etc.

Jaundice, dark urine color, abdominal pain, itchy skin, exhaustion, and loss of appetite are all signs and symptoms of liver disease. According to statistics, around 50 million people, or 4.5 percent to 9.5 percent of the overall population, suffer from liver problems. Every year, 2 million individuals die from liver diseases around the world. The best way to get rid of this disease is to receive a correct diagnosis and treatment as soon as possible. This disease's diagnosis could involve one or more of the following: blood tests, MRI, CT scans, and Ultrasounds. Doctors or practitioners determine whether or not a person is affected after observing all of these. However, detecting liver problems is always a difficult task that necessitates the use of skilled specialists and takes a long time. Machine learning plays a significant part in disease diagnosis and treatment to aid healthcare practitioners. It can extract useful data from medical databases and create a model to identify patients.

DOI: 10.17148/IJARCCE.2023.125125

Deep learning and data mining techniques have been used in numerous studies to detect people with liver problems. However, their prediction accuracy is insufficient due to the complex structure and non-linear properties of medical datasets. It has a large number of missing values and outliers, making prediction challenging.

II. RELATED WORK

"Machine learning techniques for the diagnosis of liver disease[1] A review" by K. Rajaraman et al. (2020) - This review article discusses the use of various machine learning techniques for the diagnosis of liver diseases, such as hepatitis, cirrhosis, and hepatocellular carcinoma. The authors highlight the potential of machine learning in improving the accuracy of diagnosis. [2] "Prediction of liver disease using ensemble of machine learning algorithms" by R. Gopalan and P. Ramakrishnan (2020) - This study uses an ensemble of machine learning algorithms, such as decision tree, support vector machine, and random forest, to predict liver disease. The authors achieved an accuracy of 87.5% using the ensemble method. [3] "A machine learning approach for predicting liver fibrosis in patients with chronic hepatitis B" by K. Zhang et al. (2020) - This study uses machine learning algorithms, such as logistic regression and gradient boosting, to predict liver fibrosis in patients with chronic hepatitis B.

The authors achieved an accuracy of 84.4% using the gradient boosting algorithm. [4] "Prediction of liver disease using machine learning techniques: A comparative study" by S. S. Ali et al. (2019) - This study compares the performance of various machine learning algorithms, such as decision tree, random forest, and k-nearest neighbor, in predicting liver disease. The authors achieved an accuracy of 85.5% using the random forest algorithm. [5] "Deep learning-based classification for non-invasive diagnosis of liver fibrosis stages using ultrasound images" by A. T. Al-kilidar et al. (2020) - This study uses deep learning algorithms to classify liver fibrosis stages using ultrasound images. The authors achieved an accuracy of 92.5% using the convolutional neural network algorithm. [6] "A deep learning framework for liver cancer diagnosis using multiphase CT images" by X. Wang et al. (2021) - This study uses a deep learning framework to diagnose liver cancer using multiphase CT images.

The authors achieved an accuracy of 90.6% using the convolutional neural network algorithm. [7] "Diagnosis of liver disease using ensemble of deep learning classifiers with multimodal features" by P. Shukla et al. (2020) - This study uses an ensemble of deep learning classifiers with multimodal features, such as clinical and radiological data, to diagnose liver disease. The authors achieved an accuracy of 92.3% using the ensemble method. [8] "Liver disease prediction using artificial intelligence techniques: A systematic review and meta-analysis" by A. A. Khan et al. (2021) - This systematic review and meta-analysis evaluates the performance of various artificial intelligence techniques, such as machine learning and deep learning, in predicting liver disease. The authors found that the accuracy of these techniques ranged from 70% to 98%, depending on the algorithm and dataset used.

III. EXISTING SYSTEM

The existing systems are simple and effective but are extremely vulnerable to impact. Moreover, state-of-the-art methods can detect diseases pertaining to certain organs only while some severe conditions may go completely undetected. This could lead practitioners to false assumptions and improper diagnosis and treatments provided to patients The existing systems are also not robust in predicting disease, as the characteristics of the diseases vary and datasets keep evolving with time.

The existing models cause unsatisfactory results & excessive medical costs to the customers, which may lead to these diseases going undetected in many patients.In this case, this will make the diagnosis of liver diseases to be more effective and efficient by preventing misdiagnosis of the liver disorder.The discussed limitations have been overcome to enhance the performance of disease prediction models successfully in the presented application.Developing a system with better performance than the previous works will help in preventing misdiagnosis of the disease and help in providing the best and required medication for the patient.

IV. PROPOSED SYSTEM

In the proposed system, a machine learning algorithm are proposed for detecting liver disorder. Hence, the classifiers in machine learning algorithm Logistic Regression, KNN, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes will undergo training and testing for predicting the liver diseases disorder.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125125

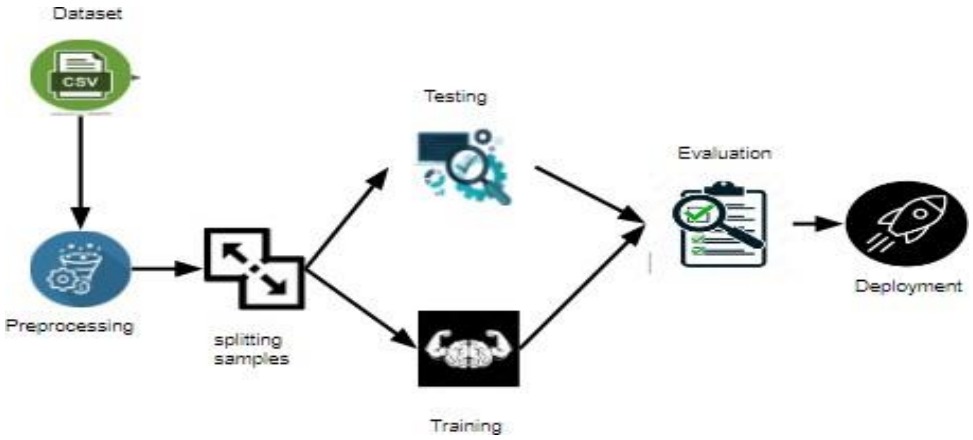

FIG 1 SYSTEM ARCHITEXTURE DIAGRAM

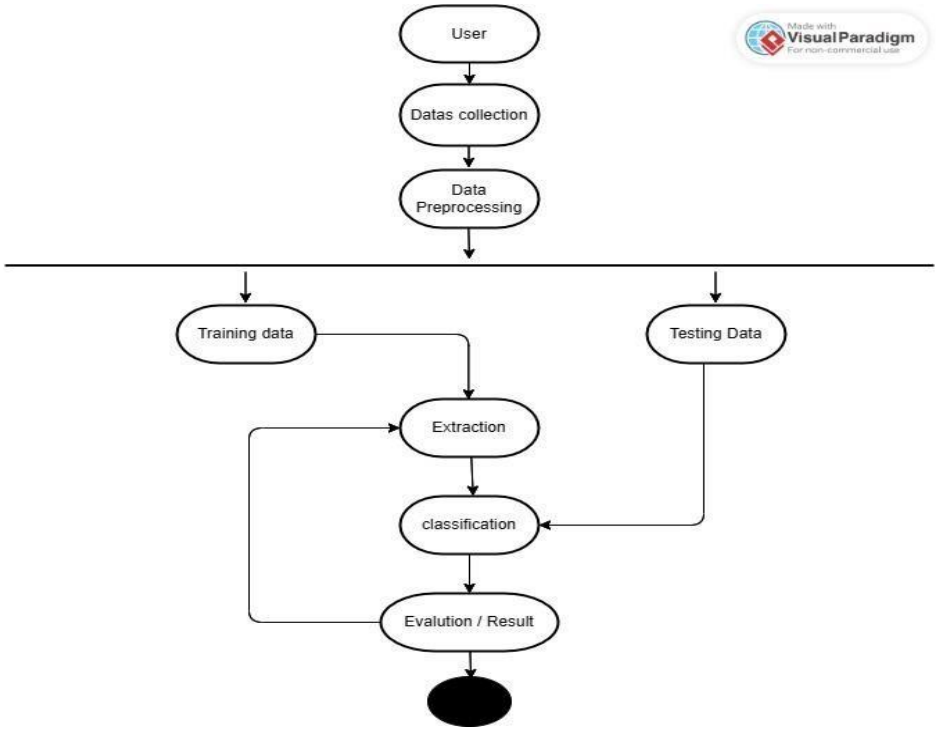

FIG 2 FLOWCHART DIAGRAM

IV. IMPLEMENTATION

A. MACHINE LEARNING

Machine learning (ML) is an area of study focused on comprehending and developing "learning" methods, or methods that use data to enhance performance on a particular set of tasks. It is assumed to be a component of artificial intelligence. Without being expressly programmed to do so, machine learning algorithms create a model from sample data, also referred to as training data, in order to make predictions or decisions. Machine learning algorithms are used in a broad range of applications, including computer vision, speech recognition, email filtering, medicine, and agriculture, where it is challenging or impractical to create conventional algorithms that can perform the required tasks. Computational statistics, which centers on using computers to make predictions, and a subset of machine learning are closely related.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125125

B. DATASET ANALYSIS

The required csv file which contains Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin Albumin and Globulin Ratio, and Dataset is download from Kaggle. A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format. Data tables are presented in Comma Delimited, CSV text file format. Although this file format allows for the data table to be easily retrieved into a variety of applications, they are best viewed within one that will allow one to easily manipulate data that is in columnar format.

C. DATASET PRE-PROCESSING

In this module, begin by attempting to transform data from a predetermined form to one that is significantly more desirable and usable. Using, machine learning techniques, mathematical modelling, and statistical knowledge, the entire process can be automated. The output of this entire process can be in any desired form, including graphs, movies, charts, tables, photographs, and many more, depending on the task, are executing and the requirements of the machine.

D. REMOVING NULL VALUES

Now this will remove the null values or drop the null values from the columns depending on how much value is present or how many are there. These elements are represented by NaN (not a number) or None in the dataset. Whatever the reasons, it makes computing cumbersome and inaccurate. As a result, find these missing data and replace them with usable elements. To indicate missing or null values, Pandas handle None and NaN similarly to each other. There are various helpful utilities for finding, getting rid of, and replacing null values in Pandas DataFrame to make this convention easier. In Pandas DataFrame, and use the functions isnull() and not null to check for missing values (). Both functions aid in determining whether or not a value is NaN. To find null values in a series, these functions can also be used with Pandas Series. Used the dropna() function to remove null values from a dataset. This function removes columns and rows of data with null values. This can also use the following methods to remove NaN values from a NumPy array:
Method 1: Use isnan() Method 2: Use isfinite() Method 3: Use logical_not()

D. DROP DUPLICATE:

Dropping duplicates means removing rows from a dataset that are identical or nearly identical to another row based on a specified set of columns. For example, if have a dataset of liver disease parameters with columns for Age, Gender, Total_Bilirubin, Direct_Bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase, Total_Protiens, Albumin, Albumin_and_Globulin_Ratio, Dataset, then might want to drop duplicate rows that represent multiple entries for the same liver disease parameters from the same dataset with the same columns. This ensures that analysis is based on unique liver disease parameters rather than counting the same liver disease parameters multiple times. To drop duplicates in Python, you can use the drop_duplicates() method of a Pandas DataFrame. By default, this method considers all columns in the DataFrame, but you can specify a subset of columns using the subset parameter.

E. REMOVING OUTLIERS:

Removing outliers means removing data points that are significantly different from other data points in a dataset. Outliers can be caused by measurement errors, data entry errors, or represent genuine extreme values that are far from the typical values in the dataset. Outliers can distort statistical analyses and machine learning models, so it is often important to remove them before proceeding with analysis. The approach to removing outliers can vary depending on the specific context and goals of the analysis. One common method is to define a threshold based on the distribution of the data and remove any data points that fall outside that threshold. For example, one approach could be to remove any data point that is more than 3 standard deviations away from the mean. In Python, you can use various libraries and techniques to remove outliers from a dataset, including: Z-score: Calculate the z-score for each data point, and remove any data points with a z-score greater than a threshold (usually 2 or 3 standard deviations away from the mean). Percentile: Remove any data points that fall outside a specified percentile range, such as the 5th to 95th percentile. Interquartile range (IQR): Calculate the IQR for each column of data, and remove any data points that fall outside 1.5 times the IQR below the first quartile or above the third quartile.

F. MODEL TRAINING & TESTING

For training the model, the six machine learning models that are used are Logistic Regression, KNN, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes. These models are trained using set of training data which is from csv file. The data set is split into training and testing for the model. In the dataset, 70% of dataset were used for training the models and the remaining 30% of datasets were used for testing the accuracy of the models. The labels undergo X train and Y train as well as X test and Y test. For training and testing both positive and negative data are concatenated. The six models are trained for 70% of dataset and tested for remaining 30% of data. Using the confusion matrix, can predict the accuracy for all the six models. Comparing the six models SVM gives the higher accuracy for predicting liver disorder.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125125

G. MODEL EVALUATION

In this module, the trained machine learning model is tested using the .csv dataset file, by pre-establishing connectivity with the dataset file. This module, then gets accuracies for the used classifiers and find out one which results in higher accuracy, in comparison with other classifiers being used. The classifiers code is running on the PyCharm, that is collected data is trained and tested. The ROC curve metric function is applied to get the best classifier model to find to predict the liver disease disorder. Figures and tables must be cantered in the column. Graphics must not use stipple fill patterns because they may not be reproduced properly. Please use only *SOLID FILL* colors which contrast well both on screen and on a black-and-white hardcopy.

VI. RESULT AND DISCUSSION

In this work, a model for the detection of liver diseases disorder using machine learning models, namely Logistic Regression, KNN, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes are proposed to detected liver disorder with the help of .csv file. The proposed model present comparable results with higher accuracy compared to existing model.

Accuracy, Precision, Recall and F1 score

A. Accuracy

Accuracy is a metric that generally describe how the model performs across all classes. It is useful when all classes ae of equal importance. It is calculated as the ratio between the number of correct predictions to the total number of predictions.

B. Precision

The precision is calculated as the ratio between the number of positive samples correctly classified to the total number of samples classified as positive (either correctly or incorrectly). The precision measures the model's accuracy in classifying a sample as positive.

C. Recall

The recall is calculated as the ration between the number of positive samples correctly classified as positive to the total number of positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

D. F1 score

F1 score is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

Model		Precision	Recall	F1 score	Support	Accuracy
Logistic Regression	0	0.45	0.29	0.35	49	0.69
	1	0.75	0.86	0.80	121	
KNN	0	0.37	0.39	0.38	49	0.63
	1	0.75	0.73	0.74	121	
Support Vector Machine	0	0.00	0.00	0.00	49	0.71
	1	0.71	1.00	0.83	121	
Decision Tree	0	0.33	0.10	0.16	49	0.68
	1	0.72	0.92	0.80	121	
Random Forest	0	0.38	0.33	0.35	49	0.65
	1	0.74	0.79	0.76	121	
Naive Bayes	0	0.39	0.96	0.55	49	0.55
	1	0.96	0.39	0.55	121	

Table I performance metrices for classifiers

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125125

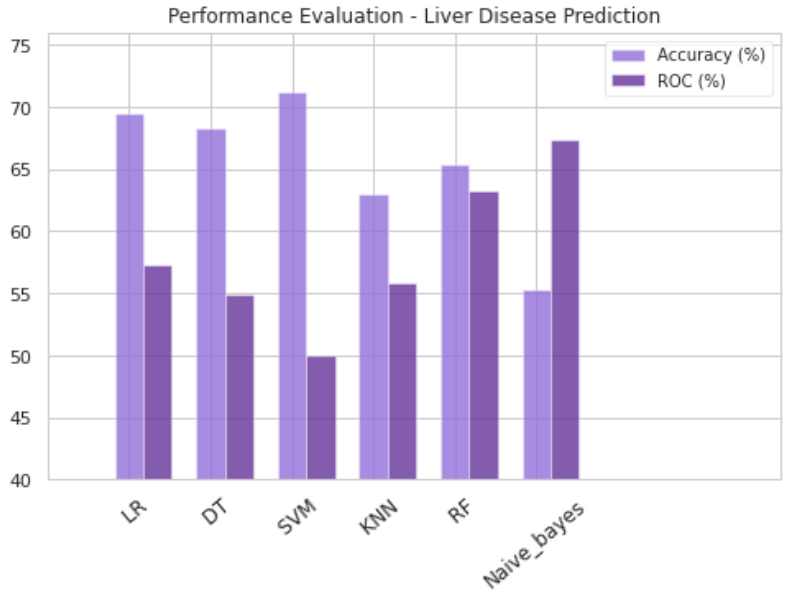

FIG.3 SAMPLE OUTPUT

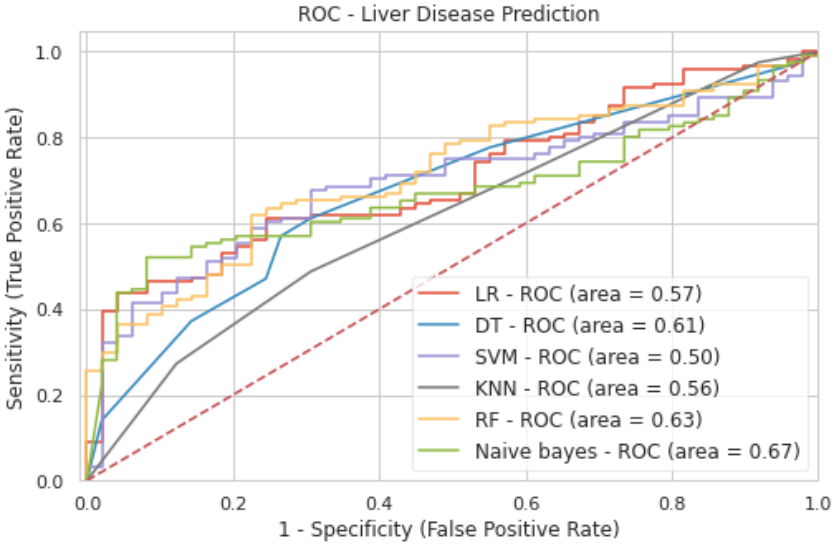

FIG.4 SAMPLE OUTPUT

	Model	Score
2	SVM	71.18
0	Logistic Regression	69.41
3	Decision Tree Classifier	68.24
4	Random Forest Classifier	65.29
1	KNN	62.94
5	Naive bayes	55.29

FIG.5 SAMPLE OUTPUT

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125125

VII.CONCLUSION

The early detection of the disease aids in the prevention and progression of the condition. Machine learning approaches aid in early disease diagnosis and the identification of relevant causative factors. This research will aid in the identification of patients with liver illness. When a patient's test results are projected to be positive, their reports and data might be thoroughly examined. In this project I have used six classifiers: Logistic Regression, KNN, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes. Comparing the above six algorithms, Support Vector Machine has the highest accuracy of 71.18% to find out the liver disorder by using the metrices precision, recall, f1-score. And ROC is used to plot the graph for Specificity (False Positive Rate) over Sensitivity (True Positive Rate).

REFERENCES

1. Liva Faes; Xiaoxuan Liu; Pearse A Kean, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis", The Lancet Digital Health, Volume 1, Issue 6, October 2019.
2. Anupama C.G.; Lakshmi C, "A comprehensive review on the crop prediction algorithms", Dept. of SWE, SRM Institute of Science and Technology, India, March 2021.
3. Pushpa Singh; Narendra Singh; Krishna Kant Singh; Akansha Singh," Diagnosing of disease using machine learning", 2021, Machine Learning and the Internet of Medical Things in Healthcare Pages 89-111, 2021.
4. AlexanderSelvikvag; undervold; Arvid Lundervold;Katarzyna Węgrzyn-Wolska, "An overview of deep learning in medical imaging focusing on MRI", Zeitschrift für Medizinische Physik, Volume 29, Issue 2, Pages 102-127, May 2019.
5. Yuanyuan Pan; Minghuan Fu; Biao Cheng; Xuefei Tao; Jing Guo,"Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform, IEEE Access (Volume: 8), May 2020.
6. Hiraj Dahiwade; Gajanan Patle; Ektaa Meshram," Designing DiseasePrediction Model Using Machine Learning Approach", 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019.
7. Durr-e-maknoon Nisar; Meshrif Alruily; Sultan H. Almotir; RashidAmin-" healthcare techniques thro deep learning"
8. Enhanced Deep Learning Assisted Convolutional Neural Networkfor Heart Disease Prediction on Internet of Medical Things Platform YUANYUANPAN, MINGHUAN FU, BIAO CHENG, XUEFEI TAO, AND JING GUO.
9. Early-Stage Risk Prediction of Non-Communicable Disease UsingMachine Learning in Health CPSRahatara Ferdousi; M.AnwarHossain; Abdulmotale b El Saddik. 35
10. Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS RAHATARA FERDOUSI, M. ANWAR HOSSAIN, (Senior Member, IEEE), AND ABDULMOTALEB EL SADDIK, (Fellow, IEEE)
11. ahatara Ferdousi; M. Anwar Hossain; Abdulmotaleb El Saddik,"Early- Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS", IEEE Special Section On AI And IoT Convergence For Smart Health, July 2021.
12. Alice Othmania; Abdul Rahman Taleb; Hazem Abdelkawy; Abdenour Hadid, "Age estimation from faces using deep learning: a comparative analysis", Computer Vision and Image Understanding, Volume 196, 102961,