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

# PREDICTION OF LIVER DISEASE USING MACHINE LEARNING

**[1]Chappa Kushal Kumar, [2]Jaya Shankar, [3]Allen Jacinth, [4]Varun Naidu, [5]Bhupinder Singh**

[1,2,3,4]Student, Department of Computer Science and Engineering , Chandigarh University, Punjab, India
[5]Academic Coordinator, Department of Computer Science and Engineering, Chandigarh University, Punjab, India

***Abstract :*** Successful treatment for liver illness necessitates an early diagnosis. Doctors may find it challenging to diagnose the disease in its infancy due to the disease's weak symptoms. The disease's symptoms typically manifest later. This work use machine learning algorithms for recognizing signs of liver illness in order to address this issue. The main objective of this investigation is to separate heart disease patients from people with good health using the classification method. We intend to construct a graphical user interface in Python to aid in the detection of liver disease in patients by the medical community. The GUI is an easy device for healthcare professionals to make use of for screening for liver disease**.**

***Keywords* – Machine Learning , Liver Patients, Classification Algorithms**

## INTRODUCTION

A medical issue that can turn fatal is liver illness. Hepatitis can be harmful, but if treatment is not received when the illness is still in its early stages, it can be deadly. The primary cause for mortality in many developing countries is this illness. The disease's severity and inadequate detection and therapy are the causes. Early identification of the illness could at times raise concerns. A specialist with a comprehensive understanding of the disease danger and the patient's financial capacity are required for the diagnosis or identification of the illness. It is challenging for clinicians to make a broad assessment of the illness since the symptoms frequently coincide alongside symptoms of other illnesses. Models of machine learning can assist with performing diagnosis the disease. Methods based on machine learning such as SVM and KNN are frequently utilized in predicting the development of liver disease in an attempt to decrease the disease's consequences.

Nowadays, machines give healthcare providers more data, which aids in the precise identification and diagnosis of disorders. A wide range of machine learning methods, such artificial neural networks (ANN), support vector machines (SVM), naive Bayes's (NB), convolutional neural networks (CNN), decision tree algorithms, and K closest neighbours (KNN), are used for developing models for illness prediction [2–12]. To increase the correctness of the disease the diagnosis, advanced prediction models significantly decrease the chance of mortality rates. Early detection of illnesses can be helped by models utilizing automated learning.

Listed below are some challenges with determining liver disease: It might be challenging to correctly identify people with liver disease because its symptoms often overlap with those of other illnesses. 2) If the illness is not identified in its early stages, it intensifies and could result in complications or even death. 3) Utilizing laboratory examinations for determining liver disease Consider through the following issues: The aim is to establish liver illness prediction models using SVM and KNN learning algorithms, which will aid healthcare workers with making options via early liver disease recognition. 2. Will be able to assess the efficacy and accuracy of models utilized for forecasting liver disease. 3. Incorporate precision, accuracy, and confusion.

## LITERATURE REVIEW

The subject matter contains data about predictive machine learning methods for predicting disease of the liver predicting. The authors of [3] indicated an approach based on machine learning for forecasting conditions in the liver utilizing SVM categorisation. In the course of their research, scientific researchers hypothesized that cellular traits were the main root cause of inflammation in the liver. The SVM classification procedure predicts liver illness with greater precision, at 68.75%. Another study [4] employed neural systems (ANN) to create algorithms using machine learning for recognizing signs of liver illness. The SVM method of classification found 71% efficient in predicting liver illness. [5] offers an enhanced diagnostic examine based on a decision tree identification algorithm. The authors depend on the random forest model to construct an illnesses forecasting or diagnosis algorithm. The model was constructed using information retrieved from the University of California, Irvine Machine Learning Repository. For training and evaluation of age, sex, total Bilirubin is direct bilirubin, and overall bilirubin levels in liver patients. The liver illness diagnosis rate was 69.30% utilizing a decision tree distributions learning model. [6] examined multiple states classifying methods for pain mathematical modelling, including naive Bayes (NB), the k- nearest-neighbors method (KNN), and the use of artificial neural networks (SVM). Compared to the artificial neural network (ANN) approach, liver disease grouping is more precise. Based on evaluation of data, we chose SVM to create our liver disease forecast model. The literature has provided an idea for a multidimensional logistical regression (MLP)-based predicting procedure for chronic liver disease. [7]. The Multiple Laboratory Procedure (MLP) approach has been applied to 239 specimens retrieved from medical records. The studies in this article demonstrate that MLP outperforms the NB algorithm. Paper [7] recommended proposed MLP-based liver disease diagnosis model. The MLP method was used to 239 samples drawn from medical files. The research results presented in the present paper indicate that MLP surpasses the NB algorithm. A further investigation hired the following machine learning methods, including logistic regression, KNN, among others + SVM to predict liver disease, and a KNN and SVM-based model was generated. In detecting liver illness, three approaches to classification were compared: the logistic regression method, KNN, among others and SVM [9]. Accuracy ratings and matrices of confusion are used to judge the performance of classification techniques. Range: 71%-73%. In [10], SVM and the algorithm for backpropagation have been used for building a model that can be learned for hepatic disease forecasting. A examination of model accuracy in forecasting reveals that the regression model surpasses SVM, with 73.2% accuracy against 71% SVM accuracy. In fact, SVM surpasses various other categorization techniques used for liver disorders. There is still opportunity to make larger advancements. Data demonstrate that a variety of different factors can contribute to elevate arterial pressure. Human conduct such as cigarette smoking, alcohol consumption, and having a genealogy of your own are all indicators of risk. For the purpose of assisting clinicians diagnose liver illness while minimizing mortality, SVM and NB-based forecasting algorithms have been put forth in the scientific literature [11-14]. The performance of SVM and NB had been assessed by applying different metrics involving consistency score, accuracy, and f- score measurement, and outcomes from the experiment discovered that SVM outperforms NB at more profound anticipatory thinking. In terms of physiological health conditions [12], the technique known as SVM improves machine learning techniques for recognizing symptoms of liver disease. Studies have revealed that this strategy is 73% successful. The authors referenced the Indian Liver Disease Database. The success rate of the model is contingent on the number of strategies for learning made use of to train the kernel of the SVM .The support vector machine, commonly referred a stance has been displayed to be a feasible algorithmic learning tool for putting together disease- centric models from frameworks with adequate degree of precision. There's a caution. The authors also compared their ANN-based model with contemporary models devised with algorithmic learning techniques such as SVM and NB. The corresponding findings establish that the degree of precision of ANN far surpasses 70%. In [15], the authors that have been suggested a system for classification for liver disease prediction that incorporates Naive Bayes and Support Vector Machine. The study's authors sought out the performance and forward-looking performance of naive Bayes algorithms and support vector machines, respectively. The findings from experiments showcase that the support vector machine superior to Naive Bayes for inflammation in the liver categorization. However, SVM has a more prolonged time of processing than it previously was Naive Bayes. The research study [16] predicted a disease classification model using logistic regression. The authors addressed the academic achievements of the liver disease examinations model and recognized that it was 74% credible in liver infectious diseases categorization purposes In their study, the scientists was additionally compared logistic regression's dexterity to other machine learning strategies which incorporates the use of support vector machines and neural networks that are artificial. In which these models' accomplishments underwent scrutiny, logistic regression consistently excelled over support vector machines and artificial neural networks, respectively, in relation to liver disease accuracy in classification. In accordance with the scientific research [1-16], we chose the assistance vector machine algorithms and the number K that are most appropriate at figuring out complications of liver disease.

## IMPLEMENTATION

### i. DATA PRE-PROCESSING

Prioritizing data is an important step in solving any machine learning problem. Often data used for machine learning problems needs to be processed/cleaned/transformed so that machine learning algorithms can be trained on them. Assigning zero values, coding categorical variables, scaling, etc. There are several commonly used preprocessing techniques such as: This process is easy to understand. But once we're done with the data, things seem to get cumbersome. Each case is different and presents unique challenges. All attributes except gender are real numbers. The last row is the "disease" label ("1" represents the presence of disease, "2" represents the absence of disease). The total number of data points was 583, of which 416 were collected for liver patients and 167 for non-liver patients. In the description of this data, it was determined that some values of the Albumin and Globulin Ratio column were empty. Rows containing null values are replaced with the middle value of the row.

### ii. CLASSIFICATION TECHNIQUES

Prioritizing data is an important step in solving any machine learning problem. Often data used for machine learning problems needs to be processed/cleaned/transformed so that machine learning algorithms can be trained on them. Assigning zero values, coding categorical variables, scaling, etc. There are several commonly used preprocessing techniques such as: This process is easy to understand. But once we're done with the data, things seem to get cumbersome. Each case is different and presents unique challenges. All attributes except gender are real numbers. The last row is the "disease" label ("1" represents the presence of disease, "2" represents the absence of disease). The total number of data points was 583, of which 416 were collected for liver patients and 167 for non-liver patients. In the description of this data, it was determined that some values of the Albumin and Globulin Ratio column were empty. Rows containing null values are replaced with the middle value of the row.

#### a) SVM

The purpose of SVM is to find the best hyperplane that divides the data into various groups. The scikit-learn package in Python is used to implement SVM. Preliminary data is divided into testing and training data and constitutes 25% and 75% of the total data set. Support vector machines create a hyperplane or a series of hyperplanes in high-dimensional or infinite-dimensional space. Good separation is achieved by the overall plane with the largest distance to the nearest training data point of each class (known as the good function), because the larger the margin, the lower the error of the classifier.

#### b) LOGISTIC REGRESSION

Among the simplest models of categorization is the logistic regression model. It is useful for developers trying to find links between variables because of its parametric nature, which allows for some interpretability by looking at the variables. A parameter vector = (0, 1... p) can fully represent a parametric model. A parametric model is typically represented by the line y = kx + m, where k = m. Custom sizes can be used to generate any design. A parametric model with predictor variable coefficients labelled as interval 0, and parameters expressed as 0 +1 + X1 +... PXp, is what is embodied in a logistic regression. We represent the variable number above in X as a vector for simplicity's sake. Although the regression model utilised, the name "logistic regression" may be a little misleading.

#### c) K-NN

This section covers the implementation specifics of the KNN algorithm. The whole KNN model is trained using the provided data. When a prediction is required for unknown data, the KNN method locates k related occurrences in the dataset. Predictive features from these related occurrences are then collected and returned as forecasts for unseen events. Similarity measurement varies by data type; Euclidean distance is usually employed for numerical data, but Hamming distance might be used for categorical or binary data. The method known as KNN belongs to the group of instance-based and lazy learning algorithms. Instance-based algorithms model problems and develop predictions by analysing samples or data points. The KNN method works particularly well for sampling because the model continues to incorporate all training observations. The model functions as a dynamic learning algorithm, using internal rivalry between data samples to make predictions. This approach employs an objective similarity assessment of profiles, allowing each profile instance to compete for similarity with the unseen profile and thereby contribute to the predictions phase.

#### d) Artificial Neural Networks

A the backpropagation neural network was created, using 10 input neurons in the data entry layer to represent the entire number of characteristics in the file. The input approach uses the rectified linear unit activation function, but the output method uses a sigmoid function to incorporate a set. The goal is to find an adequate amount of patients with proven liver illness. To attain

the desired results, changes to the neural network model are required, notably in the areas of learning rate, momentum rate, and hidden neurons. These parameters are essential elements of the backpropagation neural network. The rate of learning regulates the capacity of the system to learn, whereas the acceleration rate determines the system's learning speed. In order to achieve superior outcomes, To improve the output of the result production process, the number of neurons in the latent process is carefully examined to find the optimal configuration that accurately represents the input data. The hidden layers neuron count was determined, and the output layer employed the sigmoid function for smooth activation without complex derivatives. Pythons Keras library and the TensorFlow backend were utilized to construct the neural network. the table below outlines the features of the backpropagation neural network.

## MATERIAL AND METHODS

This chapter discusses the application of machine learning techniques, the datasets used to develop and test liver disease prediction models, the programming languages used for model development, and the experimental methodologies used. An application that teaches and assesses intricate models for disease forecasting is called Indian Epidemiology Online. This sentence can be rephrased as follows: Machine learning techniques such as SVM and KNN are utilized to develop models that tackle classification issues related to liver disorders. SVM and KNN algorithms were used in experimental studies to develop a machine learning model for predicting liver disorders, while Python was used for programming. The major techniques for predicting liver illness are machine learning, support vector machines, and KNN are shown in Figure 1.

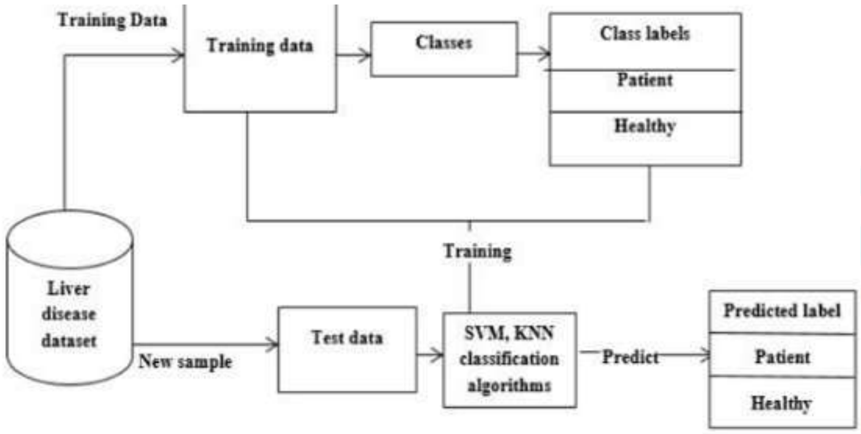
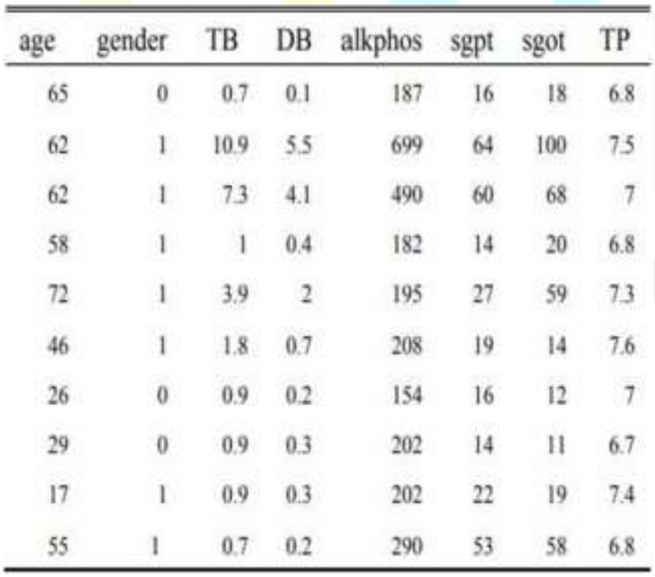


fig 1 SVM and KNN model for liver disease detection

### i) DATA SET DESCRIPTION

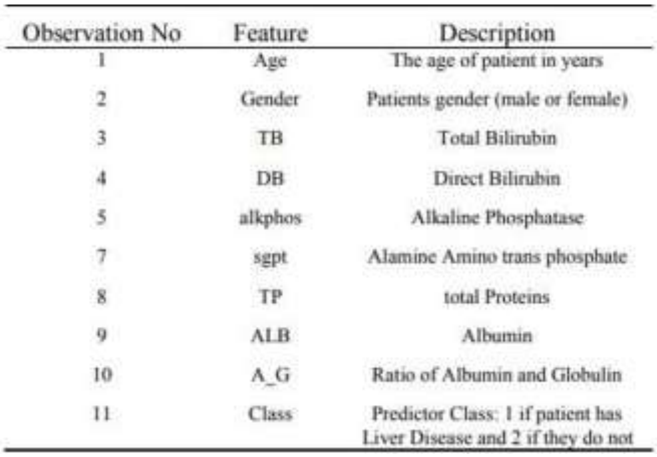
An overview of the information in the online True Indian Liver Disease Database may be seen in figure 2. This file, which has a total of 583 records and 11 attributes, is separated into two groups or labels: innocuous liver disease (patients with liver illness are represented by figure 2) and malignancy liver disease (patients with liver disease are indicated by figure 3). Figure 3 displays albumin, age (years), gender (male or female), and other parameters.

There are 416 interested parties. Out of the entire 583 samples within the storage facility, 117 samples, or 20% of the total number of samples, are used for testing, and 466 samples, or the remaining 80%, are utilised for training the model. The characteristics of the data are provided in figure 2.



| age | gender | TB | DB | alkphos | sgpt | sgot | TP |
|---|---|---|---|---|---|---|---|
| 65 | 0 | 0.7 | 0.1 | 187 | 16 | 18 | 6.8 |
| 62 | 1 | 10.9 | 5.5 | 699 | 64 | 100 | 7.5 |
| 62 | 1 | 7.3 | 4.1 | 490 | 60 | 68 | 7 |
| 58 | 1 | 1 | 0.4 | 182 | 14 | 20 | 6.8 |
| 72 | 1 | 3.9 | 2 | 195 | 27 | 59 | 7.3 |
| 46 | 1 | 1.8 | 0.7 | 208 | 19 | 14 | 7.6 |
| 26 | 0 | 0.9 | 0.2 | 154 | 16 | 12 | 7 |
| 29 | 0 | 0.9 | 0.3 | 202 | 14 | 11 | 6.7 |
| 17 | 1 | 0.9 | 0.3 | 202 | 22 | 19 | 7.4 |
| 55 | 1 | 0.7 | 0.2 | 290 | 53 | 58 | 6.8 |

fig 2 sample liver dataset used in classification



| Observation No | Feature | Description |
|---|---|---|
| 1 | Age | The age of patient in years |
| 2 | Gender | Patients gender (male or female) |
| 3 | TB | Total Bilirubin |
| 4 | DB | Direct Bilirubin |
| 5 | alkphos | Alkaline Phosphatase |
| 7 | sgpt | Alamine Amino trans phosphate |
| 8 | TP | total Proteins |
| 9 | ALB | Albumin |
| 10 | A_G | Ratio of Albumin and Globulin |
| 11 | Class | Predictor Class: 1 if patient has Liver Disease and 2 if they do not |

fig 3 liver dataset feature description

## ii) RESULTS AND DISCUSSIONS

In an experimental investigation, three performance measures were analysed to compare K-neighbour and support vector systems, and the training model was tested on liver samples. Accuracy scores, confusion matrices, and ROC (receiver operating characteristic) curves were among the performance metrics employed in the experiment. The findings suggest that SVM outperforms K neighbours.

The accuracy score of a model for learning is the most crucial metric for assessing its efficacy. This statistic assesses how effectively the model matches texts that are unknown to classes that are known. The working model was assigned a label. The effectiveness of liver tests according to SVM and KNN learning algorithms is assessed using corrective measures. After testing the model on random samples, the research revealed that it was 74.52% accurate in five Using the KNN approach, the accuracy of the Random SVM test was 70.93%. The accuracy of the two models used in a randomised experiment can be seen in Figure 4. Support vector machine (SVM) beat K-nearest neighbour (KNN) method for random evaluation of the model (Figure 4).

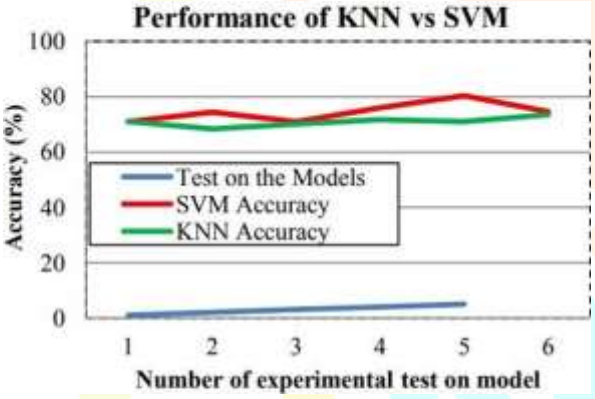


fig 4 performance of knn vs svm

1) Our study's primary goal is to forecast liver disorders using a variety of machine learning techniques. For prediction, we employ neural networks, logistic regression, K-nearest neighbour (K-NN), and vector machines (SVM). Everyone was more dependable. It is evident that the following definitions apply to each algorithm: accuracy, precision, sensitivity, and specificity.

$$Accuracy = \frac{no.of\ TP + no.of\ TN}{no.of\ TP + FP + FN + TN}$$

fig 5 formula for accuracy

$$Specificity = \frac{no.of\ TN}{no.of\ TN + FP}$$

fig 8 formula for specificity

$$Sensitivity = \frac{no.of\ TP}{no.of\ TP + no.of\ FN}$$

fig 6 formula for sensitivity

$$Precision = \frac{no.of\ TP}{no.of\ TP + FP}$$

fig 7 formula for precision

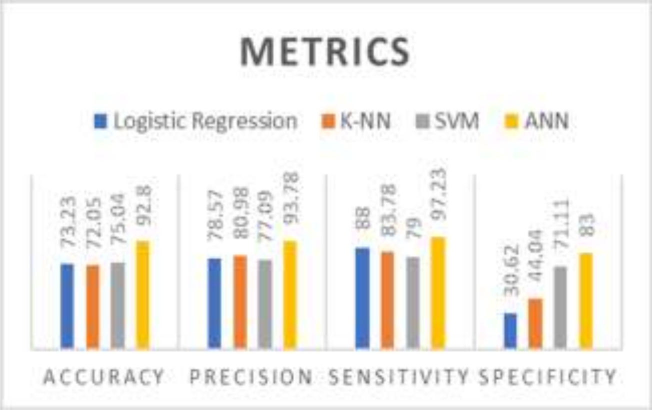


fig 9 bar graph represents the evaluation metrics

2) Receiver Operating Characteristic Curve (ROC)

CNN and SVM models have been contrasted using receiver operating characteristic curves especially patient categories, such as positive (TP) or patients with elevated risk. Since this is the main health category in which we want the model to perform well, the ROC metric is crucial for evaluating how well SVM and KNN models predict patients in this category of patients.

Figure 10 displays the receiver operating characteristic, or ROC curve, of the support vector machine. The rate of true positives (TPR) and the rate of false positives (FPR) are used in this test to evaluate the support vector machine's performance. TPR is the y- axis and FPR is the x-axis in Figure 10, indicating that higher the ROC curve area, the better the model. For those with and without liver disease, the area under the curve is 0.78. The amount shown is much higher than the KNN-ROC curve's centre in the vicinity of the model displayed in Figure 11. The ROC curve is steep in contrast with the support vector machine (ROC) curve and vertical in comparison to the KNN ROC curve, as seen in Figure11. In addition, the area under the curve of the KNN model is less than that of the SVM model. Consequently, in comparison to the SVM model, the KNN model did not perform well in forecasting quality of life (TP) or liver disease.

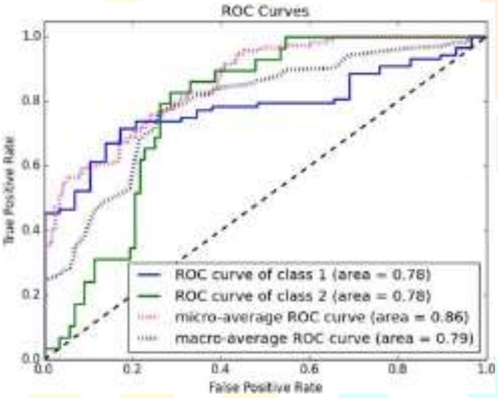


fig 10 roc curve of svm model

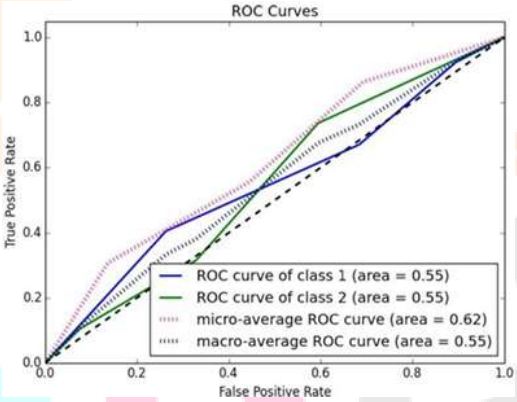


fig 11 roc curve of knn model for liver disease prediction

3) Confusion matrix for KNN, SVM, Logistic, ANN on liver disease classification

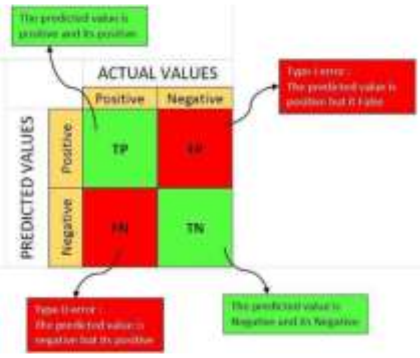

The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type I error : The predicted value is positive but it False
Type II error : The predicted value is negative but its positive
The predicted value is Negative and its Negative

fig 12 diagramatical representation of

confusion matrix

In the picture below, the confusion matrix displays the analysis of real versus projected liver disease data using KNN and SVM, ANN, and logistic regression. The performance of logistic regression is 73.23%, the success rate of K-NN is 72.5%, the performance of SVM is 75%, and the performance of ANN is 92% compared to all other models.

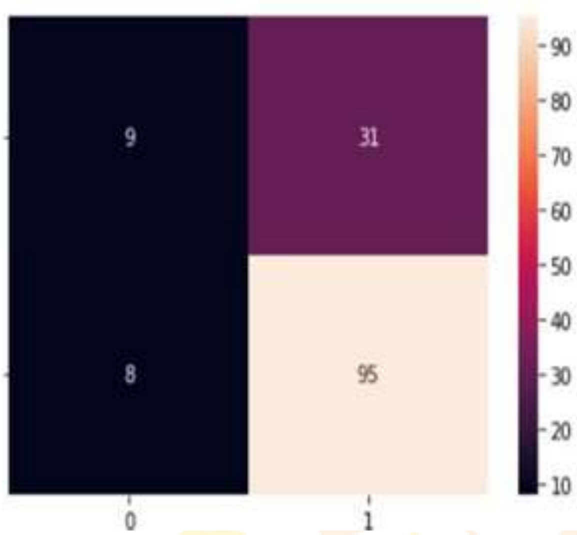


fig 13 confusion matrix of logistic regression

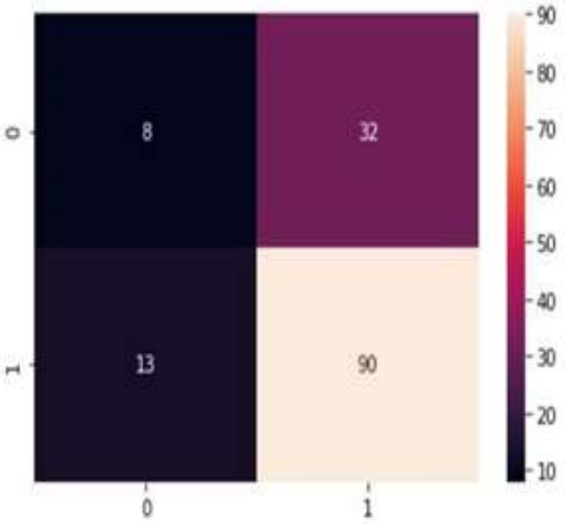


fig 14 confusion matrix of knn

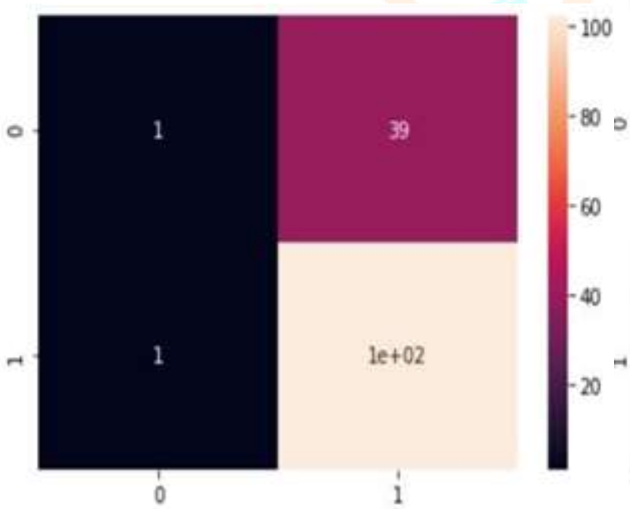


fig 15 confusion matrix of svm

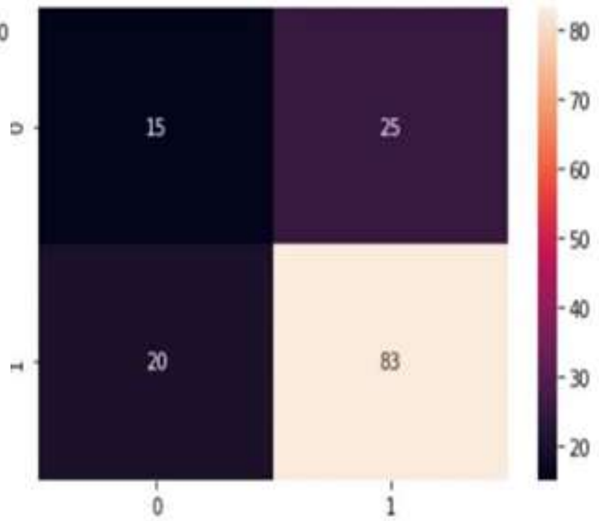


fig 16 confusion matrix of ann

## CONCLUSION

In order to predict liver illness, this study looked into four machine learning algorithms: logistic regression, K nearest neighbour (KNN), artificial neural networks (ANN), and support vector machine (SVM). We examine machine learning strategies using test data that is encoded by real and fuzzy matrices. The accuracy of the model for forecasting liver disease was demonstrated by the subsequent results: neural network was 92.8%, logistic regression was 73.23%, KNN was 72.05%, and SVM was 75.04%. In general, liver disease was more accurately predicted by the SVM algorithm. Neural connections are highly effective, as proved by contrasting the results of this research with past research.

## REFERENCES

[1] Tsehay Admassu Assegie, Pramod Sekharan Nai, Handwritten digits recognition with decision tree classification: a machine learning approach, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019.

[2] Yi-ming Lei, Xi-mei Zhao, Wei-dong Guo, Cirrhosis Recognition of Liver Ultrasound Images Based on SVM and Uniform LBP
Feature, IEEE, 2015.

[3] Sumedh Sontakke, Jay Lohokare, Reshul Dani, Diagnosis of Liver Diseases using Machine Learning, 978-1-5090- 3404-8/17/$31.00, IEEE, 2017.

[4] Nazmun Nahar, Ferdous Ara, International
Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018.

[5] Shambel Kefelegn, Pooja Kamat, Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey, International Journal of Pure and Applied Mathematics Volume 118 No. 9 2018.
[6] Yuan Cao, Zhi-De Hu, Xiao-Fei Liu, An-Mei Deng, Cheng- Jin Hu, An MLP Classifier for Prediction of HBV-Induced Liver Cirrhosis Using Routinely Available Clinical Parameters, Hindawi Publishing Corporation Disease Markers Volume 35, 2013.
[7] A Novel Computer-Aided Diagnosis Framework Using Deep Learning for Classification of Fatty Liver Disease in Ultrasound Imaging, 20th
International Conference on e-Health Networking,
Applications and Services (Healthcom), IEEE, 2018.
[8] Kanza Hamid, Amina Asif, Machine Learning with Abstention for Automated Liver Disease Diagnosis. International Conference on Frontiers of Information Technology, IEEE. 2017.
[9] Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury, Prediction of Liver Disease using Classification Algorithms, 4th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018.
[10] Sumedh Sontakke, Jay Lohokare, Reshul Dani, Diagnosis of Liver Diseases using Machine
Learning, International Conference on Emerging Trends & Innovation in ICT (ICEI) Pune Institute of Computer Technology, Pune, India, Feb 3-5, IEEE, 2017.
[11] Dr. S. Vijayaran, Mr.S.Dhayanand, Liver Disease
Prediction using SVM and Naïve Bayes
Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.
[12] Esraa M. Hashem, Mai S. Mabrouk, A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis, American Journal of
Intelligent System.
[13] Ebenezer Obaloluwa Olaniyi, Khasman Adan, Liver Disease Diagnosis Based on Neural Networks, Advances in Computational
Intelligence, 2017
[14] Assegie Tsehay Admassu, A support vector based heart disease prediction, journal of software engineering and intelligent systems December 2019.
[15] Syed Hasan Adil1, Mansoor Ebrahim, Kamran Raza, Liver Patient Classification using Logistic Regression, 4th International Conference on Computer and Information Science, IEEE, 2018. [16] Esraa M.Hashem, Mai S. Mabrouk, A Study of support vector machine algorithm for liver disease diagnosis, Biomedical Engineering, Misr University for Science and Technology (MUST University), 6th of October, Egypt, 2019.