ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter568102020.pdf
https://doi.org/10.30534/ijeter/2020/568102020

# An Intelligent Approach for Prediction of Liver Disease using Machine Learning Models

**Ram Prasad Reddy Sadi[1], PanduRanga Vital Terlapu[2], Soumya Bonela[3]**
[1]Department of Computer Science & Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, reddysadi@gmail.com
[2]Dept. of Computer Science and Technology, Aditya Institute of Technology and Management, Tekkali, Srikakulam, Andhra Pradesh, India, vital2927@gmail.com
[3]Department of Computer Science & Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, bonelasoumya77@gmail.com

## ABSTRACT

Liver disease (LD) is a common disease in the world. The functionality of the liver is very crucial in the human body where it impacts much physical functionality like the manufacture of protein, Metabolism of iron and sugar, and blood clotting. In the present decade, the research on prediction and prevention of LD with Data Mining and artificial intelligence concepts is very important. For this, artificial inelegancy concepts play a vital role. Many researchers have to utilize machine learning (ML) models for predictions of diseases. In this paper, we present the empirical statistical analysis for prevent the LD and apply efficient ML models for predictions of liver diseases in early with low cost. The data set is collected from hospital and reputed clinical centers of Andhra Pradesh, India during 2018 to 2020. The data set contains personal and clinical information. We apply reputed 5 ML models that are KNN, SVM, RF, Naïve Bayes, and AdaBoost. As per performance analysis, the KNN and AdaBoost models perform highly than other experimental models with accuracy 1(100%) for predicting LDs. The remaining also performs well, where their accuracy values are above 0.86(86%).

**Key words:** Liver Disease, Machine Learning, Prediction, Andhra Pradesh, Patient data

## 1. INTRODUCTION

Liver is one of the significant major organs in human body. The normal liver occupies the right upper quadrant, extending from the fifth inter costal space in the mid clavicular line down to the right costal margin. The lower margin descends below the costal margin during inspiration [5]. The average weight of liver is 1800 gm in men and 1400 gm in women. The liver is the second largest and the heaviest organ in the body and serves a key role in critical metabolic pathways and synthetic functions [6]. The liver is the biggest organ of the body encased inside the correct lower rib confine underneath the stomach. It is totally secured by instinctive peritoneum just as totally secured by a thick unpredictable connective tissue layer that lies profound to the peritoneum. Liver is partitioned in to two head projections, a huge right flap, and a little left projection isolated by falciform tendon. The correct flap is considered by numerous anatomists to incorporate a second rate quadrate projection and a back quadrate flap. Liver has five surfaces as Anterior, back, prevalent, mediocre, and right [9][10].

Acute liver disorder most regularly effects from acute huge liver cell corruption made by toxic drugs and viral hepatitis and harmful medications and substance additionally it follows the acute greasy difference in the liver Acute liver disappointment is portrayed by Hypoglycemia, Jaundice, Electrolyte and AST, LDH,ALT. There are three sorts of acute liver disorder that are FHF, Acute or CHF and Sub-acute HF. FHF is a symptom, described byserious encephalopathy ensuing on monstrous Carcinoma of the liver. The subsequent one is Acute or CHF; this may result from protein over-burden sepsis or intercession with medications or medical procedures. The third one is Sub-acute Hepatic Failure – It is characterized as an acute disappointment happening in patients without previous liver ailment, in whom the indications of encephalopathy grow over about two months after the beginning of the sickness [11][12][13].

ML is part of artificial intelligence (AI). It provides tools and methods that can solve diagnostic and prognosticative problems in various medical domains. It is being used for the investigation of the significance of clinical parameters and of their mixes for forecast, for the extraction of clinical information for examine, for treatment arranging, and for generally speaking patient administration. The fruitful fortunate execution of ML models can encourage the combination of system based frameworks in the clinical space giving chances to improve the work of clinical specialists and in this way improve the productivity and adequacy of clinical consideration[14][15].

## 2. RELATED WORK

Deutschmann et al. [29] carried out a study on identifying chitinase-3-like protein 1 as novel neutrophil antigenic target in crohn's disease. Early detection of disease is most important for saving the life and money. Machine learning algorithms are widely used in predicting the diseases based on their characteristics. PhaniMadhuri et al. [19] proposed a model for early detection of skin disease. The proposed framework was able to classify the images of skin diseases with more certainty. The system also indicates the individuals the risk of the disease in due duration if proper car is not taken. NarasingaRao et al. [20] explored the process of predicting the chronic kidney

disease using C4.5 algorithm based on the features {Haemoglobin, Blood glucose Random, Albumin, Scum createnine, Sodium, hypertension, diabetes mellitus, class}. Razia et al. [22] carried out a survey on the prediction of malaria disease by considering parasites life stages, erythrocytes segmentation, density estimation of parasites and other symptoms. Sajana et al. [27] carried out a survey on the malaria disease dataset to predict the chance of human beings affected with that disease using machine learning and image processing techniques. The authors observed that machine learning techniques has a broader usage for identifying the criticality of the disease. Sajana et al. [28] further explored their research on the malaria disease set to improve the accuracy of the suggested model in [27]. The authors contributed three models using decision trees (C 4.5), NaïveBayesian Networks and Radial Basis function and compared their results.

Razia et al. [21] studied the problem of thyroid disease using Support Vector Machine, Multiple Linear Regression, Naive Bayes and Decision trees. The authors concluded that the results generated using decision trees could be used to identify the disease more accurately as compared to the other models. Shinde et al. [23] presented a fourfold structural model for health prediction using machine learning. This study helps the new researchers in getting an overview on the usage of machine learning techniques in the domain of health. Bommadevara et al. [24] applied machine learning techniques for predicting the heart disease using the machine learning algorithms. The authors concluded that the primary cause for heart disease is THAL and reported that only 5% of the individuals had high probability of suffering from a heart attack. They also claimed that aged people in the range of 50-60 also had a high chance of getting a hear stroke. Srinivas et al. [25] used artificial neural networks in proposing a safety system for the human beings suffering from health related issues such as heart problem and to increase their longevity. Rajesh et al. [26] in their work explored the heart disease set using the machine learning algorithms like Naïve Bayes algorithm and decision trees. Naïve Bayes algorithm was used to analyse the dataset to identify the risk factors and decision trees were used to predict the accuracy of the disease.

Detection and forecasting of liver diseases (LD) with ML models and neural networks are very crucial for the researchers and analysts. Dhingra et al. (2020) [1] analyzed UCI ML repository Indian liver dataset with ML models like SVM, logistic regression, NB and ANN. In this, they found that ANN model is the better performed model with 80.7% of accuracy than other experimental models. El-Shafeiy et al., (2018) [2]researched on Egyptian LD data set with hybrid ML models that it describes as an ensemble (C5.0 + NB + SVM) classifier. This is proposed to classify the data with combination of Boosted C5.0, SVM and NB ML algorithms chosen rules. Their proposed model performs 97.2% of accuracy. Kaiaqzek et al., (2019) [3] proposed novel methodology to identification of HCC (Hepatocellular carcinoma). In this, they used C-SVC type SVM with 2level GA optimizer and get the 88.49% of accuracy. Abdar et al.,(2017) [4] applied new DT model on UCI ILPD and they also applied CHAID and C5.0 models for creating the rules in LD. They conclude that C5.0 via Boosting model had an accuracy of 93.75% and CHAID model performed 65.00% of accuracy. Gatos et al., (2017) [5] studied on CLD (Chronic liver disease) utilizing SWE images along with CAD system. In this work, they focused on 5 cluster segmentation of 35 features. They analyzed total 126 patients data that 70 CLD and 56 non-CLD. As per classification analysis, the highest accuracy was found in SVM model with 87.3% comparative other experimental set up models.

Fatty liver disease (FLD) is very complicated disease in clinical level; it is related with mortality. Wu et al., (2019) [7] researched on FLD with noble ML methods like NB, RF, LR and ANN models. In this, they found ANN model was very accurate than other ML algorithms. LD is one of the significant ailments on the planet, Liver is one of the immense strong organ in the human body; and is likewise viewed as an organ in light of the fact that, among its numerous capacities, it makes and secretes bile. The liver auditoriums are indispensable job in numerous physical capacities from protein assembling and blood thickening to fat, sugar and iron digestion. Liver issue infections are any issue of liver reason that purpose behind disorder. Ansari et al., (2011) [8] exhibits an ANN based methodology for the analysis of hepatitis infection. The dataset utilized for this reason for existing is taken from the UCI ML database. Both supervised and neural models have been dissected with various structures, learning and enactment capacities. It is presumed that the regulated model performed superior to the unaided one. The paper compares the consequences of the past examinations on the analysis of hepatitis which utilize the equivalent dataset. Lee et al., [9] separate the characterization of liver malady into progressive multiclass clustering. The SVM includes state-of-the-art ML. The classifier is a piece of CADx, which helps radiologists in precisely diagnosing liver sickness. They figure separating between cysts, hematoma, cavernous hemangioma(CH), and normal tissue, and apply SVM to ordering the sicknesses.Sivakumar et al. [30] proposed new mixed mode database miner (MMDBM) algorithm for successful analysis of breast cancer dataset. The proposed techniques design best upon supervised learning in quest (SLIQ) and decision tree algorithms. This is suitable for handling both numerical and categorical data. Raghav et al. [31] proposed a system, in which resemblance factor is evaluated from the extracted keywords. In order to identify the difference between SARS affected and others, the proposed scheme fetches the inputs from user's displayed in the form of text. Deep RNN model is used to process the data. The J48graft algorithm is used to carry the classification based on the type of infection and symptoms of each user. The basic objective of the model is to identify the SARS disease at an initial stage.

## 3. PROPOSED MODEL

Fig.1shows the proposed model for Liver Disease Identification using Machine Learning (ML) Algorithms. In this, we collected the Liver Disease (LD) and non-LD patients' data from state of Andhra Pradesh, India. The whole information stored into the secondary storage section. In this process, avoid unnecessary information for the experiment. For this, we choose cleaning and preprocess the information and get the pure data set. This data set stored into the secondary storage section as *.csv format. As per dataset, we want to conduct the statistical

analysis and fit the ML model for predicting LD. The statistical reports are very useful for the doctors, analytics and decision makers for the preventions about LD. We evaluate the ML models utilizing performance parameters.

### 3.1 Description of Liver Data Set

The table shows the description of the Liver Data set collected from Andhra Pradesh during 2018 march to 2020 February with 16 features attributes and one 2 classes (Diseased and non-Diseased)attribute. We choose 1640 instances for the experiment. In this, 586 attributes are related to non-liver disease and 874 instances are related to liver disease. The detailed information about attributes described in the table 1. It describes attribute name, Type and description of attributes with values and ranges.

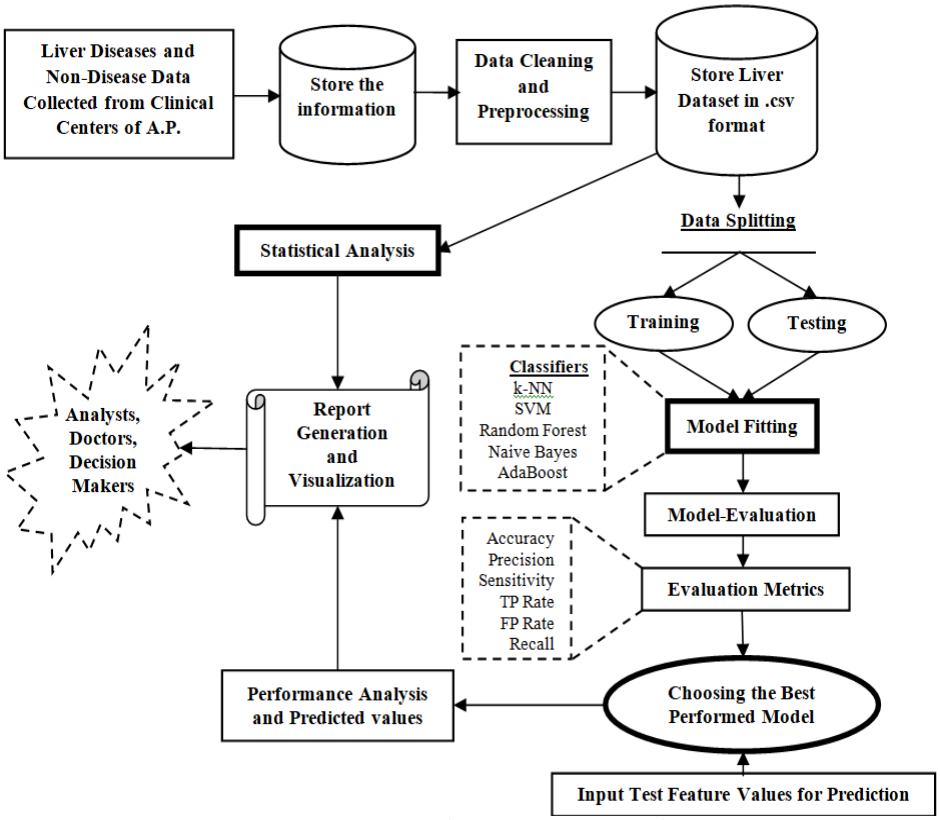


**Figure 1:** Proposal model for Liver Disease Identification

**Table 1.**Description of Andhra Pradesh Liver(APL) Data Set

| Sl. No | Attribute | Type | Description |
|---|---|---|---|
| **1** | **Age** | Continuous | Age of the individual Range is **6 to 99** |
| **2** | **Gender** | Categorical | Sex of individual **Female – 0 Male- 1** |
| **3** | **Smoke** | Categorical | Habit of smoke, values are **NO-0 YES-1** |
| **4** | **Drink** | Categorical | Habit of smoke, values are **NO-0 YES-1** |
| **5** | **Vomiting** | Categorical | Vomiting Sensation, values are **Absent-0 Present-1** |
| **6** | **Headache/ Bone Ache** | Categorical | Headache or Bone Ache, **Absent-0 Present-1** |
| **7** | **Fever** | Categorical | Fever, **Absent-0 Present-1** |
| **8** | **BP** | Categorical | Blood Pressure, **Normal-0 Low-1 High-2** |
| **9** | **Total_Bilirubin** | Continuous | Total_Bilirubin,range is **0.4 to 75** |
| **10** | **Direct_Bilirubin** | Continuous | Direct_Bilirubin range **is 0.1 to 19.7** |
| **11** | **Alkaline_Phosphotase** | Continuous | Alkaline Phosphates range **is 10 to 4929** |
| **12** | **Alamine_Aminotransferase** | Continuous | Alamine Aminotransferase – range **10 to 2000** |
| **13** | **Aspartate_Aminotransferase** | Continuous | Aspartate Aminotransferase – range **5 to 4929** |
| **14** | **Total_Protiens** | Continuous | Total_Protiens –range is **0.9 to 7.7** |
| **15** | **Albumin** | Continuous | Albumin-range –range is**0.9 to 7.7** |
| **16** | **A-G_Ratio** | Continuous | Albumin_and_Globulin_Ratio –range is **0.3 to 4.0** |
| **17** | **Diagnosis(Class)** | Categorical | Non-Liver Disease (Class 0) and Liver Disease (Class1) |

### 3.2 Naïve Bayes (NB) Model

It expects that the presence of an unambiguous aspect of a class is autonomous of every other aspect. As per Bayes theorem, the contingent probability is given by the Equations (1) and (2).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{.....(1)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{.....(2)}$$

### 3.3 SVM Model

An additional fantastic ML model is support vector (SVM) machine that can be utilized for both regression problems as well as classification. In this, 'n' features are addressed on the n-dimensional plane spaces with every segment described by the assessment of a particular coordinate. A data segment containing n qualities is plotted on this n-dimensional space [16]. The fact of the matter is to discover a hyper-plane which orders and builds the edge in an n-dimensional space (shown in figure 2)

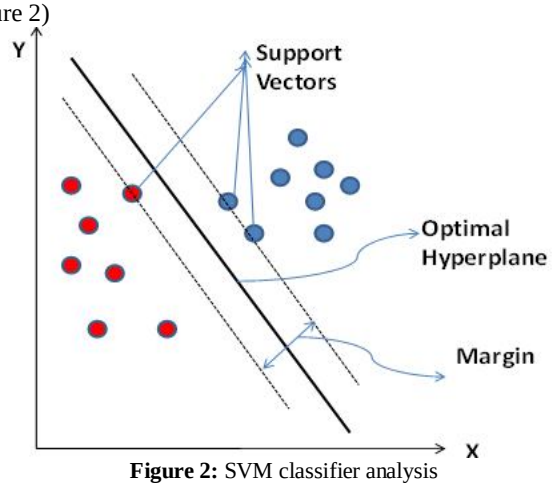


**Figure 2:** SVM classifier analysis

### 3.4 K-Nearest Neighbors' (k-NN) Classification

The k-NN is a nonparametric statistic technique fitted for regression problems as well as classification. It conceives the k nearest points of data in the preparation models. The yield contrasts dependent on the way that KNN is utilized for regression problems as well as classification. The yield predicts the class to which an information point has a place dependent on how intently it matches with the k closest neighbors. This is one of the case based learning, or sluggish learning calculations, on the grounds that the capacity considers the nearby information focuses and all calculation is conceded until arrangement. This calculation utilizes separation capacity to ascertain the nearby inexact with the K Nearest neighbors [17]. For uninterrupted factors, Minkowski (eq.5), Euclidean (eq.3) Manhattan (eq. 4) and measures of outdistance are utilized and hamming outdistance for straight out factors appeared in conditional equations (3, 4,5).

### 3.5 Confusion Matrix

In this, we express the AP Liver Disease Confusion Matrix analysis structure for each ML model. The table 2 depicts problem with2 classes that are 0 specifies NLD and 1specifies LD. It is construed with True positive, True Negative, False Negative and False positive that are specified with cells like (0, 0), (0, 1), (1, 0) and (1, 1) relatively. The diagonal values of the matrix represent the correctly classified instances by that ML model and non-diagonal values have represented the incorrectly classified instances. Confusion matrix is very curial for calculations of performance parameters like accuracy of model, recall, precision, F1-value and so on [18].

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{.....(3)}$$

$$Mahattan\ Distance = \sum_{i=1}^{k} |x_i - y_i| \quad \text{.....(4)}$$

$$Minkowski\ Distance = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \quad \text{.....(5)}$$

**Table 2.**Analysis of Confusion Matrix Structure

| | | Predicted Values | |
|---|---|---|---|
| Actual Values | **Classes** | **Non-Liver Disease (0)** | **Liver Disease (1)** |
| | **Non-Liver Disease (0)** | (0,0) | (0,1) |
| | **Liver Disease (1)** | (1,0) | (1,1) |

### 3.6 Performance Parameters

Performance specifies the working model efficiency. We will measure this with many parametric measures like classification accuracy, rates and ratios of truly and falsely classified instance and so on. The equations 6 to 17 describe the parameters like TPR, FNR, FPR, F1 Score and etc. These equations are clearly specified with their parametric values of the model performance. Every parameter specifies their individual behavioral value reflects to the particular data set with particular ML model.

$$TPR = \frac{\sum True\ Positive}{\sum Condition\ Positive} \quad \text{.....(6)}$$

$$FNR = \frac{\sum False\ Nagative}{\sum Condition\ Positive} \quad \text{.....(7)}$$

$$FPR = \frac{\sum False\ Positive}{\sum Condition\ Negative} \quad \text{.....(8)}$$

$$SPC\ or\ TNR = \frac{\sum True\ Nagative}{\sum Condition\ Negative} \quad \text{.....(9)}$$

$$Prevalence = \frac{\sum Condition\ Positive}{\sum Total\ Population}. \quad \text{.....(10)}$$

$$PPV\ or\ PRC = \frac{\sum True\ Positive}{\sum Predicted\ Condition\ Positive} \quad \text{.....(11)}$$

$$FOR = \frac{\sum False\ Negative}{\sum Predicted\ Condition\ Negative} \quad \text{.....(12)}$$

$$Accuracy(ACC) = \frac{\sum True\ Positive + \sum True\ Negative}{\sum TotalPopulation} \quad \text{.....(13)}$$

$$FDR = \frac{\sum False\ Positive}{\sum Predicted\ Condition\ Positive} \quad \text{.....(14)}$$

$$NPV = \frac{\sum True\ Negative}{\sum Predicted\ Condition\ Negative} \quad .....(15)$$

$$DOR = \frac{LR+}{LR-} \quad .....(16)$$

$$F_1 score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad .....(17)$$

## 4. Results and discussions

In this section, we promote necessary simulation results that are correlation feature attributes of LD on class attribute (LD Patients (1) and non-LD patients (0)), statistical analysis reports and results, and MLs performance analysis and predicted values.

### 4.1 Correlation Feature Attributes Analysis:

The figure 3 shows correlation Liver attribute analyzing with values between -1 and +1, and colors (Red and Blue). As per correlation values analysis, the value is one thatit specifies highly correlated attributes (indicated dark red) and minus values are declared as under correlated attributes (indicated blue color). Neutral correlated attribute values are zero or very nearer to zero (specified color is light blue).

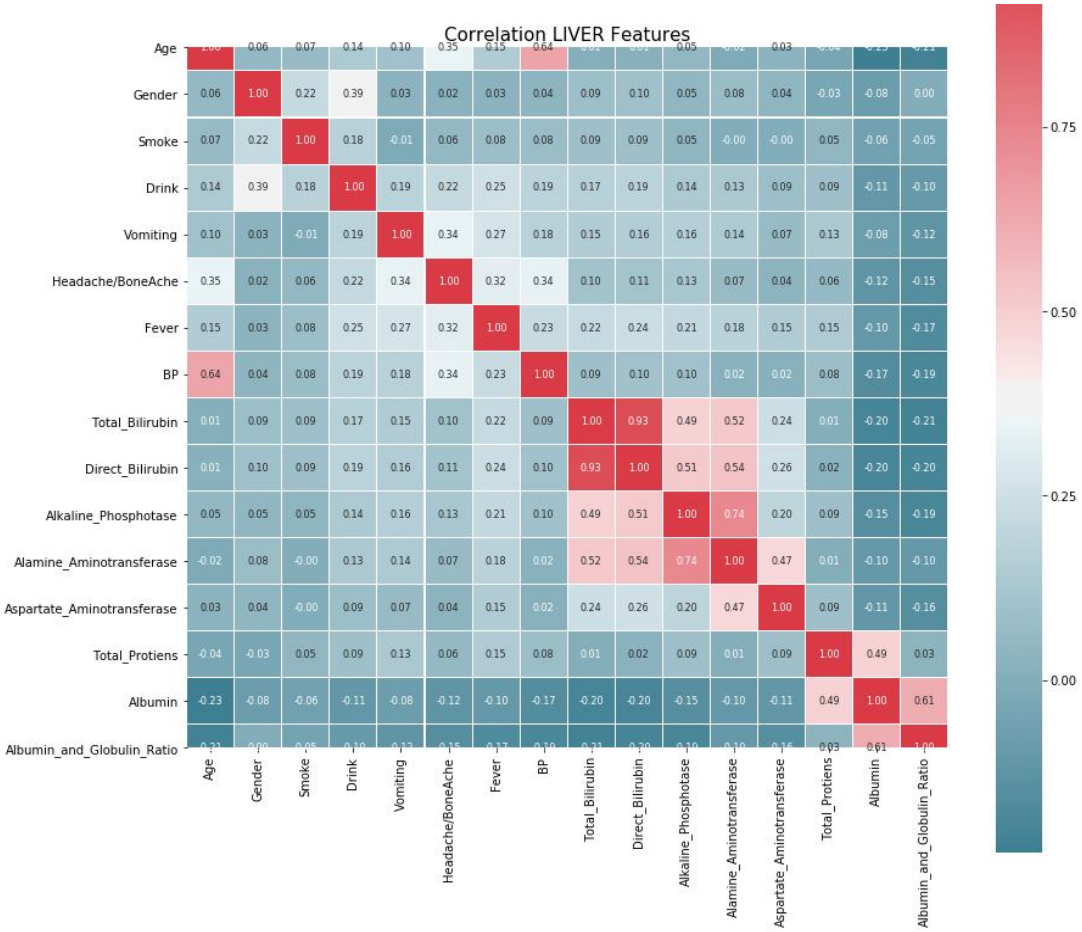


**Figure 3:** Correlation feature attributes of A.P. Liver (APL) Data set

The 'BP' attribute is correlated with the feature attribute'Age' that the value is 0.64.Some of the feature attributes like 'Total-Bilirubin', Direct-Bilirubin,,Alkaline-Phosphatases' and 'Alanine –Aminotransferase' are correlated to each other where the correlation value is greater than 0.49. Some other attributes also correlated to other that are 'Total-Proteins', 'Albumin' and 'Albumin_and_Globulin_Ratio' that Albumin_and_Globulin_Ratio is correlated to Albumin with 0.61 correlation value and 'Total-Proteins' attribute is correlated with 'Albumin' with 0.49 correlated value.

### 4.2 Statistical Analysis

The table 3 describes about Categorical (nominal) attributes of APL data set. As per ratio of gender, most of instances are male than female that there are 685 male instances and only 191 female belonging to the LD instances set (total 876) because of habits, life style or work environment. The more number of smokers and drinkers instances are in LD set than non-LD. The detailed analysis is shown in table 3 about symptoms ofliver disease patients like vomiting, head and bone ache, fever, and BP.

**Table 3.**Categorical Attribute Statistical nominal Values

| Attribute Statistical Counting Values | | | | |
|---|---|---|---|---|
| Attributes | **Attribute Value** | **Non-Liver Disease(NLD)** | **Liver Disease(LD)** | **Total-Dataset(TD)** |
| Gender | Male | 414 | 685 | 1099 |
| | Female | 174 | 191 | 365 |
| smoke | No (0) | 318 | 412 | 730 |
| | Yes (1) | 270 | 464 | 734 |
| Drink | No (0) | 500 | 418 | 918 |
| | Yes (1) | 88 | 458 | 546 |
| Vomiting | Absent (0) | 472 | 245 | 717 |
| | Present (1) | 116 | 631 | 747 |
| Headache/ Bone Ache | Absent (0) | 444 | 293 | 737 |
| | Present (1) | 144 | 583 | 727 |
| Fever | Absent (0) | 470 | 210 | 680 |
| | Present (1) | 118 | 666 | 784 |
| BP | Normal (0) | 398 | 368 | 766 |
| | Low BP (1) | 91 | 182 | 273 |
| | High BP (2) | 100 | 327 | 427 |

**Table 4.**IntegralAttributesStatisticalValues

| Mean and Median Attributes Statistics | | | | | | |
|---|---|---|---|---|---|---|
| | **Mean Values** | | | **Median** | | |
| Attributes | **Non-Liver Disease(0)** | **Liver Disease(1)** | **Total** | **Non-Liver Disease(0)** | **Liver Disease(1)** | **Total** |
| Age | 41.75427 | 45.63616 | 44.07808 | 40 | 46 | 45 |
| TB | 1.219283 | 4.65389 | 3.275342 | 0.9 | 1.8 | 1.3 |
| DB | 0.384642 | 2.199199 | 1.47089 | 0.2 | 0.8 | 0.3 |
| AP | 96.87884 | 325.1705 | 233.5411 | 49 | 194 | 151 |
| AA | 36.3686 | 186.4611 | 126.2185 | 28 | 74 | 48 |
| AAT | 114.6672 | 213.7586 | 173.9863 | 93 | 168 | 126 |
| TP | 5.312799 | 6.237071 | 5.866096 | 5.5 | 6.4 | 6.2 |
| Albumin | 3.268601 | 3.071739 | 3.150753 | 3.2 | 3 | 3.1 |
| AG_Ratio | 1.074198 | 0.922197 | 0.983205 | 1 | 0.9 | 1 |

The table 4 projects the mean and median values of the attribute age and clinical result attributes. As per observations of clinical results, the mean and median LD values of the attributes like 'Total-Bilirubin', Direct-Bilirubin, 'Alkaline-Phosphatases' and 'Alanine –Aminotransferase' are very high than the Non-LD. The values of Aspartate-Aminotransferase and Total-Proteins are also some high mean and median values in LD comparative Non-LD.

### 4.3 Experimental set up for Machine Learning (ML) models on APLD dataset

For the experiment, we apply 5 ML algorithms that are SVM, k-NN, Naive Bayes (NB), AdaBoost (AB) and Random Forest (RF) on APLD data set. In this, we analyze the confusion matrices of ML models, parameters of performance like classification accuracy, AUC, F1-score, precision and recall, and ROC analysis. Finally, we compare the ML models' accuracy for predicting the LD.

#### 4.3.1 Confusion Matrices Analysis

Figure 4 depicts the confusion matrices of experimental ML models. The confusion matrix is constructed with combination of true and predicted elements of LD and NLD. In this, the '0' indicates the NLD and LD is specified with '1'. In other words, it is combination of true positive, true negative, false positive and false negative values. The model SVM with RBF kernel classifies correctly of LD and N-LD out of 1460 total instances, and incorrectly classified instances are 198. The confusion matrix of SVM is shown in Figure 4(A) in detail. Figure 4(B) describes about confusion matrix of k-NN model. 1460 (586 + 874) instances are classified correctly by k-NNwith 100% accuracy. The model Naive Bayes is classified 1296 correctly out of 1460 instances. The confusion matrix of NB is shown in Figure 4(C) in detail. Figure 4(D) describes about confusion matrix of AdaBoost model and 4 (E) describes Random Forest Tree(RF Tree).

#### 4.3.2 Accuracy Analysis

In this, we want to analyze the Non-Liver Diseases (NLD), LD and average weighted classes NLD and LD.

#### 4.3.2.1 Target Class NLD classification Analysis:

Table 5 shows the class NLD performance parameters analysis. In this, the class NLD classifies 100% by the k-NN and AdaBoost algorithms where CA (classification accuracy) and AUC values are 1. The random forest model also performs better way where the CA value is 0.987 and AUC value is 0.999(nearer to 1). Remaining models SVM with RBF kernel and Naïve Byes perform below 0.887 value of accuracy. The detailed analysis is shown in the table 5.

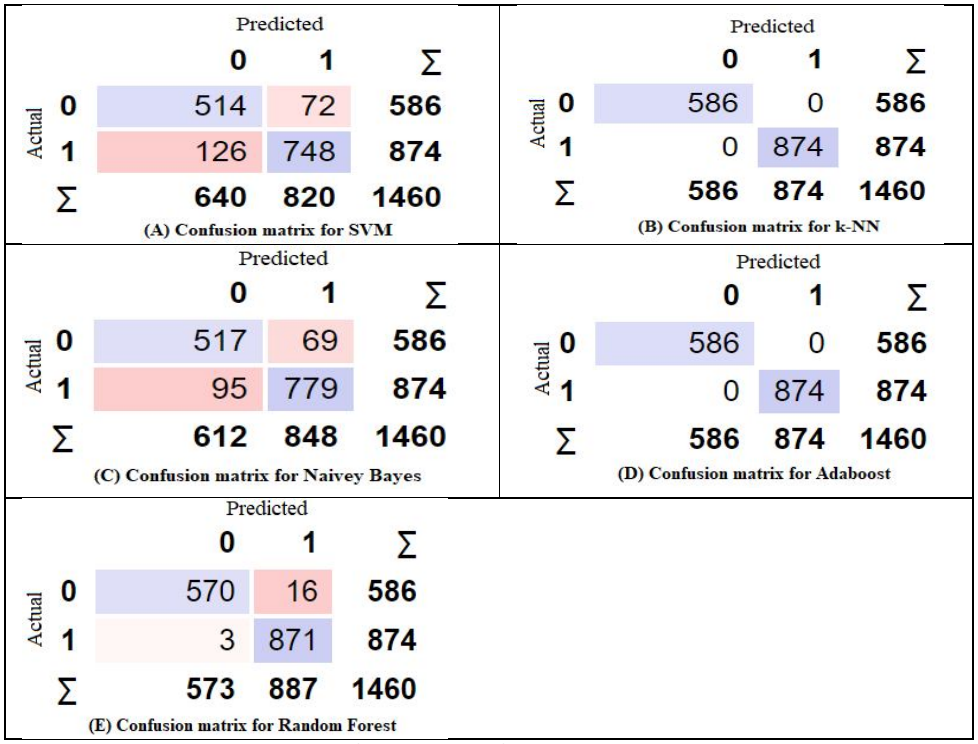


**Figure 4:** Confusion matrices for experimental ML models

**Table 5.**Accuracy specifies for the APLD Dataset on target class NLD

| Model | AUC | CA | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| k-NN | 1 | 1 | 1 | 1 | 1 |
| SVM | 0.950791 | 0.864384 | 0.838499 | 0.803125 | 0.877133 |
| Random Forest | 0.999195 | 0.986986 | 0.983607 | 0.994764 | 0.972696 |
| Naive Bayes | 0.959648 | 0.887671 | 0.863105 | 0.844771 | 0.882253 |
| AdaBoost | 1 | 1 | 1 | 1 | 1 |

#### 4.3.2.2 Target Class LD classification Analysis:

Table 6 shows the class LD performance parameters analysis. In this, the class LD classifies 100% by the k-NN and AdaBoost algorithms where CA (classification accuracy) and AUC values are 1. The random forest model also performs better way where the CA value is 0.987 and AUC value is 0.999(nearer to 1). Remaining models SVM with RBF kernel and Naïve Byes perform below 0.887 value of accuracy. The detailed analysis is shown in the table.

**Table 6.**Accuracy specifies for the APLD Dataset on target class LD

| Model | AUC | CA | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| k-NN | 1 | 1 | 1 | 1 | 1 |
| SVM | 0.950791 | 0.864384 | 0.883117 | 0.912195 | 0.855835 |
| Random Forest | 0.999194 | 0.986986 | 0.989211 | 0.981962 | 0.996568 |
| Naive Bayes | 0.959648 | 0.887671 | 0.904762 | 0.918632 | 0.891304 |
| AdaBoost | 1 | 1 | 1 | 1 | 1 |

#### 4.3.2.3 Average Classes LD and NLD classification Analysis:

Table 7 shows the average classesof LD and NLD performance parameters analysis. In this, the average of classes LD and NLD classifies 100% by the k-NN and AdaBoost algorithms where CA (classification accuracy) and AUC values are 1. The random forest model also performs better way where the CA value is 0.986 and AUC value is 0.999(nearer to 1). Remaining models SVM with RBF kernel and Naïve Byes perform below 0.887 value of accuracy.

**Table 7.**Accuracy specifies for the APLD Dataset on average target classes LD and NLD

| Model | AUC | CA | F1-Score | Precision | Recall |
|---|---|---|---|---|---|
| k-NN | 1 | 1 | 1 | 1 | 1 |
| SVM | 0.950791 | 0.864384 | 0.865209 | 0.868418 | 0.864384 |
| Random Forest | 0.999194 | 0.986986 | 0.986961 | 0.9871 | 0.986986 |
| Naive Bayes | 0.959648 | 0.887671 | 0.888042 | 0.888987 | 0.887671 |
| AdaBoost | 1 | 1 | 1 | 1 | 1 |

### 4.4 Receive Operator Characteristic (ROC) Curves

The figure 5 shows the ROC curves of the experimental ML Models on target class 0 (Non-Liver Disease). Each ML model ROC specified with different colors. The AdaBoost and k-NN model's AUC values are one and indication colors are dark green and orange. The Random Forest Tree specified with pink color and the AUC value is 0.999195 nearer to one. The Naïve Bayes model indicated with violet color and AUC values is 0.959648 better then SVM model (AUC value is 0.95079).

The figure 6 shows the ROC curves of the experimental ML Models on target class 1 (Liver Disease). Each ML model ROC specified with different colors. The AdaBoost and k-NN model's AUC values are one and indication colors are dark green and orange. The Random Forest Tree specified with pink color and the AUC value is 0.999194 nearer to one. The Naïve Bayes model indicated with violet color and AUC values is 0.959648 better than SVM model (AUC value is 0.95079).

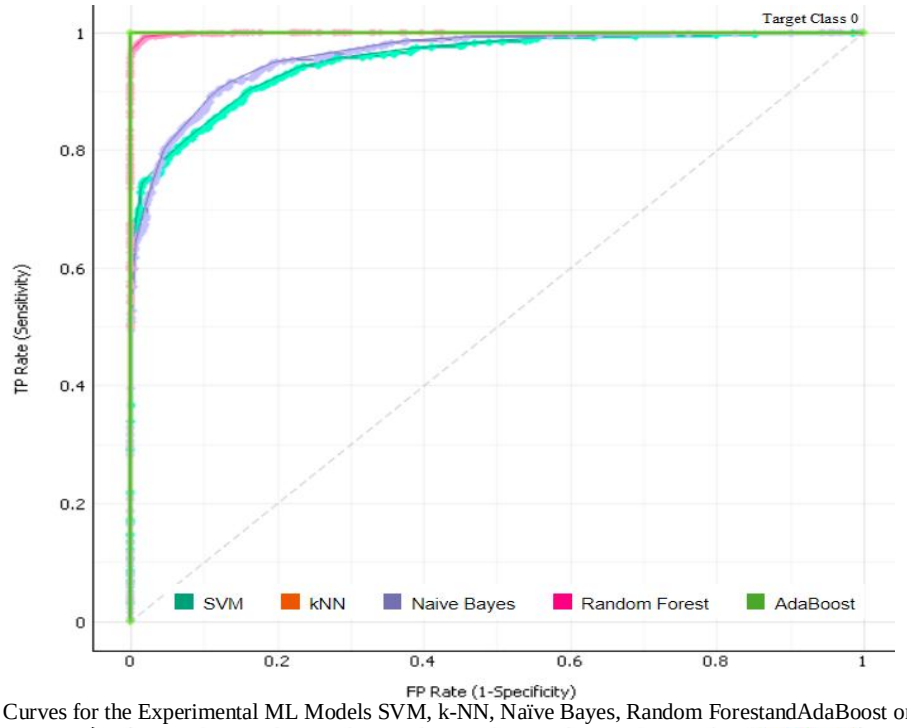


**Figure 5:** ROC Curves for the Experimental ML Models SVM, k-NN, Naïve Bayes, Random ForestandAdaBoost on target class 0

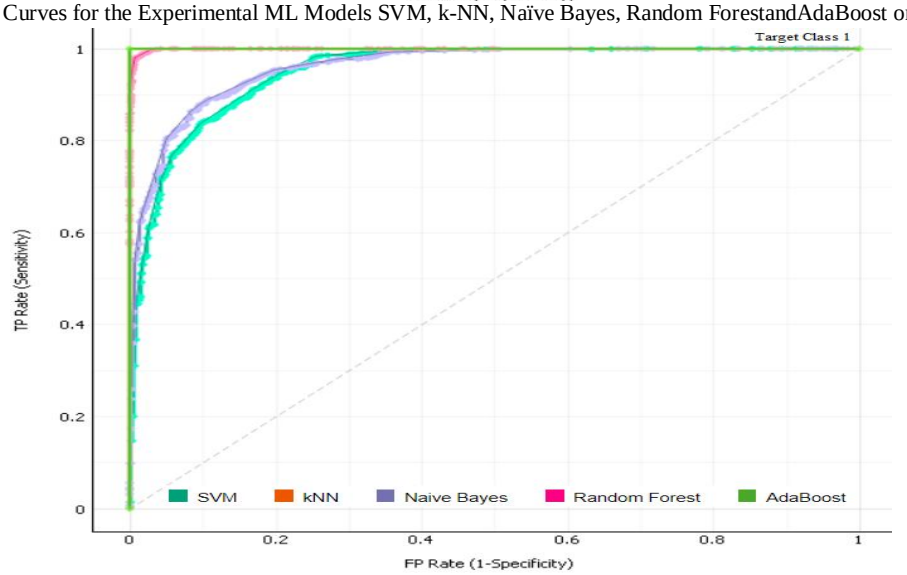


**Figure 6:** ROC Curves for the Experimental ML Models SVM, k-NN, Naïve Bayes, Random ForestandAdaBoost on target class 1

### 4.5 ML Models Comparative Analysis

The figure 7 shows the comparative accuracy analysis of the experimental ML models. As per analysis, the k-NN and AdaBoost algorithms are very accurate to predict the Liver Diseases where the classification accuracy (CA) and AUC values are one. All experimental models are performed well with above 0.86 CA and 0.95 AUC values. The random forest algorithm is also well noted with 0.986 CA and 0.999 AUC value.

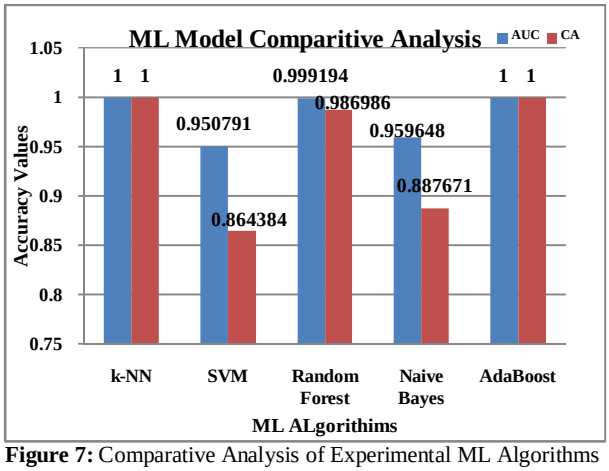


**Figure 7:** Comparative Analysis of Experimental ML Algorithms

### 5. Conclusion

Protection, prediction and preventions are very necessary for any health caresystem. Liver disease (LD) is common disease in the world. For the protection and preventions, statistical analysis role is very crucial. In this research, empirical statistical analysis about LD patients is in detailed with personal as well as clinical information. As per observation, we found that life style of patient and habits are main factors for occurring LD. We also observed some symptoms like vomiting, fever and head/bone ache indication of LD. In clinical tests of patients, most correlated attributes were Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase and Alanine Aminotransferase for prediction of LD. As per ML performance analysis, the K-NN and AdaBoost models were very accurate. So, this model will very useful to predict the LD with APLD data set. Further, we will researched on LD with Deep Learning models for identification of LD with scanning and CAD images

### REFERENCES


1. Dhingra, S., Singh, I., Subburaj, R., &Diwakar, S. ANN Model for Liver Disorder Detection. In Advances in Data Sciences, Security and Applications, 161-167, 2020.
2. El-Shafeiy, E. A., El-Desouky, A. I., &Elghamrawy, S. M. Prediction of Liver Diseases Based on Machine Learning Technique for Big Data. In International Conference on Advanced Machine Learning Technologies and Applications, 362-374, 2018.
3. Książek, W., Abdar, M., Acharya, U. R., &Pławiak, P. A novel machine learning approach for early detection of hepatocellular carcinoma patients. Cognitive Systems Research, 54, 116-127, 2019.
4. Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251, 2017.
5. Gatos, I., Tsantis, S., Spiliopoulos, S., Karnabatidis, D., Theotokas, I., Zoumpoulis, P. &Kagadis, G. C. A machine-learning algorithm toward color analysis for chronic liver disease classification, employing ultrasound shear wave elastography. Ultrasound in medicine & biology, 43(9), 1797-1810, 2017.
6. Bean, R. B., & Baker, W. Some racial characteristics of the liver weight in man. American Journal of Physical Anthropology, 2(2), 167-173, 1919.
7. Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N. & Li, Y. C. J. Prediction of fatty liver disease using machine learning algorithms. Computer methods and programs in biomedicine, 170, 23-29, 2019.
8. Ansari, S., Shafi, I., Ansari, A., Ahmad, J., & Shah, S. I. Diagnosis of liver disease induced by hepatitis virus using artificial neural networks. In 2011 IEEE 14th International Multitopic Conference, 8-12. IEEE, 2011.
9. Lee, C. C., Chen, S. H., & Chiang, Y. C. (2007). Classification of liver disease from CT images using a support vector machine. Journal of Advanced Computational Intelligence and Intelligent Informatics, 11(4), 396-402, 2007.
10. Sinnatamby, C. S. Last's Anatomy e-Book: Regional and Applied. Elsevier Health Sciences, 2011.
11. Bernal, W., Auzinger, G., Dhawan, A., &Wendon, J. Acute liver failure. The Lancet, 376(9736), 190-201, 2010.
12. Rueff, B., &Benhamou, J. P. Acute hepatic necrosis and fulminant hepatic failure. Gut, 14(10), 805, 1973.
13. Lee, W. M. Acute liver failure. New England Journal of Medicine, 329(25), 1862-1872, 1993.
14. Iyyanki, M., &Jayanthi, P. Machine Learning for Health Data Analytics: A Few Case Studies of Application of Regression. In Challenges and Applications for Implementing Machine Learning in Computer Vision, 241-270, 2020.
15. Das, S., &Sanyal, M. K. Application of AI and Soft Computing in Healthcare: A Review and Speculation. Vol, 8, 21.
16. Kim, H. C., Pang, S., Je, H. M., Kim, D., & Bang, S. Y. Constructing support vector machine ensemble. Pattern recognition, 36(12), 2757-2767, 2003.
17. Kozma, L. k-Nearest Neighbors algorithm (kNN). Helsinki University of Technology, 2008.
18. Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. Confusion Matrix-based Feature Selection. MAICS, 710, 120-127, 2011.
19. PhaniMadhuri, N., Meghana, A., PrasadaRao, P. V. R. D., &Prem Kumar, P. Ailment prognosis and propose antidote for skin using deep learning. International Journal of Innovative Technology and Exploring Engineering, 8(4), 70-74, 2019.
20. NarasingaRao, M. R., Sajana, T., Bhavana, N., Sai Ram, M., & Nikhil Krishna, C. Prediction of chronic kidney disease using machine learning technique.

Journal of Advanced Research in Dynamical and Control Systems, 10, 328-332, 2018.

21. Razia, S., SwathiPrathyusha, P., Vamsi Krishna, N., &SathyaSumana, N. A comparative study of machine learning algorithms on thyroid disease prediction. International Journal of Engineering and Technology (UAE), 7(2.8), 315-319, 2018.
22. Razia, S., &Narasinga Rao, M. R. A neuro computing frame work for thyroid disease diagnosis using machine learning techniques. Journal of Theoretical and Applied Information Technology, 95(9), 1996-2005, 2017.
23. Shinde, S. A., &Rajeswari, P. R. Intelligent health risk prediction systems using machine learning: A review. International Journal of Engineering and Technology(UAE), 7(3), 1019-1023, 2018.
24. Bommadevara, H. S. A., Sowmya, Y., &Pradeepini, G. Heart disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(5), 270-272, 2019.
25. Srinivas, V., Aditya, K., Prasanth, G., Babukarthik, R. G., Satheeshkumar, S., &Sambasivam, G. A novel approach for prediction of heart disease: Machine learning techniques. International Journal of Engineering and Technology(UAE), 7(2.32), 108-110, 2018.
26. Rajesh, N., Maneesha, T., Hafeez, S., & Krishna, H. Prediction of heart disease using machine learning algorithms. International Journal of Engineering and Technology(UAE), 7(2.32), 363-366, 2018.
27. Sajana, T., &Narasingarao, M. R. Machine learning techniques for malaria disease diagnosis - A review. Journal of Advanced Research in Dynamical and Control Systems, 9(6), 349-369, 2017.
28. Sajana, T., &Narasinga Rao, M. R. A comparative study on imbalanced malaria disease diagnosis using machine learning techniques. Journal of Advanced Research in Dynamical and Control Systems, 10, 552-561, 2018.
29. Deutschmann, C; Sowa, M; Murugaiyan, J; Roesler, U; Rober, N; Conrad, K; Laass, MW; Bogdanos, D; Sipeki, N; Papp, M; Rodiger, S; Roggenbuck, D; Schierack, P. Identification of Chitinase-3-Like Protein 1 as a Novel Neutrophil Antigenic Target in Crohn's Disease, Journal Of Crohn’s& Colitis, 13(7), 894-904, 2019.
30. Sivakumar, S.; Nayak, SoumyaRanjan; Vidyanandini, S.; Kumar, J. Ashok; Palai, G. An empirical study of supervised learning methods for breast cancer diseases, Optik, 175, 105-114, 2018.
31. Raghav, R. S.; Dhavachelvan, P. Bigdata fog based cyber physical system for classifying, identifying and prevention of SARS disease Journal of Intelligent & Fuzzy Systems, 36(5), 4361-4373, 2019.