Journal of Positive School Psychology http://journalppw.com
2022, Vol.6, No.4, 5489-5496

Analysis of Chronic Disease (Liver) Prediction Using Machine Learning

Jayavarapu Yajurved[1]
B. Tech,4thYear,
Department of Computer science and Engineering
SRM Institute of Science and
Technology, Chengalpattu, India

Padhi Sai Prasad[2]
B. Tech 4th year,
Department of Computer science and Engineering
SRM Institute of Science and Technology
Chengalpattu, India.

Dr. Umamaheswari KM[3*]
Department of Computing Technologies
SRM Institute of Science and Technology
Chengalpattu, India.
umamahek@srmist.edu.in

Abstract—
Liver disease is becoming one of the most fatal diseases in several countries. Patients with the Liver disease have been continuously increasing because of excessive consumption of alcohol, inhaling of harmful gases, intake of contaminated food, pickles, and drugs. These days, AI strategies have generally been utilized in clinical science for guaranteeing precision. In this work, we have precisely built computational model structure procedures for liver infection forecast. We utilized some effective characterization calculations (Random Forest, Logistic Regression-NN, and Support Vector Machine) for chronic liver disease patients which lasts over six months. We proposed an investigation model to predict liver infection with a high exactness value. Then, we analysed the good and bad values using a machine learning classifier which improvises the classification resultant. We examined that; the Support Vector Machine has been giving better outcomes contrasted with other classification models.

Keywords*—Liver Diseases, Machine Learning, K-NN, SVM, Classification, decision making.*

I. INTRODUCTION

In India Cirrhosis of the liver is a big health issue. According to the latest WHO data, Liver disease mortality in India reached 259,749 in 2017, or 2.95 percent of total deaths, accounting for 18.3 percent of all cirrhosis deaths worldwide [1]. The intricacy of the symptoms in the early stages of liver disease makes it difficult to identify. Because the liver continues to operate even when partially damaged, problems with liver illnesses are frequently not recognized until it is too late. An early diagnosis may be able to save a person's life. It is possible to identify the early signs of certain diseases, even if they are not visible to a

doctor's trained eye. Patients who are diagnosed early have a far greater chance of living a long life. Liver failure is associated with a high risk in Indians. India is expected to become the world's capital of liver disease by 2025 [2]. The epidemic of liver infections in India is caused by a desk-tied lifestyle, increased alcohol consumption, and smoking. There are about 100 types of liver infections.

The human liver is a fascinating internal organ that can perform about 500 different functions. Its main functions are boosting immunity, detoxifying and purifying, and producing proteins and hormones. It maintains blood sugar levels and prevents blood loss due to coagulopathy. In more complex metabolic activities, the liver is known as the internal organ and the liver can regenerate and repair its tissues. Failure of such organs can cause serious health problems[3]. According to a survey of the main causes of medical accidents in India, liver disease is on the top10 list of diseases, and worldwide, India was ranked 63rd in liver disease in 2017. Liver diagnosis is made by both imaging and liver functiontests.

Theliveristhebiggestorganoftheframeanditsmiles are importantfordigestingmealsandfreeingthepoisonous detailoftheframe.Thevirusesandalcoholuseleadthe liver closer to liver harm and lead a human to a life-threatening condition. There are many kinds of liver illnesses while hepatitis, cirrhosis, liver tumors, liver cancers, and plenty more. Among them liver illnesses and cirrhosis as the principal reason for death. Therefore, liver sickness is one of the foremost fitness issues withinside the global Every year, around 2 million human beings died internationally due to liver sickness. According to the Global Burden of Disease (GBD) project, posted in BMC Medicine, 1,000,000 peopleareidiedin2010duetocirrhosis,and a millionare affected by liver cancers Machine gaining knowledge has made a sizeable effect in the biomedical area of liver sickness prediction and prognosis [4]. With developments in the domain of artificial intelligence (AI) and ever-increasing technological improvements, innovative ML and DL approaches may evaluate medical data more effectively. Benefiting from this overall growth, researchers are constructing stronger ML and DL models to cope with progressively more complicated and abundant medical data. Medical professionals can benefit from ML in clinical settings as well. Clinicians' prognostication, diagnosis, imaging interpretation, and therapy can be aided by the use of machine learning (ML) prediction and classification algorithms [5]. The target of this investigation is to improve the presentation of the classification model for liver judgments by utilizing diverse classification calculations that have not been endeavored beforehand. The outcomes are expected to supplement past findings in accomplishing a comprehensive correlation of exactness rate. To accomplish this, the trials in this paper will use a comparable dataset as the past research, just as comparable assessment rules, which are accuracy, precision, and recall [6]. It is hard to formulate and cannot be optimized after the hyper-parameters have been tuned sufficiently, GA is one of the meta-heuristic algorithms that help in optimizing the hyperparameter. GA is based on biologically inspired operators like crossover, mutation, and selection. In place of Gradient Descent GD approaches like error backpropagation, GA is utilized as an alternate learning methodology [7]. In this study, the principle element is to expect the outcomes more: efficiently and decrease the fee of prognosis withinside the scientific zone Therefore, we used extraordinary type strategies forthe type of sufferers have liversickness.

To make the best possible clinical treatment decision the doctor needs to make the right decision to predict the various illnesses. If the doctor makes the wrong decision, it can lead to interruptions in medical treatment or death. We know that medical services have always been a big commercial perspective. Business flows are always carried out in these areas. Patients are always looking foragoodplatformforbetterservice.However, thereis no100%affordableplatformforallpatients.Therefore, there is a need for a comprehensive platform for problem-solving in health care and medical care in this area. Here, the main idea for improving medical services is to focus on the early detection of chronic diseases, focus on treatment and lead a betterlife.

This paper is summarized as follows: In section 2 describes the related words according to the classification of liver infection. Section 3 relates the proposed system model and the architectural flow of the system. Section 4 illustrates the results and analysis of the model as well as in sesection 5 the conclusion is portrayed.

II RELATED WORK

Machine Learning has pulled in an immense measure of investigation and has been applied indifferent fields. In medication, ML has demonstrated its force where it has been utilized to takecare of numerous crisis issues like

malignant growth treatment, coronary illness, dengue feveranalysis, etc. A few remarkable strategies such as decision tree, Logistic regression, Random Forest, and K-NN calculations have been utilized for liver examination.

Rajeswari et.al compares the ML Classification algorithm namely Naive Bayes, FT, and K-Star to obtain theaccuracy to predict the liver disorder. The algorithm is evaluated using the UCI repository which consists of 345 instances with 7 different attributes. Among these ML algorithms, FT outperforms well compare to the other three methods which provide 97.10% accuracy with less computational time [8]. To predict liver disorders, Rahman et al. [9] examined six alternative classification techniques: Logistic Regression (LR), KNN, Decision Tree (DT), Support Vector Machines (SVM), NB, and Random Forest (RF). Their investigation revealed that LR had the highest accuracy, with a score of 75%. A Comparative analysis of data classification accuracy using Liver disorder data in different scenarios helps the author to compare predictive performances of several classification algorithms like J.48, SVM, and Random Forest quantitatively. By analyzing the results multi-layer perception gives the overall best classification result with an accuracy of 71.59% than other classifiers[10]. The author compares the classification algorithm with original liver patients' datasets collected from the UCI repository and takes only the critical attributes which are obtained by features selection methods namely Random Forest, which outperforms well when compared to other feature selection methods. The results of our experiments show that the Random Forest algorithm is superior to all other methods that use feature selection with 71.8696% accuracy.

To break down the information of liver sicknesses utilizing Particle Swarm optimization with K-means classification. The presence of inflammation is identified using main two facets. The proposed evaluation upgraded the better result when contrasted with existing classification algorithms. The PSO-Kstar is the best reckoning for the arrangement of liver diagnosis as it improved the presentation in forecast exactness result in earlier stage [11]. The Pearson Correlation method is used to describe the relationship among the minerals which are present in the liver. Although, it tends the classifier to calculate the factual measures such as accuracy, sensitivity, specificity, and precision [12]. The Classification algorithm doesn't give high priority to utilizing duplication in liver sickness problems. Although the Case-Based Reasoning (CBR) and Classification and Regression Tree (CART) procedures could be helpful to distinguish liver infection [13]. The region of interest is taken out by utilizing the iso-shape structure strategy. Utilizing staggered fractal highlights and multi-space wavelet-includes the highlights are separated for better segregation limit. swarm advancement strategy is utilized to get better grouping by forming five cross approval mistakes[14]. The boosted classification 5.0 will be advanced utilizing hereditary calculation for that the liver dataset. The goal of utilizing hereditary calculation is to diminish the standard and increment exactness [15].

III. METHODOLOGY

The project aims to identify appropriate machine learning algorithms that can determine if an individual has liver disease. The facts set accrued for predicting given facts is cut up into the Training set and Test set. Generally, 7:3 ratios are carried out to cut up the Training set and Test set. The Data Model which became created by the usage of Random Forest, logistic regression, Decision tree, and K-NN is carried out at the Training set and is primarily based totally on the take a look at result accuracy, Test set prediction is done. The facts which became accrued may comprise lacking values which could result in inconsistency. To benefit from higher consequences facts, want to be preprocessed as a way to enhance the performance of the algorithm. The outliers must be eliminated and additionally, variable conversion wants to be done. Figure 1 shows the workflow of liver disease prediction. Below are the subsequent steps for making the dataset effective.

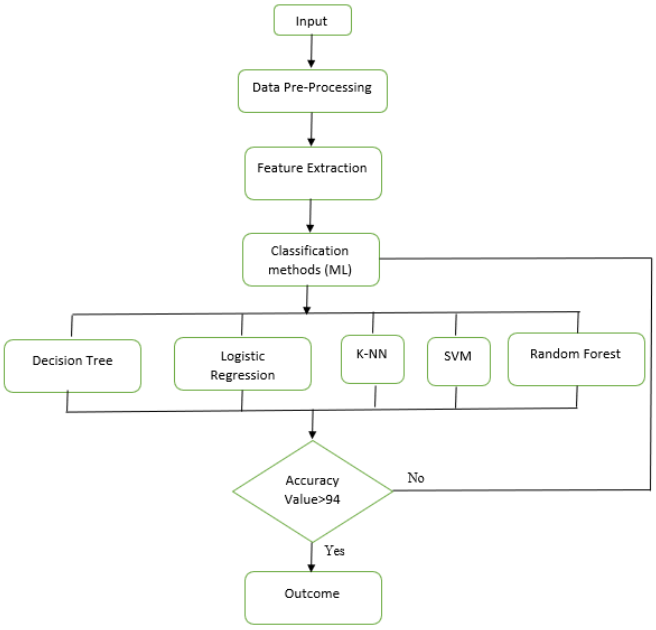


Figure 1: Architectural flow graph of liver disease prediction

3.1 Dataset Description:

The number of patients with liver disease has been steadily increasing as a result of excessive alcohol consumption, inhalation of toxic gases, and ingestion of contaminated food, foods, and medications. This dataset was used to assess prediction systems to alleviate physician burden. Thisdata collection contains 416 records of liver patients and 167 records of non-liver patients collected in Andhra Pradesh's northeast region. The "Dataset" column is a class label that is used to categorize groups as liver patients (with or without liver disease) (no disease). This data collection contains 441 records for male patients and 142 records for female patients. Any patient over the age of 89 is classified as "90." The dataset is 22.8 KB in size.[18]

3.2 Preprocessing:

The data in the real world has a lot of quality concerns, noise, inaccurate, and incomplete. It may lack relevant, specific properties and may contain missing values, as well as incorrect and spurious data. Preprocessing is essential for improving data quality. It helps to make data consistent by removing duplicates and inconsistencies. The datais to be Normalized to compare and enhance the quality of the results.The attributes of our AP datasets are listed in Table 1[16]. Albumin and Globulin Ratio attribute are the 4 missing values in AP datasets, we utilized the Replace missing value operator to handle the missing data, and we set the average value for replacement of the missing values. Following that, we used the Remove Duplicate Records operator to remove any duplicate records from our datasets. Cleaning/preparing data by analyzing univariate, bivariate, and multivariate processes by renaming the specified dataset, removing columns, etc. Data cleansing procedures and techniques vary from dataset to dataset[17-19]. The main goal of data cleaning is to detect and eliminate errors and anomalies to increase the value of the data in analysis and decision making.

Attribute	Type
Gender	Categorical
Age	Real number
Total_bilirubin	Real number
Direct_bilirubin	Real number
Indirect_bilirubin	Real number
Total_proteins	Real number
Albumin	Real number
Globulin	Real number
A/G ratio	Real number
SGPT	Integer
SGOT	Integer
ALP	Integer
Selector Field	Binomial (Class)

Table:1 Attributes of AP datasets

3.3 Feature Selection and Feature Extraction:

Both Feature selection and extraction are essential for dimensionality reduction, which helps to reduce the model complexity and overfitting.because dimensionality reduction is the crucial part of training the machine learning model.The entire dataset of liver patients is composed of all relevant or irrelevant data.attributes.Using feature selection, a subsetofliver patient datasets will be extracted from large datasets of liver patients, which comprise important features Using specific classification techniques on the important features of attributes after feature selection of UCI datasets. Here some of the important classification and regression algorithms are used to train the model and get good accuracy of the following machine learningalgorithms.

3.4 Support Vector Machine:

Popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the decision boundary that can segregate n-dimensional space into classes .So that we can easily put the new data point. SVM chooses the extreme vectors that help in creating the hyperplane[20]. These extreme cases arecalledsupport vectors. SVM algorithm can be used for Facedetection, Image classification, text categorization**,**etc. In SVM the data is labeled with the following condition:

$X_i w+b>0$, $C_i=1$ (1)

$X_i w+b<0$, $C_i=-1$ (2)

Where X is the feature vector, b represents the bias C_i is the class label, and the optimized SVM classifier in hyperplane gives as

$$minimize\ \frac{1}{2}\|w\|^2 \quad (3)$$

subject to $c_i(X_i w+b)\geq 1$ (4)

3.5 Logistic Regression:

Logistic regression comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. It is a significant machine learning algorithm because it can provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. Logistic Regression can be classified into three types.

1. Binomial: There can be only two possible types of the dependentvariables
2. Multinomial: There can be 3 or more possible unordered types of the dependentvariable
3. Ordinal: There can be 3 or more possible ordered types of dependentvariables[21]

3.6 Random Forest Classifier:

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random Forest Algorithm belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model. “The

greater number of trees in the forest leadstohigheraccuracyandpreventstheproblemof overfitting." Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.[21]

3.7 Decision Tree Classifier:

A Decision Tree is a Supervised learning technique. It is a tree-structured classifier, where internal nodes represent the features of a dataset,[22] branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are twonodes.

1.Decision Node: Decision nodes are used to make any decision and have multiplebranches of algorithms those decisions do not contain any furtherbranches.

Leaf Node: Leaf nodes are the output of those decisions and do not contain any furtherbranches.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. A decision tree simply asks a question and based on the answer of Yes or No, it further split the tree into subtrees The decision tree contains lots of layers, which makes it complex. It may have an overfitting issue, which can be resolved using the Random Forest algorithm.

3.8 K-Nearest neighbor:

K-Nearest Neighbor is one of the simplest algorithms based on the Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified in the goodwill suitecategory. Suppose there are two categories, Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particulardataset[23-24].

K-NN algorithm calculates the Euclidean distance is between the training and testing data is given as follows:

$$d(xi - xl)\sqrt{(xi1 - xl1)2 + (xi2 - xl2)^{2+\cdots+(xip-xlp)2}}$$

IV. EXPERIMENTAL RESULTS

Classificationisanimportantfactorthatsubdividesthepre-processeddataintodifferentlabelsbased on their

taxonomy. We are analyzing three different classification algorithms such asDecision tree, Support Vector Machine, and Random Forest are performed for enhancingprecision in the forecast of liver patients with feature selection.Performance parameters are the main factor that contrasts the best classifier among different classifier techniques. Thefollowing performance metrics as Accuracy, precision, Recall, and F-Score are considered to evaluate the confusion matrix in each progression [25]. We are focusing on the improvement of accuracy by comparing the three different classification algorithms. The confusion matrix is evaluated based on the,

TP addresses the quantity of accurately ordered positive examples.

FP addresses the quantity of misclassified positive cases.

FN addresses the quantity of misclassified negative occurrences.

TN addresses the quantity of accurately characterized negative cases.

Accuracy = $(TP+TN) / (TP+TN+FP+FN)$

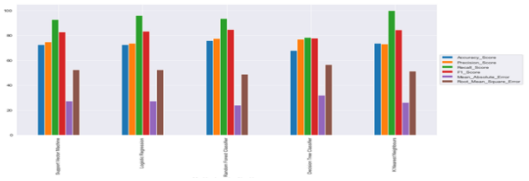
SVM changes the initial information into a higher measurement utilizing non-straight planning which coordinates a direct ideal isolating hyperplane. Anomalies anticipated in this work depend on the moderate scope ofthe liver inflammation test results. The SVM classification algorithm provides the best accuracy of the other classification algorithm.

Table1 depicts the evaluation values of performance metrics for the prediction of chronical disorder through different machine learning algorithms. Figure 2 shows the evaluation of distinct algorithms.

Table 2: Analysis of different classification Algorithms

S.no	Machine Learning Algorithm	*Accuracy* Score	*Precision* Score	*Recall* Score	*F1* Score	*Mean Absolute* error	Root Mean Square error
1	SVM	72.571429	74.838710	92.8	82.857143	27.428571	52.372294
2	LR	72.571429	73.619632	96.0	83.333333	27.428571	52.372294
3	RFC	76.000000	77.483444	93.6	84.782609	24.000000	48.989795
4	DTC	68.000000	77.165354	78.4	77.777778	32.000000	56.568542
5	K-NN	73.714286	73.099415	100.0	84.459459	26.285714	51.269596

Figure 2: Performance graph for the prediction of liver diseases using distinct ML algorithms



V. Conclusion

Diseases of the liver and heart become more common over time. With continuous technological advances, these will increase in the future. Today, people are becoming more health-conscious and are taking yoga and dance classes. Still, the sedentary lifestyle and luxury are constantly being introduced and improved. The problem will last for a long time. So, in such a scenario, our project is very useful to society. The dataset used in this project gave 96% accuracy in the SVM model. While it may be difficult to achieve such accuracy with such large datasets, the conclusion of this project is clear that liver risk can be predicted. In the future, philosophy is utilized to examine the liver area into distinct compartments for better classification accuracy. However, the technique requires further improvement generally to include the excretion of the liver into various parts: renal cortex, renal segment, renal medulla, and renal pelvis.

References:

[1] Mishra, Debakanta, Kaibalya R. Dash, ChittaranjanKhatua, SubhenduPanigrahi, Prasanta K. Parida, Sambit K. Behera, Rakesh K. Barik et al. "A study on the temporal trends in the etiology of cirrhosis of the liver in coastal eastern Odisha." *Euroasian Journal of Hepato-Gastroenterology* 10, no. 1 (2020):

[2] Sontakke, Sumedh, Jay Lohokare, and Reshul Dani. "Diagnosis of liver diseases using machine learning." In *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, pp. 129-133. IEEE, 2017.

[3] Priya, M. Banu, P. Laura Juliet, and P. R. Tamilselvi. "Performance analysis of liver disease prediction using machine learning algorithms." *International Research Journal of Engineering and Technology (IRJET)* 5, no. 1 (2018): 206-211.

[4]https://bmcmedicine.biomedcentral.com/articles

[5] Garg, Arunim, and Vijay Mago. "Role of machine learning in medical research: A survey." *Computer Science Review* 40 (2021): 100370.

[6] SinaBahramirad, Aida Mustapha, Maryam Eshraghi Classification of Liver Disease Diagnosis: A Comparative Study ISBN: 978-1-4673-5256-7/13/$31.00 ©2013 IEEE

[7] Rezai, Bahram, and Ebrahim Allahkarami. "Application of Neural Networks in Wastewater Degradation Process for the Prediction of Removal Efficiency of Pollutants." In *Soft Computing Techniques in Solid Waste and Wastewater Management*, pp. 75-93. Elsevier, 2021.

[8] Rajeswari, P., & Reena, G. (2010). Analysis of Liver Disorder Using Data Mining Algorithm. Global Journal of Computer Science and Technology, Retrieved from https://computerresearch.org/index.php/computer/article/view/652

[9] **A.** K. M. Rahman, F.M. Shamrat, ZarrinTasnim, Joy Roy and Syed Hossain, "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms", *International Journal of Scientific & Technology Research*, vol. 8, pp. 419-422, 2019.

[10]Baitharu, Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset." *Procedia Computer Science* 85 (2016): 862-870.

[11] P. Thangarajul, R.Mehala, Performance Analysis of PSO-KStar Classifier over Liver Diseases, International Journal of Advanced Research in Computer Engineering, 2015.

[12] L. A. Auxilia, "Accuracy Prediction Using Machine Learning Techniques for Indian Patient Liver Disease," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2018, pp. 45-50, doi: 10.1109/ICOEI.2018.8553682.

[13] R.H.Lin, "An Intelligent model for liver disease diagnosis", Artificial Intelligence in Medical, Vol. 47, no. 1 (2009), PP. 53-62.

[14] Krishnamurthy, Rajesh Krishnan, Sudhakar Radhakrishnan, and Mohaideen Abdul KadharKattuva. "Particle swarm optimization-based liver disorder ultrasound image classification using multi-level and multi-domain features." International Journal of Imaging Systems and Technology (2020).

[15] Hassoon, Mafazalyaqeen, MikhakSamadiKouhi, Mariam Zomorodi-Moghadam, and MoloudAbdar. "Rule optimization of boosted c5. 0 classification using genetic algorithm for liver disease prediction." In 2017 international conference on computer and applications (icca), pp. 299-305. IEEE, 2017.

[16] Bahramirad, Sina, Aida Mustapha, and Maryam Eshraghi. "Classification of liver disease diagnosis: a comparative study." In *2013 Second International Conference on Informatics & Applications (ICIA)*, pp. 42-46. IEEE, 2013. [17] Esfahanian, Parsa, and Mohammad Akhavan. "Gacnn: Training deep convolutional neural networks with genetic algorithm." *arXiv preprint arXiv:1909.13354* (2019).
[18] Jeyalakshmi, K., and R. Rangaraj. "Accurate liver disease prediction system using convolutional neural network." *Indian Journal of Science and Technology* 14, no. 17 (2021): 1406-1421.
[19] Schelter, Sebastian. "Data Validation and Data Cleaning."
[20] Himani Rani, Dr. Gaurav Gupta,"Prediction Analysis Techniques of Data Mining: A Review", IJCSMC, Vol 8,Issue5, May 2019, Pp:15-22.
[21] [RF and LR]Chieh-Chen Wu , Wen-Chun Yeh , Wen-Ding Hsu , Md. Mohaimenul Islam , Phung Anh (Alex) Nguyen , Tahmin Nasrin Posly , Yao-Chin Wang, "Prediction of fatty liver disease using machine learning algorithms ", Computer Methods and Programs in Biomedicine, Vol.170, Pages 23-29, March2019.
[22] Pillai, N. Sowri Raja, K. Kamurunnissa Bee, and J. Kiruthika. "Prediction of heart disease using rnn algorithm." *International Research Journal of Engineering and Technology* 5 (2019).
[23] Rahman, AKM Sazzadur, FM Javed Mehedi Shamrat, ZarrinTasnim, Joy Roy, and Syed Akhter Hossain. "A comparative study on liver disease prediction using supervised machine learning algorithms." *International Journal of Scientific & Technology Research* 8, no. 11 (2019): 419-422.
[24] Singh, Smriti Mukesh, and Dinesh B. Hanchate. "Improving disease prediction by machine learning." *Int. J. Res. Eng. Technol* 5 (2018): 1542-1548.
[25]Benteng Ma, Xiang Li, Yong Xia, Yanning Zhang, Autonomous deep learning: A genetic DCNN designer for image classification, Neurocomputing, Volume 379, 2020, https://doi.org/10.1016/j.neucom.2019