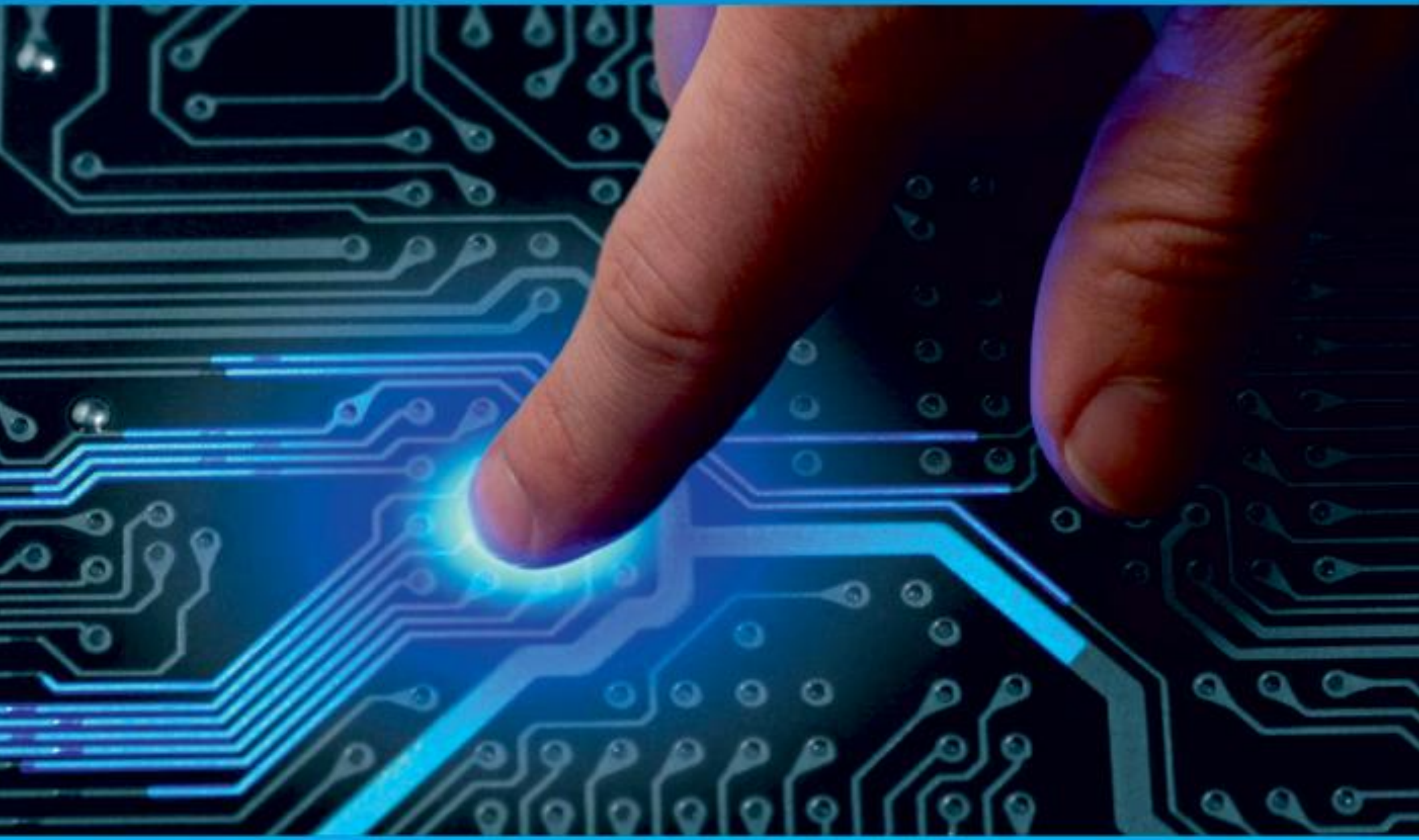
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204208 |

Liver Disease Diagnosis Using ML & DL Algorithms

VK.Sreedhar, K.Lahari, C.Rahithya, P.Pushpanjali, G.Jyothirmai, C.Pavan

Assistant Professor, Dept. of ECE, GATE Institute of Technology, Gooty, India

UG Student, Dept. of ECE, GATE Institute of Technology, Gooty, India

UG Student, Dept. of ECE, GATE Institute of Technology, Gooty, India

UG Student, Dept. of ECE, GATE Institute of Technology, Gooty, India

UG Student, Dept. of ECE, GATE Institute of Technology, Gooty, India

UG Student, Dept. of ECE, GATE Institute of Technology, Gooty, India

ABSTRACT: Around a million deaths occur due to liver diseases globally. There are several traditional methods to diagnose liver diseases, but they are expensive. Early prediction of liver disease would benefit all individuals prone to liver diseases by providing early treatment. As technology is growing in health care, machine learning significantly affects health care for predicting conditions at early stages. This study finds how accurate machine learning is in predicting liver disease. This present study introduces the liver disease prediction (LDP) method in predicting liver disease that can be utilised by health professionals, stakeholders, students and researchers. Five algorithms, namely Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (K-NN), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART), are selected. The accuracy is compared to uncover the best classification method for predicting liver disease using R and Python. From the results, K-NN obtains the best accuracy with 91.7%, and the autoencoder network achieved 92.1% accuracy, which is above the acceptable level of accuracy and can be considered for liver disease prediction**.**

KEYWORDS: Liver disease, Machine learning, Prediction, Data analytics, Healthcare, Autoencoders.

I.INTRODUCTION

The liver is one of the most critical organs of the human body. It plays an essential role in the body’s function. Primary purposes include removing toxins from the body, fighting against infections, and balancing the hormones and secretion of bile juice (Devikanniga et al., 2020). If these functions are not performed by the liver correctly, it will result in several complications and liver diseases. Therefore if a virus infects the liver or chemicals that injure the liver are consumed,or the immune system’s dysfunction occurs, severe damage to the liver or malfunctioning may happen, which ultimately might cause death (Nahar & Ara, 2018). Liver disease is one of the most chronic and threatening diseases globally that can cause various side effects if not treated early (Dutta et al., 2022). According to World Health Organization (WHO) report in 2018, the number of deaths due to liver diseases is around one million and ranked 11th in the world with a critical number of fatalities (World Total Deaths, n.d.). As the symptoms of liver diseases cannot be visible until the condition becomes chronic, it is challenging and daunting for medical health professionals to identify liver disease at its early stages (Devikanniga et al., 2020). In addition, the traditional testing methods like sonography, MRI scans and CT scans that are available for detecting liver diseases are expensive and harmful with numerous side effects (Joloudari et al., 2019). Thus, a significant constraint found by health care workers is to predict liver diseases at an early stage, at minimal cost and at the same time provide a better health care system to treat liver diseases. Severe liver diseases include problems with indigestion, dry mouth, pain in the abdomen, skin colour turning yellow, numbness, memory loss and fainting problems (Shaheamlung et al., 2020). Unnoticed at the initial stages, these symptoms are only visible when the disease turns chronic. However, even though the liver is partially infected, it can still function (Devikanniga et al., 2020). Diagnosis of liver diseases can be divided into three stages i.e., the first stage is liver inflammation, the second is liver scarring (cirrhosis), and the final stage is liver cancer or failure. Since these scenarios are present in liver disease, early prediction is significant to provide better health for New Zealanders. If liver disease is diagnosed early, there will be a chance of early treatment and control of deaths due to liver diseases (Arbain & Balakrishnan, 2019). But when the liver fails to function, few treatments are available except liver transplantation (Shaheamlung et al., 2020), which is very expensive, particularly in New Zealand (Hepatitis C, 2021). Apparently, in

New Zealand, 35 - 40% of the population are not diagnosed with Hepatitis C at the early stages because of the asymptomatic behaviour of liver disease. Unfortunately, most of these individuals do not know the risks linked to liver disease. Due to the asymptomatic behaviour and higher costs of liver disease treatment, it is essential to prevent or diagnose early for better treatment. With advancements in biomedical sciences, the health care system has significantly improved by predicting disease using machine learning techniques (El-Shafeiy et al., 2018). Machine Learning algorithms are one of the potential solutions to this problem due to their handling large amounts of data and employing different approaches like classification, association and clustering, which benefits in realistic arbitration of disease prediction (Naseem et al., 2020). There are different learning techniques in ML methods, one of which is supervised learning. Supervised learning techniques use labelled data and map the input and output data. These supervised learning methods are widely used for prediction and classification (Osisanwo et al., 2017). Supervised learning techniques would be appropriate as this research predicts whether the patient has liver disease or has no liver disease. The supervised learning methods used in this study are Support Vector Machine (SVM) (Boser et al., 1992), Naïve Bayes (McCallum & Nigam, 1998), K-Nearest Neighbors (K-NN) (Fix & Hodges, 1951), Classification and Regression Trees (CART) (Breiman et al., 1984), and Linear Discriminant Analysis (LDA) (Kemp, 2003). The main objective of this research is to compare the accuracies using five supervised learning algorithms, i.e., SVM, Naïve Bayes, K-NN, CART, LDA and Autoencoders, for predicting whether the patient has liver disease or not. This study also proposes the liver disease prediction (LDP) method to help relevant stakeholders pursue an effective healthcare strategy. Moreover, this paper examines the techniques that indicate liver diseases at an acceptable level of accuracy and determines the methods that produce the best accuracy. This study selects a single data set of liver patients with five supervised learning techniques that are applied to that data set in R. The accuracy results from other learning techniques are also used to compare the best algorithm for predicting liver diseases. The stakeholders, including doctors, researchers, lab technicians, or companies dealing with healthcare improvements, can use these results to predict liver diseases at a lower cost and provide better health care in liver treatment.

II. LITERATURE REVIEW

In a study conducted by Vijayarani and Dhayanand (2015), the liver disease prediction applied the SVM and Naïve Bayes (using MATLAB 2013 software) on the Indian Liver Patient Records dataset having 583 instances and 11 attributes, with accuracies of 79.66% (SVM) and 61.28% (Naïve Bayes). In their findings, the time taken to execute SVM was 3210ms, almost two times the time taken by Naïve Bayes (i.e., 1670ms), without preprocessing missing values. In addition to the accuracies, they found that SVM had better performance than Naïve Bayes.

Auxilia (2018) made an accurate prediction for liver disease using different ML methods, including SVM, Random Forest, Decision Trees, Artificial Intelligence and Naïve Bayes. The research was conducted using R on the Indian Liver Patient Records dataset, with 583 instances and 11 attributes. The accuracies were obtained from SVM (77%), Random Forest (77%), Decision Trees (81%), Artificial Intelligence (71%), and Naïve Bayes (37%), with the highest accuracy from the Decision Trees algorithm, and least with Naïve Bayes.

Wu et al. (2019) did a prediction analysis on patients having Fatty Liver Disease (FLD). The research collected 700 patient records from New Taipei Hospital, which had screening tests for fatty liver disease; out of 700 patients, 577 records were considered depending on the patient's age and sufficient data. Of those 577 patients, 377 had fatty liver disease, and the remaining had no fatty liver disease.

The dataset contains patient health details of age, gender, systolic and diastolic blood pressure, abdominal girth, glucose level, triglyceride, HDL-C, SGOT-AST, and SGPT-ALT. Synthetic Minority Over-Sampling Technique (SMOTE) was applied at the data preprocessing stage, and normalisation was done. Four ML algorithms, namely Random Forest, Naïve Bayes, Artificial Neural Network and Logistic Regression with 3, 5, and 10-fold cross-validation, were applied in the next step.

In addition to the accuracies, the area under the receiver operating curve for all the algorithms was observed. Random Forest had given the best accuracy with all the cross-validations from all the results. Singh et al. (2020) focused their research on predicting liver disease using different classification methods with feature selection and implementing software for easy prediction. The study was conducted on the Indian Liver Patient Records dataset.

Some attributes were removed during the feature selection phase using the Correlation-based Feature Selection Subset Evaluator with the Greedy Stepwise search method in WEKA. Only five (5) attributes were selected through this method: Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, and Aspartate Aminotransferase.

With this, six different classification methods were applied: Logistic Regression, Naïve Bayes, Sequential Minimal Optimization (SMO), Random Forest, Instant based Classification (IBk), and Logistic Regression has provided the highest accuracy with 74.36%. The least accuracy was produced by Naïve Bayes (55.9%).

Most of the past research concentrated on just the analysis but not the preprocessing part for this Indian Liver Patient Records dataset. So, this research bridges the gap by considering preprocessing as a significant stage in data analysis. Moreover, several other algorithms are also applied in this research.

III. PROPOSED REASERCH METHODOLY

The proposed liver disease prediction (LDP) method used in this research is based on SEMMA (Santos & Azevedo, 2005), which stands for Sample, Explore, Modify, Model, and Assess (Azevedo & Santos, 2008).

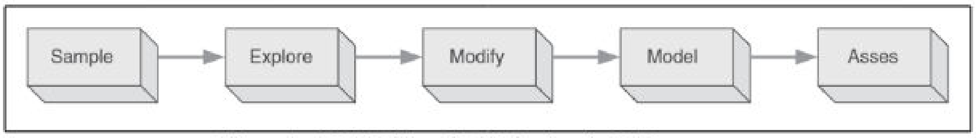


Figure 1: SEMMA lifecycle (Mariscal et al., 2010)

SEMMA lifecycle (see Figure 1) is a simple process to understand, aiming to get the solutions quickly for data mining problems and determine business goals. This methodology has developed by an institute named SAS Institute.

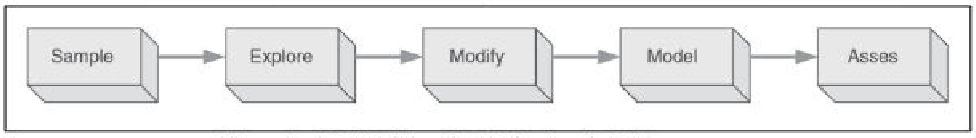


Figure 1: SEMMA lifecycle (Mariscal et al., 2010)

The LDP method involved in this research are Sample, Explore, Modify, Data preprocessing, Model, Assess and Results. Along with these steps from the SEMMA lifecycle, two more steps, Data preprocessing and Results, are added to this research process.

IV. EXPERIMENTAL ENVIRONMENT

Experimental settings and parameter settings :

The data analysis of applying algorithms and finding the accuracy is done using R with version 1.4.1717. The investigation starts by loading the dataset into R and then modifying and preprocessing the data. Then five different algorithms, namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbours (K-NN), and Classification and Regression Trees (CART), are applied to the dataset. For all the algorithms, the seed is set to 7 and a cross-fold validation of 10. For K-NN, the k value is set to 3.

Classifiers :

Support Vector Machine (SVM)

SVM is a supervised machine learning technique that strives to search for a hyperplane with maximum margin. Then it separates the linearly independent variables onto either side of the hyperplane and classifies the data (Devikanniga et al., 2020).

Linear Discriminant Analysis (LDA)

LDA reduces the dimensions at the preprocessing stage and classifies the data. LDA organises the data by mutating the attributes to lower-dimensional space, magnifying the within-class and between-class variance ratios and providing greater class separation (Tharwat et al., 2017).

Naïve Bayes

Naïve Bayes is one of the basic probabilistic classifiers which classifies the specific class with the given tuple. It is categorised by hypothesising that every attribute has a solitary effect on the class attribute by not depending on other attribute values (Passi & Pandey, 2018).

K-Nearest Neighbours (K-NN)

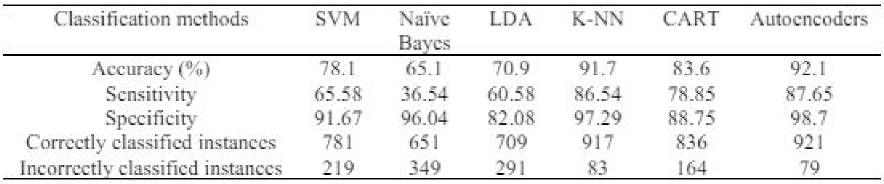
K-NN is one of the most straightforward and efficient classification methods. This method predicts the test data point label with the superior class of its k most identical points of training data (Zhang et al., 2017).

Autoencoder Network

Autoencoder is a special type of Artificial Neural Network that uses an unsupervised approach for learning features, making it more efficient for small datasets with overlapping features. Autoencoders networks (deep and narrow) successfully solved complex classification problems (Tirumala, 2020).

V. EXPERIMENTAL RESULT

After applying different algorithms to the liver disease data set, accuracy, sensitivity, specificity, and confusion matrix are recorded.



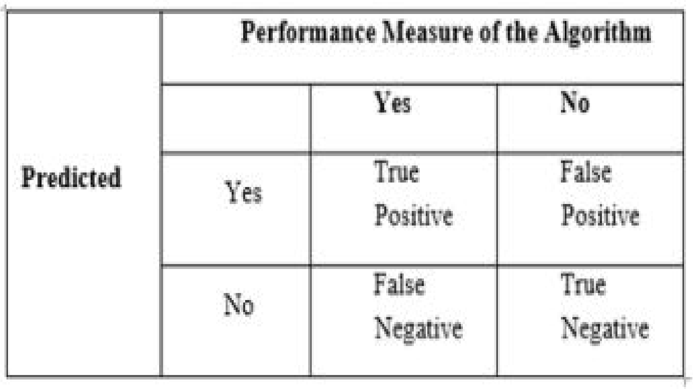
Classification methods	SVM	Naïve Bayes	LDA	K-NN	CART	Autoencoders
Accuracy (%)	78.1	65.1	70.9	91.7	83.6	92.1
Sensitivity	65.58	36.54	60.58	86.54	78.85	87.65
Specificity	91.67	96.04	82.08	97.29	88.75	98.7
Correctly classified instances	781	651	709	917	836	921
Incorrectly classified instances	219	349	291	83	164	79

Results of different experimented algorithms

The results obtained from the experiment, except for two algorithms, SVM and LDA, the rest three algorithms gave an acceptable level of accuracy above 75%. Autoencoders (3 layered) achieved 92.1% (921 correctly classified instance) accuracy, with K-NN achieving an almost similar level of accuracy with correctly classified instances to 917. The lowest accuracy is for Naïve Bayes, 65.1%, with only 651 correctly classified instances.

Confusion Matrix

The confusion matrix is used to anticipate the behavioural structure of supervised learning algorithms. It is a square matrix and represents actual and predicted class values. The rows in the confusion matrix represent the actual values, and the columns represent the predicted values. In binary classification (see Figures 6a and 6b), a 2*2 matrix represents the confusion matrix consisting of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Caelen, 2017).

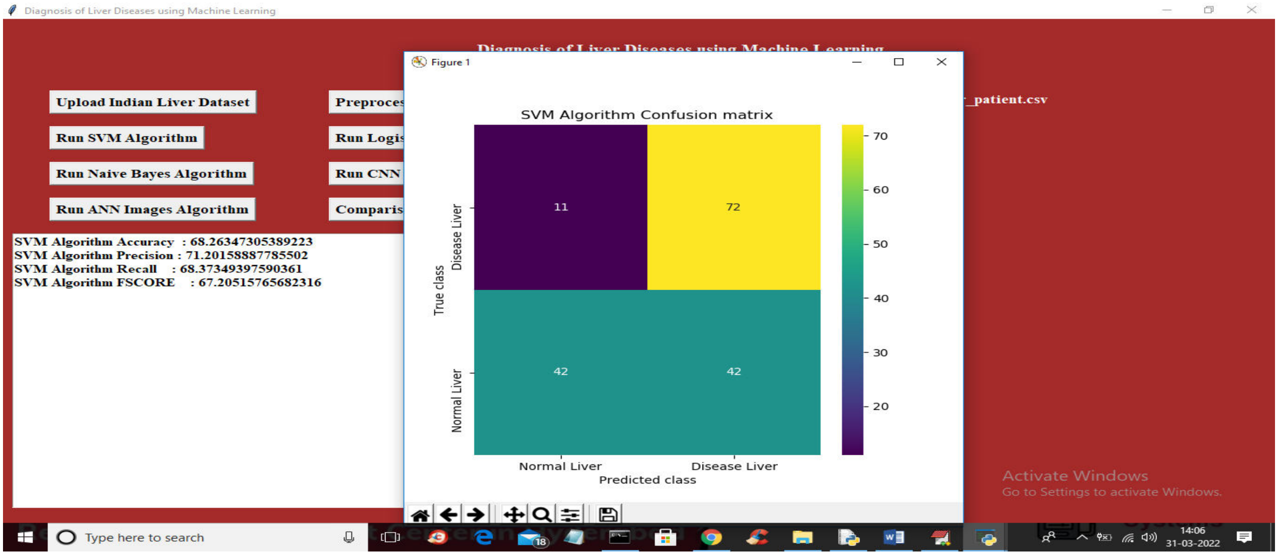


	Performance Measure of the Algorithm		
		Yes	No
Predicted	Yes	True Positive	False Positive
	No	False Negative	True Negative

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

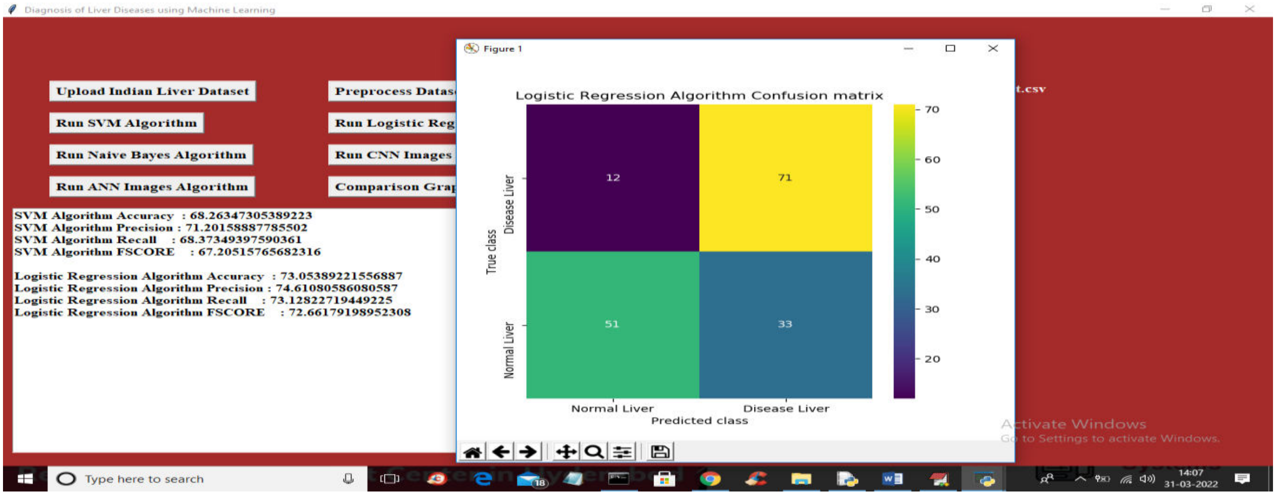
|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204208 |



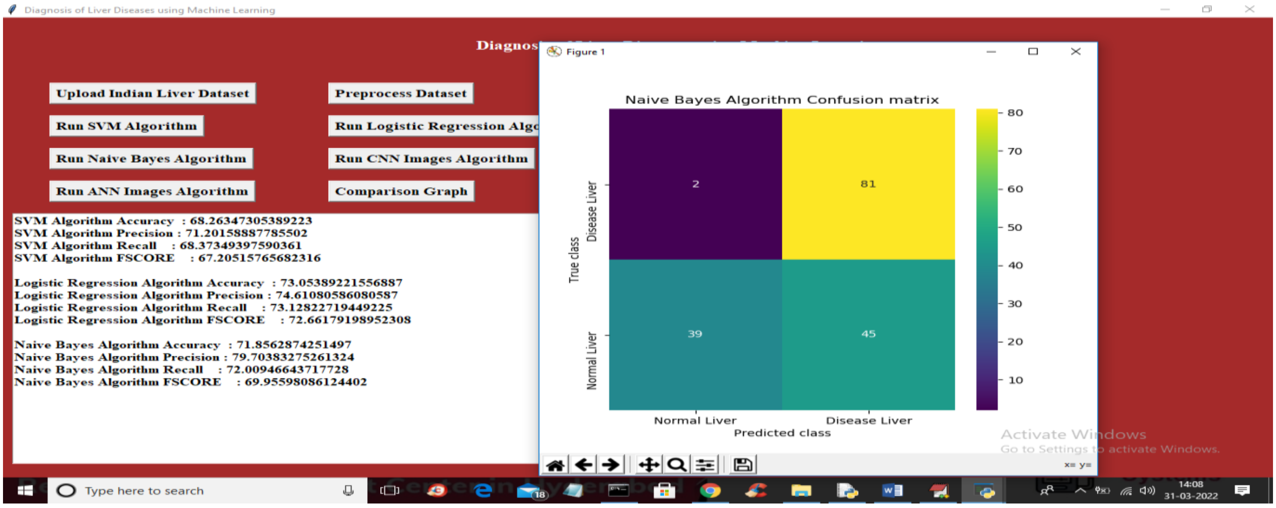
```
SVM Algorithm Accuracy  : 68.26347305389223
SVM Algorithm Precision : 71.20158887785502
SVM Algorithm Recall    : 68.37349397590361
SVM Algorithm FSCORE    : 67.20515765682316
```

SVM BASED CONFUSION MATRIX



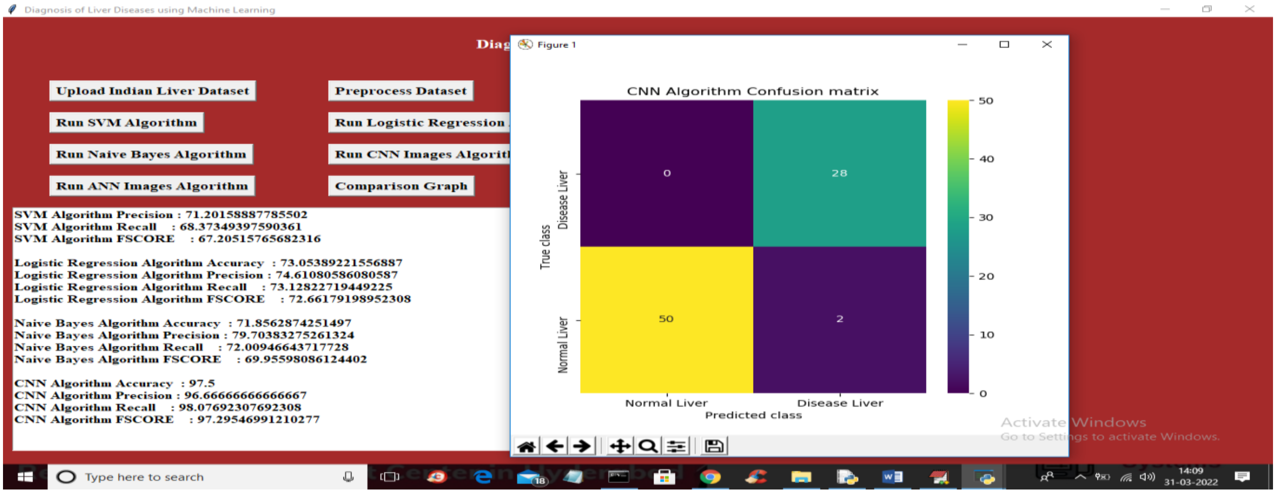
```
SVM Algorithm Accuracy  : 68.26347305389223
SVM Algorithm Precision : 71.20158887785502
SVM Algorithm Recall    : 68.37349397590361
SVM Algorithm FSCORE    : 67.20515765682316

Logistic Regression Algorithm Accuracy  : 73.05389221556887
Logistic Regression Algorithm Precision : 74.61080586080587
Logistic Regression Algorithm Recall    : 73.12822719449225
Logistic Regression Algorithm FSCORE    : 72.66179198952308
```

LR BASED CONFUSION MATRIX



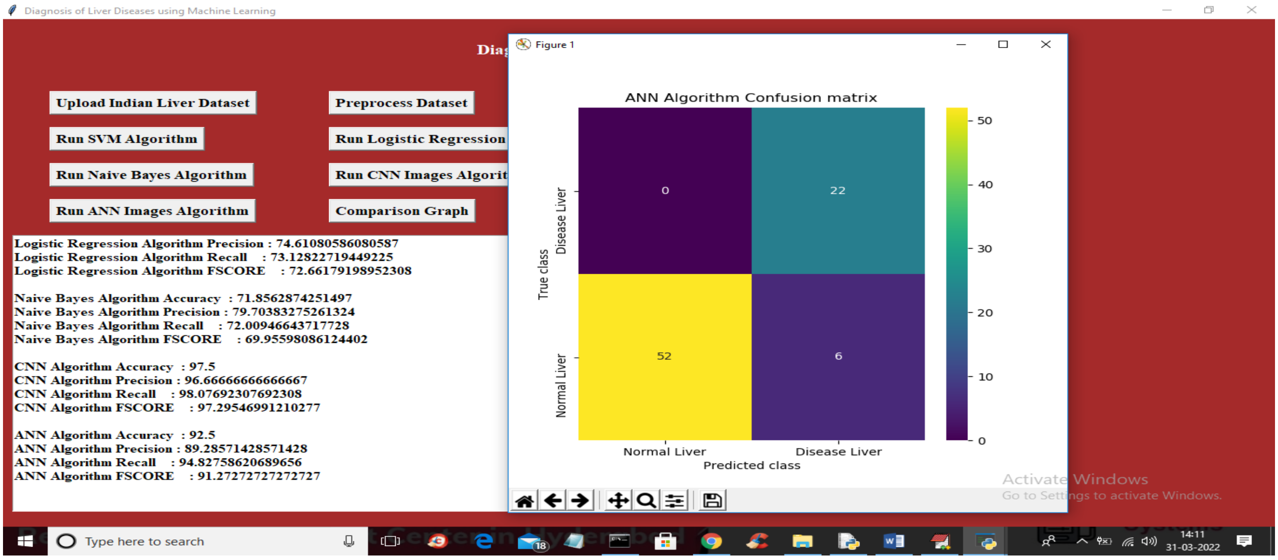
```
SVM Algorithm Accuracy  : 68.26347305389223
SVM Algorithm Precision : 71.20158887785502
SVM Algorithm Recall    : 68.37349397590361
SVM Algorithm FSCORE    : 67.20515765682316

Logistic Regression Algorithm Accuracy  : 73.05389221556887
Logistic Regression Algorithm Precision : 74.61080586080587
Logistic Regression Algorithm Recall    : 73.12822719449225
Logistic Regression Algorithm FSCORE    : 72.66179198952308

Naive Bayes Algorithm Accuracy  : 71.8562874251497
Naive Bayes Algorithm Precision : 79.70383275261324
Naive Bayes Algorithm Recall    : 72.00946643717728
Naive Bayes Algorithm FSCORE    : 69.95598086124402
```

NAVE BAYES BASED CONFUSION MATRIX



```
SVM Algorithm Precision : 71.20158887785502
SVM Algorithm Recall    : 68.37349397590361
SVM Algorithm FSCORE    : 67.20515765682316

Logistic Regression Algorithm Accuracy  : 73.05389221556887
Logistic Regression Algorithm Precision : 74.61080586080587
Logistic Regression Algorithm Recall    : 73.12822719449225
Logistic Regression Algorithm FSCORE    : 72.66179198952308

Naive Bayes Algorithm Accuracy  : 71.8562874251497
Naive Bayes Algorithm Precision : 79.70383275261324
Naive Bayes Algorithm Recall    : 72.00946643717728
Naive Bayes Algorithm FSCORE    : 69.95598086124402

CNN Algorithm Accuracy  : 97.5
CNN Algorithm Precision : 96.66666666666667
CNN Algorithm Recall    : 98.07692307692308
CNN Algorithm FSCORE    : 97.29546991210277
```

CNN BASED CONFUSION MATRIX

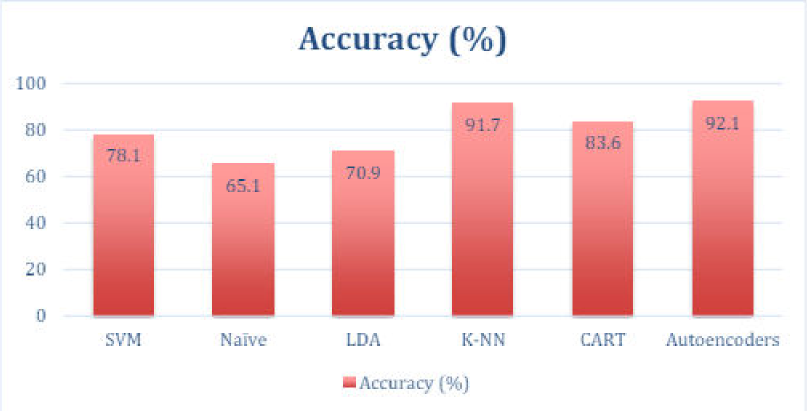


```
Logistic Regression Algorithm Precision : 74.61080586080587
Logistic Regression Algorithm Recall    : 73.12822719449225
Logistic Regression Algorithm FSCORE    : 72.66179198952308

Naive Bayes Algorithm Accuracy  : 71.8562874251497
Naive Bayes Algorithm Precision : 79.70383275261324
Naive Bayes Algorithm Recall    : 72.00946643717728
Naive Bayes Algorithm FSCORE    : 69.95598086124402

CNN Algorithm Accuracy  : 97.5
CNN Algorithm Precision : 96.66666666666667
CNN Algorithm Recall    : 98.07692307692308
CNN Algorithm FSCORE    : 97.29546991210277

ANN Algorithm Accuracy  : 92.5
ANN Algorithm Precision : 89.28571428571428
ANN Algorithm Recall    : 94.82758620689656
ANN Algorithm FSCORE    : 91.27272727272727
```

ANN BASED CONFUSION MATRIX

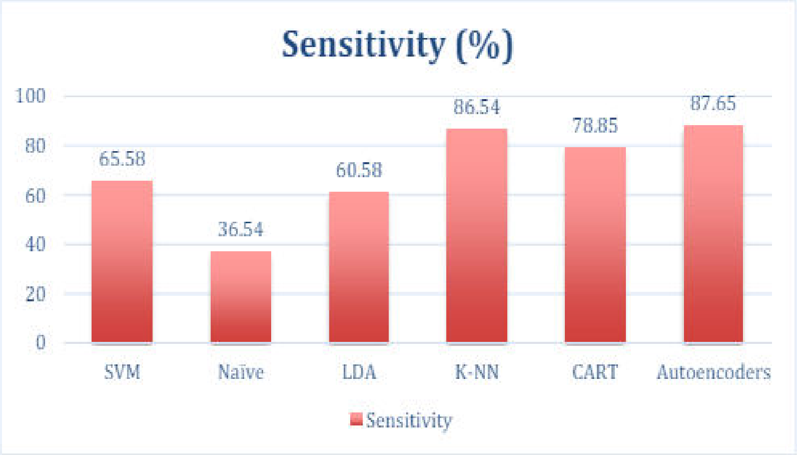
Accuracy

Accuracy is the value of correctly classified instances in both classes (Wu et al., 2019). Accuracy = TN+TP/(TP+FP+TN+FN) The example calculation of accuracy for K-NN = 450+467/ (1000) = 0.917. So, the accuracy of KNN is 91.7%. The rest of the accuracies for other algorithms is shown in Figure below.



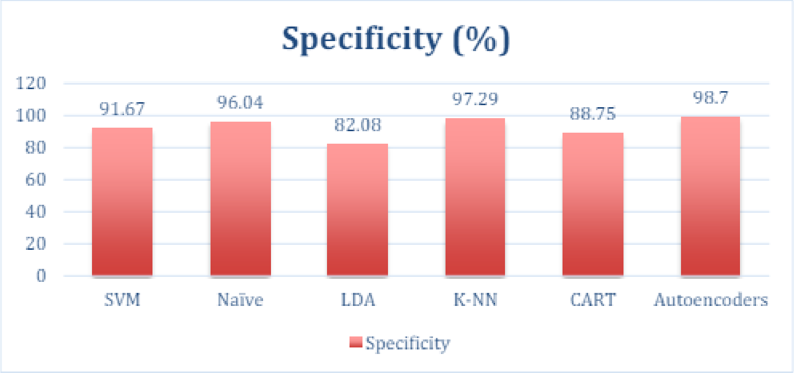
Sensitivity

Sensitivity is the value of correctly classified positive instances (Coenen, 2012). It says how well the algorithm correctly classified the patient has liver disease.



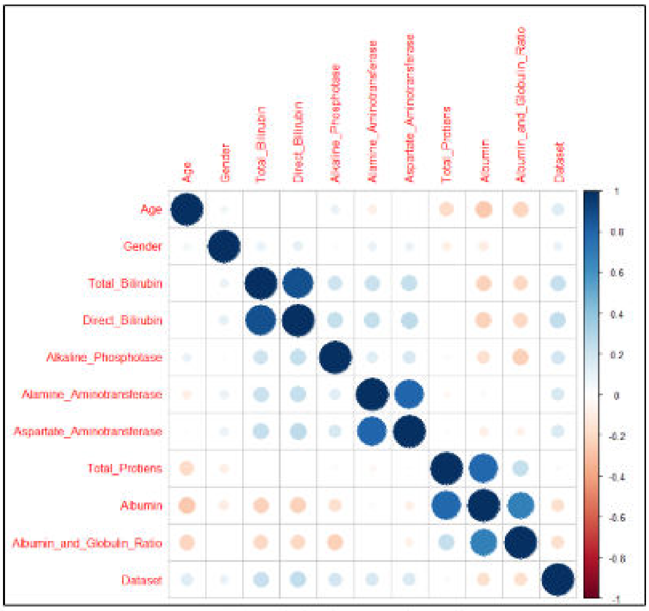
Specificity

Specificity is the value of correctly classified negative instances (Coenen, 2012). It says how well the algorithm has correctly classified that patient has no liver disease. Specificity = TN/(TN+FP)



Correlation between the Attributes

In the correlation plot, it can be observed that some attributes, namely total proteins, albumin and globulin ratio, and albumin, are not much likely correlated with the class attribute, and the remaining attributes are significantly correlated with the class attribute. This correlation plot is generated using Pearson Correlation in R to know how likely the attributes are correlated.



The proposed liver disease prediction (LDP) method has provided the right path for liver disease detection. From the results of this study, after balancing the dataset, SVM has 78.1%, and Naïve Bayes has 65.1%. This balancing of the dataset using ROSE significantly changes the accuracy compared to the accuracies produced by Auxilia (2018), which is 77% for SVM and 37% for Naïve Bayes.

V.CONCLUSION

Since the liver disease is not easy to diagnose, given the delicate nature of its signs, this research is pertinent in determining the algorithms that have better accuracy in predicting this dreadful disease. The stages in the proposed LDP method provide a better alignment of each phase. Once the dataset is selected, the preprocessing step is conducted by replacing the missing values and balancing the dataset. After that, using R, five different supervised learning methods are applied (i.e., SVM, Naïve Bayes, KNN, LDA, and CART), and the accuracy with confusion matrix metrics are recorded. The result shows that K-NN has a better accuracy of 91.7% for liver disease prediction. Autoencoders are applied in this research as a test case for understanding the classification ability of unsupervised algorithms over other traditional approaches. In this study, the autoencoder with 3-layers achieved an accuracy of 92.1%, slightly higher than K-NN due to its ability to ascertain overlapping features better than conventional K-NNs. Most of the algorithms are more than the acceptable level of accuracy, which is 75%. The results from this study would be able to assist health care professionals and relevant stakeholders in the early detection of liver disease.

REFERENCES

[1] Rong-Ho Lin, "An Intelligent Model for Liver Disease Diagnosis," Artificial Intelligence in Medicine, 2009"
[2] Ryan Rifkin, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Chen-Hsiang Yeang, Micheal Angelo, Christine Ladd, Micheal Reich, Eva Latulippe, Jill P Merisov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S Lander, Todd R Golub, "An Analytical Method For Multi-Class Molecular Cancer Classification ", 2003
[3] Akin Ozcivit and Arif Gulten "Classifier Ensemble Construction With Rotation Forest To Improve Medical Diagnosis Performance Of Machine Learning Algorithms",2011
[4] Kun-Hong Liu and De-Shuang Huang. "Cancer classification using Rotation forest", Computers in Biology and Medicine, 2008
[5] BendiVenkataRamana, Prof. M.Surendra Prasad Babu and Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis". International Journal of Engineering Reasearch and Development, 2012
[6] V.N. Vapnik, "Statistical Learning Theory", Wiley Publications, 1998
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Delving Deep into Rectifiers", Microsoft Research, 2009
[8] Beilharz TH, Preiss T: Translational profiling: the genome-wide measure of the nascent proteome. Brief Funct Genomic Proteomic, 2009.
[9] Gros F: From the messenger RNA saga to the transcriptome era. C R Biol. 2003, 326: 893-900.

[10] Shackel NA, Gorrell MD, McCaughan GW: Gene array analysis and the liver. Hepatology. 2002, 36: 1313-1325. 10.1053/jhep.2002.36950.
[11] Yano N, Habib NA, Fadden KJ, Yamashita H, Mitry R, Jauregui H, Kane A, Endoh M, Rifai A: Profiling the adult human liver transcriptome: analysis by cDNA array hybridization. J Hepatol. 2001, 35: 178-186. 10.1016/S0168-8278(01)00104-0.
[12] Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt_Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S: Intra- and interspecific variation in primate gene expression patterns. Science. 2002, 296: 340-343. 10.1126/science.1068996.
[13] Nicholas A Shackel, Devanshi Seth, Paul S Haber, Mark D Gorrell and Geoffrey W McCaughan, “The Hepatic Transcriptome in human Liver Disease”. 10.1186/1476-5926-5-6, BioMedCentral, 2006
[14] World Health Rankings, www.worldlifeexpectancy.com
[15] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29 CS231n : Convolutional Neural Networks for Visual Recognition

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details