Computative Analysis of Various Techniques for Classification of Liver Disease

J. Sujith , P. Karthik Kumar , S. Joshi Manohar Reddy , Aniruddha Kanhe

Department of Electronics and Communication Engineering, NIT Puducherry, Puducherry.

E-mail: sujithjulakanti2002@gmail.com , polamreddykarthikkumar@gmail.com , joshimanoharsomireddy321@gmail.com. , aniruddhakanhe@nitpy.ac.in

Abstract. This paper presents a software-engineered approach using a classification algorithm for the classification of liver disease. The ILPD dataset is used for the proposed work. Different attributes of liver patient records such as direct bilirubin, age, sex, total bilirubin, alphos, albumin, sgpt, globulin ratio, sgot are used to classify liver disease. The proposed Convolution Neural Network classification technique shows an accuracy of 67% and a precision of 71%. Various classification algorithms such as CNN, RNN, ANN, and logistic regression are executed on the liver patient dataset and their accuracy is determined.

1. Introduction

Healthcare today is such a vital aspect for every human being that there is a necessity for easily accessible medical services for all. An essential part of the human body is the liver. Among the many jobs of the liver are breaking down red blood cells and storing vitamins, minerals, glycogen, etc. It is situated in the upper right corner of our belly. Due to the lack of equipment in many hospitals in India, it can be an efficient way to the classification of liver disease, and also it reduces staff requirements. Nowadays, deep-learning techniques are widely being used in many healthcare systems like managing medical data, helping in medical diagnosis, and detecting diseases at an early stage.

2. Literature Review

Most of the earlier models, such as random forest(RF), logistic regression, decision tree, SVM, and linear regression, are based on machine learning. This paper's main objective is to analyze deep learning methods such as CNN, RNN, and ANN.

P. Kuppan et al. [1] proposed the Decision Table, Naive Bayes, and J48 to analyze data related to liver disease. However, attributes such as patient history, smoking, diabetes, obesity, alcohol consumption, and smoking were used. The 35-to-65-year-old group is most affected, with 26% of these people having a disability caused by alcohol, 22% by smoking, and 4% and 5% by obesity and diabetes, respectively.

A. Gulia et al. [2] proposed Bayesian networks, SVM, random forests, and J48, to classify liver patient data. The algorithm is tested on the UCI repository provided by the Center of Machine Learning and Intelligent Systems. The Random Forest algorithm proposed in [2] shows a classification accuracy of 71.8%.

Y. Kumar et al. [3] proposed a classification algorithm based on rules for predicting liver disease. Absent RBC, the acceptance efficacy of all popular algorithms was analyzed. The algorithm uses 20 rules for classifying liver disease. Decision tree-based outperformed the rule-based classification.

J. Singh et al. [4] proposed algorithms with random forests, naive Bayes classifiers, SVM, J48, and logistic regression to classify liver patient data. Indian liver patient dataset was used. The three-step analysis and logistic regression algorithm proposed in [4] shows an accuracy of 71.87%.

Vijayarani et al. [5] used a classification algorithm to predict liver disease. Uses algorithms proposed in [5] such as Naive Bayes and SVM. The UCI repository is where the data set was obtained & employs fields such as Sgot, Gender, ALP, ALB, and DB. Based on their current work, Support Vector Machines (SVM) are the most accurate, and Naive Bayes is faster when it comes to execution.

A. Sateesh. [8] worked on building machine learning models using ILPD. Random Forest (RF) algorithms are used to predict disease using various preprocessing techniques. Univariate and bivariate analyses are used to analyze for outliers, skewness, and imbalance. Then a suitable algorithm is used to eliminate outliers and various oversampling and undersampling techniques are used to balance the data.

The Literature suggests that with logistic regression, maximum accuracy can be achieved i.e, 74.36%.

3. Proposed Methodology

We are classifying liver disease in the proposed work using feature selection techniques and classification algorithms for evaluating liver disease. The following list of steps includes details:

3.1. Pre-processing

The ILPD dataset was prepared in ARFF(Attribute Relation File Format) format. The data were changed to the format required for the execution of several classifiers. To prepare the dataset in a standard format, it was also necessary to remove a redundant field, duplicate records, and any missing records.

3.2. Performance Evaluation

In order to compare the results of various classification methods like CNN, RNN, ANN, and logistic regression, the performance of each classifier is assessed using several performance metrics including recall, precision, accuracy, and F1-score.

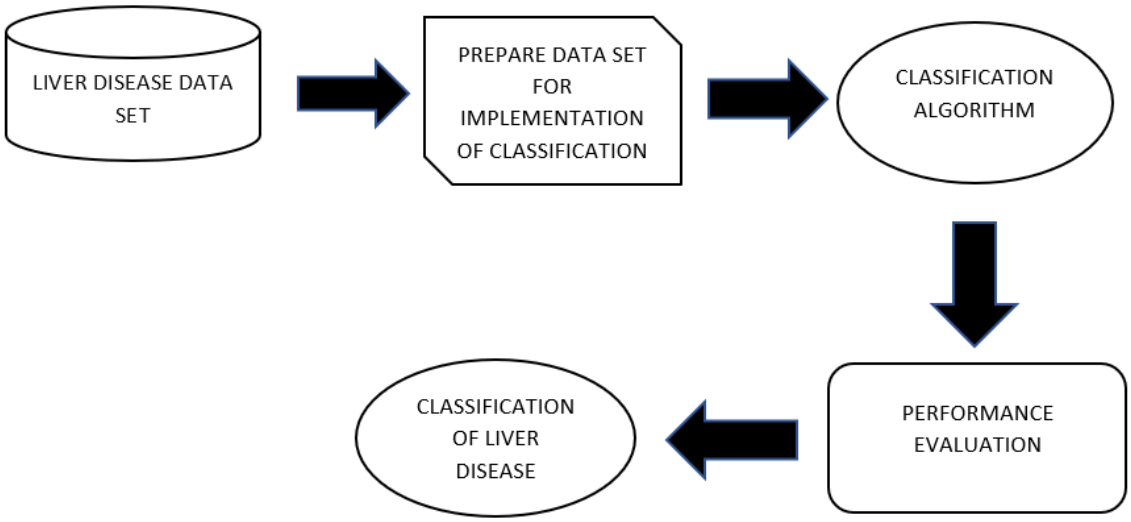

Figure 1. Work Flow Diagram

3.2.1. Accuracy: It is the ratio of predictions that were correct to those that were made in total.

$$Accuracy = \frac{TrueNegative + TruePositive}{TrueNegative + FalseNegative + TruePositive + FalsePositive}$$

3.2.2. Precision: It is the ratio of correctly classified samples that are positive to those that are correctly classified (both negative and positive).

$$Precision = \frac{TruePositive}{FalsePositive + TruePositive}$$

3.2.3. Recall: The Recall is the ratio of correctly classified positive samples to all samples that were classified as positive.

$$Recall = \frac{TruePositive}{FalseNegative + TruePositive}$$

3.2.4. F1-score: It is the harmonic mean of Recall and Precision.

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision}$$

4. Classification Techniques

4.1. Convolutional Neural Network

CNN has frequently been used in object recognition, facial recognition, etc. CNN typically has three layers: the convolutional layer, the pooling layer, and the fully connected layer. There are 16 convolutional layers in this network and pooling is of order 2.

The convolutional Layer performs mathematical operations on the input matrix and kernel matrix. A kernel is a type of filter that is used to take features from the input matrix. The kernel is a matrix that is moved over the input data, executes a dot product operation with the

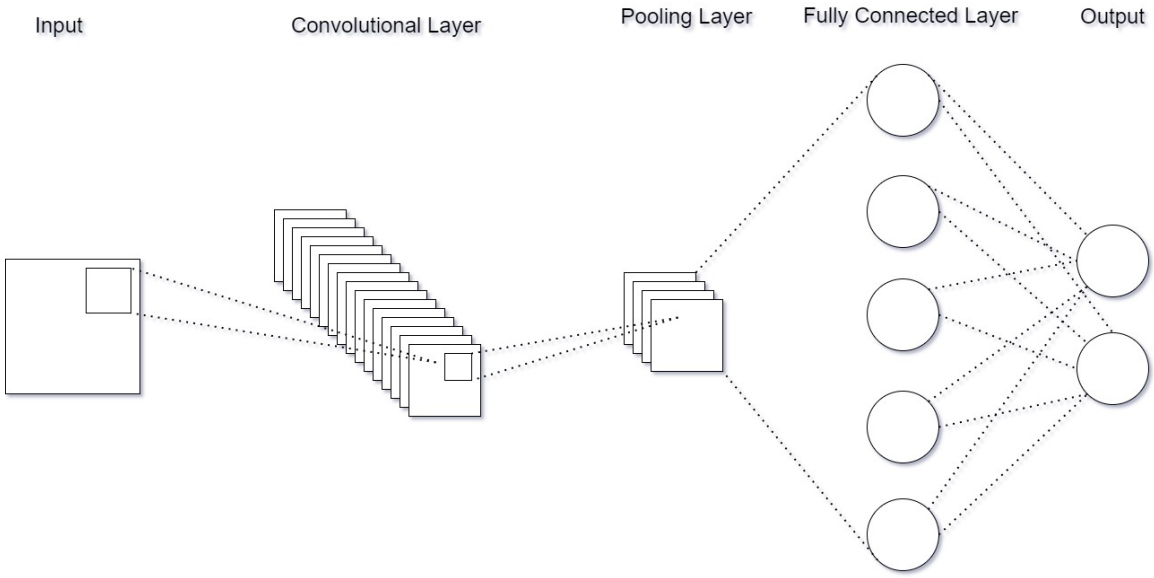

Figure 2. Convolutional Neural Network

subregion of input data, and generates a matrix of dot products as the output. On the resulting matrix, the pooling operations are carried out to cut down on the number of parameters. Finally, a fully connected layer flattens the matrix into the vector and feeds it to the final layer. In fig.2, the CNN's structure is shown.

4.2. Recurrent Neural Network

An RNN is a type of neural network that computes the current output while taking into account both the input and the output from the previous step. The hidden state of RNN, which makes use of some of the information about a sequence, is a crucial feature of RNN. There are 16 hidden layers in this network. In fig.3, the RNN's structure is shown.

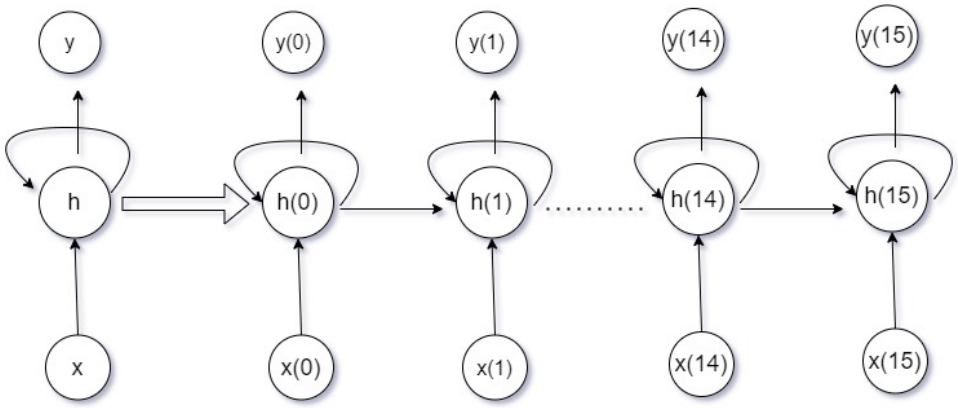

Figure 3. Recurrent Neural Network

Here x(t) is the input, h(t) is the hidden layer, and y(t) is the output.

4.3. Artificial Neural Network
ANN networks are usually based on biological neural networks that build the structure of the human brain. An artificial neural network typically comprises three layers. These three are the input layer, the hidden layer, and the output layer. In fig.4, the ANN's structure is shown.

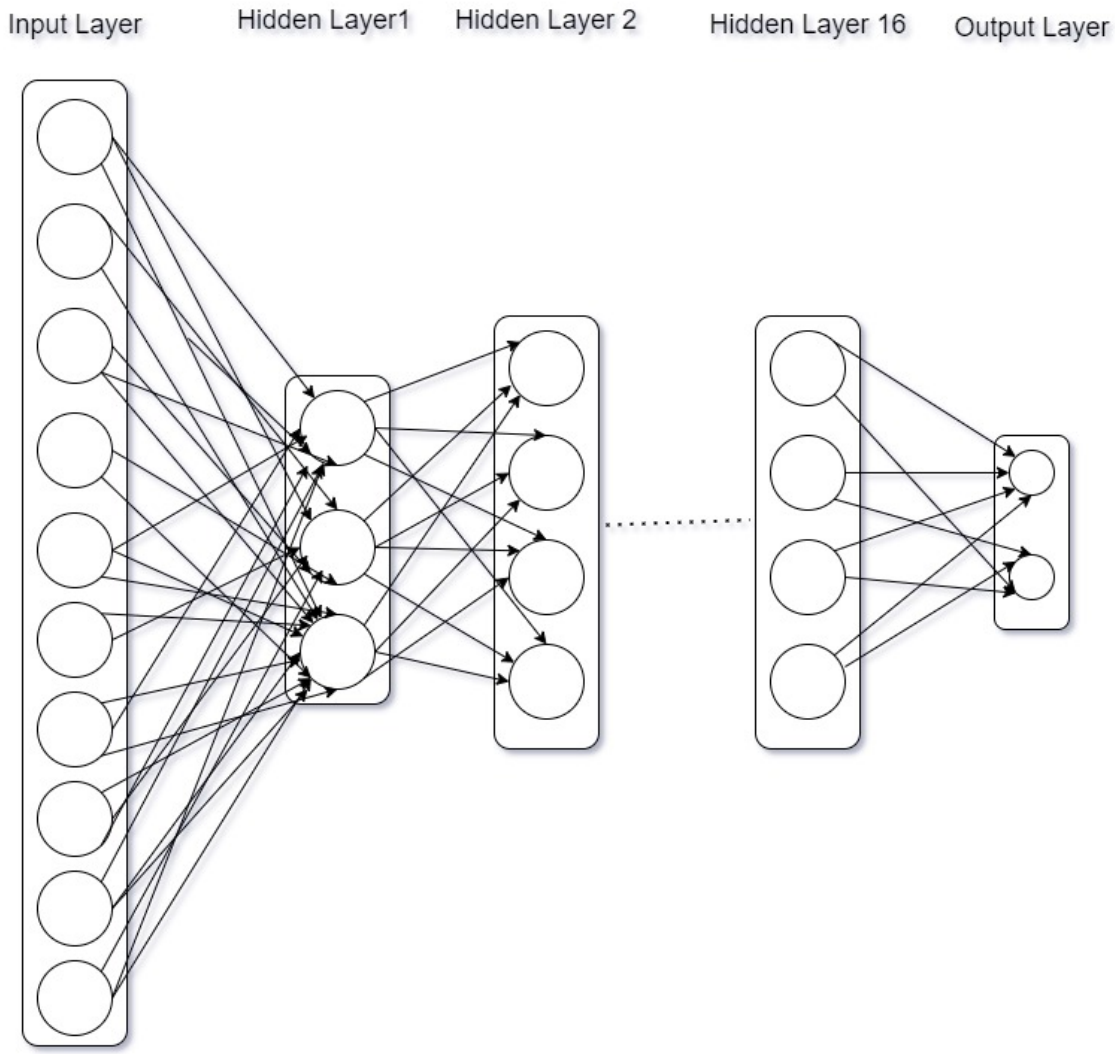

Figure 4. Artificial Neural Network

The input layer accepts the different inputs. The input data are fed to the hidden layers. There are 16 hidden layers in this network. The hidden layer calculates the resultant output based on the weights and sends it to the next hidden layer for further processing. The final hidden layer is linked to the output layer, which displays the output.

4.4. Logistic Regression[9]
It is a classification technique that predicts the output in the form of binary as Yes/No or 1/0 or true/false based on the data from the dataset containing independent variables.

5. Dataset

The proposed algorithms are evaluated using the ILPD data set, which consists of 217 liver patient records and 167 non-liver patient records obtained from Andhra Pradesh's North East region in India. The ILPD dataset contains attributes like sgpt, age, sex, total bilirubin, globulin ratio, alphos, albumin, direct bilirubin, sgot. To classify groups as having liver disease or not, the "is patient" column is used as a class label. In this data set, there are 104 records of female patients and 280 records of male patients. There are 20 patient records with patients under the age of 18, 308 patient records whose age is between 18 to 60 years, and 63 patient records whose age is more than 60 years.

6. Results

The proposed CNN, RNN, ANN, and Logistic Regression are implemented on the Jupyter Notebook version 5.0 tool and tested on the ILPD data set. The performance matrices are obtained by considering the 60% training data and 40% testing data and epochs are 10.

Tool used: Jupyter Notebook

Table 1. Convolutional Neural Network

OPTIMIZERS			
	Adam	Adamax	Adagrad
Accuracy	65%	67%	62%
Precision	70%	71%	64%
Recall	66%	70%	69%
F1-score	68%	70%	69%

Table 1 shows that for Convolutional Neural Network, the accuracy of 67% and Precision of 71% when tested with the Adamax optimizer, as the accuracy, and precision is 65% and 70% respectively when tested with the adam optimizer. For the Adagrad optimizer, the accuracy and precision are 62% and 64% respectively. So, the Adamax optimizer shows better results as compared to Adam and Adagrad optimizers.

Table 2. Recurrent Neural Network

OPTIMIZERS			
	Adam	Adamax	Adagrad
Accuracy	66%	66%	65%
Precision	70%	72%	66%
Recall	67%	65%	72%
F1-score	69%	68%	71%

Table 2 shows that for the RNN, the accuracy of 66% and Precision of 72% when tested with the Adamax optimizer. The accuracy and precision are 66% and 70% respectively when tested with the adam optimizer. For the Adagrad optimizer, the accuracy and precision are 65% and 66% respectively. So, the Adamax optimizer shows better results as compared to Adam and Adagrad optimizers.

Table 3. Artificial Neural Network

OPTIMIZERS			
	Adam	Adamax	Adagrad
Accuracy	57%	61%	51%
Precision	58%	66%	57%
Recall	57%	63%	63%
F1-score	52%	65%	60 %

Table 3 shows that for Artificial Neural Network, the accuracy of 61% and Precision of 66% when tested with the Adamax optimizer. The accuracy and precision are 57% and 58% respectively when tested with the adam optimizer. For the Adagrad optimizer, the accuracy and precision are 51% and 57% respectively. So, the Adamax optimizer shows better results as compared to Adam and Adagrad optimizers.

Table 4. Logistic Regression

Accuracy	65%
Precision	68%
Recall	65%
F1-score	67%

Table 4 shows that Logistic Regression's accuracy is about 65% and precision is 68%.

7. Conclusion

Data Processing is done on the data set to filter the unnecessary values. CNN, RNN and ANN, and Logistic Regression are considered for execution in the assessment of liver disease prediction. In this, we have implemented 16 layers of CNN, RNN, and ANN on the ILPD dataset. From the result, Convolutional Neural Networks with Adamax optimizer give 1% accuracy higher than Recurrent Neural Networks. However Recurrent Neural Network with Adamax optimizer gives 1% more precision than Convolutional Neural Network.

References

[1] Kuppan, P., Manoharan, N. (2017). A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes. International Journal of Computing Algorithm, **6(1)**,2278-239.

[2] Gulia, A., Vohra, R., Rani, P. (2014). Liver patient classification using intelligent techniques. International Journal of Computer Science and Information Technologies, **5(4)**, 5110-5115.

[3] Kumar, Y., Sahoo, G. (2013). Prediction of different types of liver diseases using rule based classification model. Technology and Health Care, **21(5),** 417-432.

[4] Singh, J., Bagga, S., Kaur, R. (2020). Software based prediction of liver disease with feature selection and classification techniques. Procedia Computer Science,**167**,1970-1980. https://doi.org/10.1016/j.procs.2020.03.226

[5] S. Vijayarani, and S. Dhayanand. (2015) "Liver disease prediction using SVM and Naïve Bayes algorithms." International Journal of Science, Engineering and Technology Research (IJSETR),**4**, pp. 816-820.

[6] Jin, H., Kim, S., Kim, J. (2014). Decision factors on effective liver patient data prediction. International Journal of Bio-Science and Bio-Technology, **6(4)**,167-178. http://dx.doi.org/10.14257/ijbsbt.2014.6.4.16

[7] Vandana, U., Ahmed, N. S. S. (2022). A Survey on Machine Learning Techniques for the Diagnosis of Liver Disease. International Journal, **7(6)**, 30-36.https://doi.org/10.46335/IJIES.2022.7.6.8

[8] Ambesange, S., Vijayalaxmi, A., Uppin, R., Patil, S., Patil, V. (2020, November). Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques. In 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM) (pp. 98-102). IEEE.

[9] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, **24(1)**, 12-18.https://doi.org/10.11613/BM.2014.003

[10] Ramana, B. V., Babu, M. S. P., Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems, **3(2)**, 101-114.

[11] Che, H., Brown, L. G., Foran, D. J., Nosher, J. L., Hacihaliloglu, I. (2021). Liver disease classification from ultrasound using multi-scale CNN. International Journal of Computer Assisted Radiology and Surgery, **16(9)**, 1537-1548.

[12] Zhang, Z. Y., Wang, Z. M., Huang, Y. (2020). Polycystic liver disease: Classification, diagnosis, treatment process, and clinical management. World journal of hepatology, **12(3)**, 72.

[13] Singh, A. S., Irfan, M., Chowdhury, A. (2018, December). Prediction of liver disease using classification algorithms. In 2018 4th international conference on computing communication and automation (ICCCA) (pp. 1-3). IEEE.

[14] Valenzuela-Vallejo, L., Mantzoros, C. S. (2022). Time to transition from a negative nomenclature describing what NAFLD is not, to a novel, pathophysiology-based, umbrella classification of fatty liver disease (FLD). Metabolism, 134, 155246.

[15] Gaber, A., Youness, H. A., Hamdy, A., Abdelaal, H. M., Hassan, A. M. (2022). Automatic classification of fatty liver disease based on supervised learning and genetic algorithm. Applied Sciences, **12(1)**, 521.