Indonesian Journal of Electrical Engineering and Computer Science
Vol. 36, No. 2, November 2024, pp. 1043~1051
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v36.i2.pp1043-1051

Comparative analysis on liver benchmark datasets and prediction using supervised learning techniques

Tilakachuri Balakrishna[1], Jagadeeswara Rao Annam[2], Dasari Haritha[3]
[1]Department of Computer Science and Engineering, University College of Engineering Kakinada (A),
Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India
[2]Department of Computer Science and Engineering (AI&ML), CVR College of Engineering (A), Telangana, India
[3]Department of Computer Science and Engineering, University College of Engineering (A),
Jawaharlal Nehru Technological University Kakinada (JNTUK), Kakinada, India

Article Info

Article history:

Received Jan 28, 2024
Revised Jun 27, 2024
Accepted Jul 14, 2024

Keywords:

Classification
Liver diagnosis
Machine learning
Medical data mining
Soft-computing

ABSTRACT

Disease diagnosis is most challenging task today. Different datasets are available in web source that contains important features to diagnose the diseases. This paper explores different classification algorithms on medical liver bench mark datasets like BUPA and Indian Liver patient dataset (ILPD). The ILPD is best fit for the model and also gives high classifier accuracy. In proposed model the following classifiers like Naïve Bayes (NB), support vector machine (SVM), K-nearest neighbor (KNN), decision tree (DT), and random forest (RF) classification, multi-layer perceptron (MLP), artificial neural network (ANN), deep belief network (DBN) and probabilistic neural network (PNN) are used. The results shown that ILPD is best dataset for all classifiers and RF classification in particular is best classifier.

Corresponding Author:

Tilakachuri Balakrishna
Department of Computer Science and Engineering, University College of Engineering Kakinada (A)
Jawaharlal Nehru Technological University Kakinada (JNTUK)
Kakinada, Andhra Pradesh, India
Email: balakrishnagec@gmail.com

1. INTRODUCTION

Medical datamining is a prominent research area to extract the meaningful data from the given features. A key challenge is to diagnosis the disease [1]. In earlier days doctors observes complete information about the tests and come to know the conclusion that the person is diagnosed with some disease. Now a days, especially in medical field is digitized and new innovation strategies comes in to the picture to diagnosis the disease without human intervention. One such innovation is classification. In this technique the model is trained with some known samples along with the class label or decision. When a new sample is supplied to the model, based on the training samples the model will generate the decision result. Liver disease accounts for two million deaths per year and is responsible 1 out of every 25 deaths worldwide. Classification of Liver disease is the most challenging task this paper explores the following traditional classification algorithms [2] like Naïve Bayes (NB), K-nearest neighbour (KNN), decision tree (DT), random forest (RF) classification model, support vector machine (SVM) on the other hand soft computing-based classification algorithms [3] like multi-layer perceptron (MLP), artificial neural network (ANN), deep belief network (DBN) and probabilistic neural network (PNN), The following Figure 1 depicts classification of supervised learning techniques.

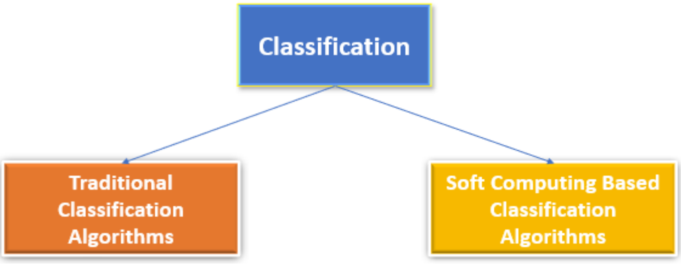
Classification
Traditional Classification Algorithms
Soft Computing Based Classification Algorithms

Figure 1. Classification of supervised learning techniques

2. RELATED WORK

In Related work best classifiers [4] like SVM, KNN, DT, SVM, PNN, and MLP are identified especially in health-related datasets. These classifiers were further classified into both traditional and soft-computing based and then techniques will be applied to liver benchmark datasets.

2.1. Traditional classification algorithms

2.1.1. NB

It is a supervised learning algorithm worked for classification algorithm based on Bayes' theorem [5] based on prior knowledge which depends on conditional probability. Bayes' theorem formulae is given as:

$$P(A|B) = (P(B|A) * P(A))/(P(B))$$

were,
Posterior probability is given as P(A|B) that is hypothesis of A on event B.
The Likelihood probability i.e., evidence given that hypothesis is true.
Advantage:

– It is a simple classifier works on small dataset.

2.1.2. SVM

The SVM is a type of supervised machine learning technique that may be used for both classification and regression tasks. Its goal is to identify a hyperplane in a space with N dimensions that effectively separates the data points into discrete classes. The size of the hyperplane is determined by the number of features. In Figure 2 depicts SVM classifier.

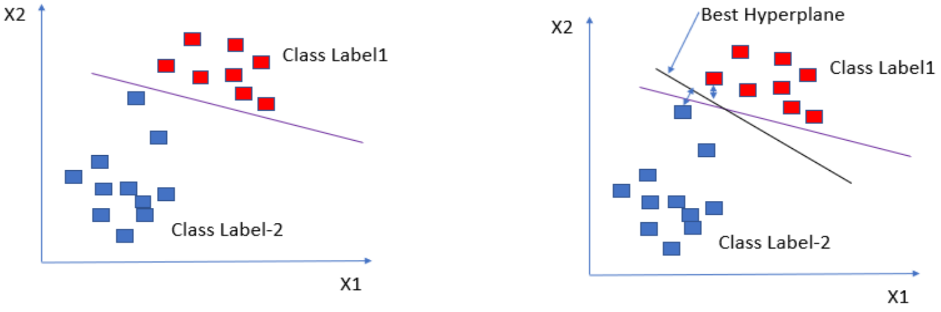

Figure 2. SVM classifier

Advantages of SVM:

i) SVMs are highly effective in handling high-dimensional instances.
ii) The approach demonstrates memory efficiency by utilizing a limited collection of training points, referred to as support vectors, in the decision function.
iii) It is possible to specify different kernel functions for the decision functions, and it is also feasible to specify customized kernels.

Battineni *et al.* [6] suggested the SVM a classification approach to predict the dementia to achieve optimized results. The proposed approach exhibits an accuracy of 68.75% and a precision of 64.18. Dementia a disorder caused by brain disease or injury is one of the major health issues challenged by health experts. It was identified the especially the senior citizen people whose age is greater than 60. It results lack of memory skills and perform daily activities. For the last two decades lots of research made by computer scientists and health experts for this disease. Still, there is an urge for detection of dementia. ECG data collected by the Kardia Mobile app can accurately forecast the risk of cardiac illness. The study [7] results revealed that SVM classifier surpassed traditional statistical methods and achieved accuracy to predict the risk of heart illness.

2.1.3. KNN classifier

This classifier serves as supervised learning technique. The KNN method utilizes the similarity between a new sample and the training examples to predict the results of the new sample as being closest or most similar to the instances in the training set. Figure 3 depicts KNN classifier.

Cardiac illness is the primary cause of deaths in India. The recorded mortality rate in AP accounted for 32% of all deaths, which is comparable to the rates observed in Canada (35%) and the USA. Therefore, it is necessary to have a decision support system that provides guidance on implementing key preventative measures in order to prevent these elevated mortality rates. The author proposed the combined KNN and genetic algorithm (GA) [8] is the outstanding classifier compared with others to diagnose heart disease. The proposed results show higher classification accuracy for many datasets particularly heart disease for A.P. A comparative analysis was conducted to categorize and predict cases of heart disease using several classifiers [9]. The goal was to achieve accurate classification while using as few attributes as possible. The dataset comprises a total of 76 attributes, which includes the class attribute. The results indicate that KNN outperforms JRip, J48, and decision table, making it the most effective classifier.

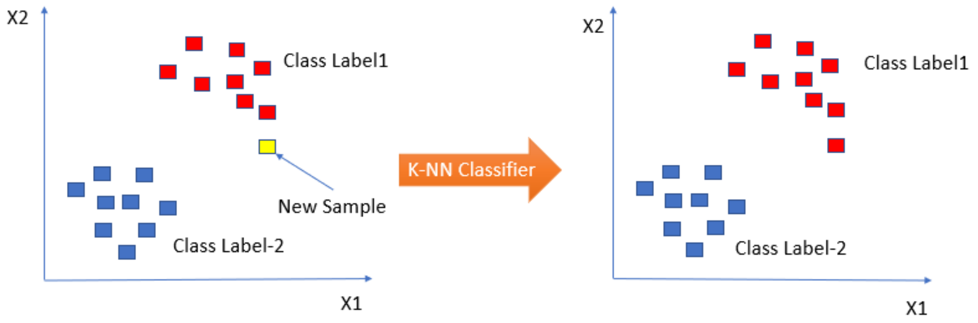

Figure 3. KNN classifier

2.1.4. DT

It is a binary tree [10] is a kind of supervised learning algorithm that is skilled in solving regression as well as classification challenges. However, it is mostly used for dealing with classification. A tree-structured classifier is used, including internal nodes signifying dataset features, branches indicating decision rules, and each leaf node indicating a result. In Figure 4 show the model of DT.

A. Attribute selection measures (ASM) in DT

When implementing a DT, a key challenge arises with the selection of the most appropriate attribute for both the root node and the sub-nodes. In order to tackle these issues, the utilization of a technique called ASM might be employed. Using this metric, we can readily determine the optimal attribute for the nodes in the tree. There are two widely used approaches for ASM (1) and (2), namely:
information gain,

$$IG(L,X) = Entropy(X) - \sum_{split} \frac{x1}{L} Entropy(x1) \tag{1}$$

gini index,

$$Entropy = -\sum_{i} P(i).log2P(i) \tag{2}$$

High entropy: more uncertainty
Low entropy: more predictability

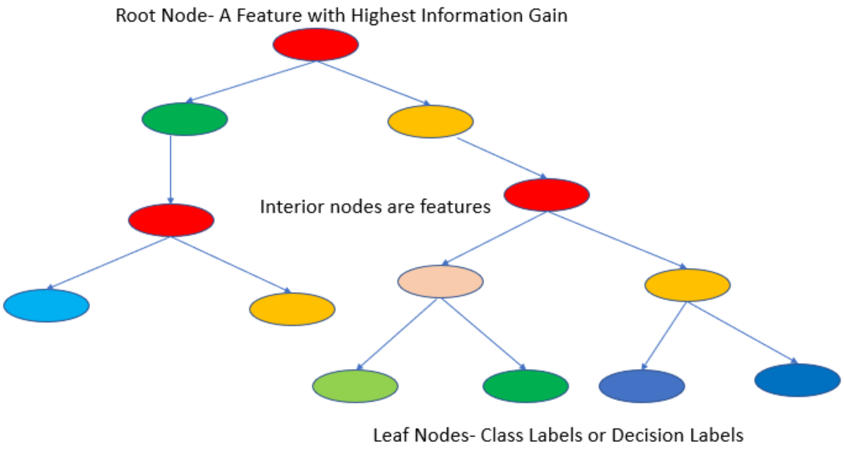
Root Node- A Feature with Highest Information Gain
Interior nodes are features
Leaf Nodes- Class Labels or Decision Labels

Figure 4. DT classifier

2.1.5. RF classifier

It is a machine learning technique employed for solving both regression and classification problems. It employs ensemble learning, a method that combines multiple classifiers to solve challenging problems shows in Figure 5. The RF method determines the result by relying on the predictions made by the DTs. It makes predictions by calculating the average or mean of the output generated by several trees. Enhancing the quantity of trees enhances the accuracy of the result.

To predict chronic kidney disease (CKD), data mining approaches are applied. They suggested that the RF classification approach [11] is the most appropriate classifier when compared to other techniques like J48, NB tree, and REP tree, based on the accuracy of categorizing examples. The KNN classifier and logistic regression [12] classifier is outstanding classifier compared to the SVM and RF classifier. The former achieved 100% accuracy in the CKD dataset and 85% accuracy in the heart disease dataset. The suggested model additionally suggested employing principal component analysis (PCA), a technique used to decrease the number of features in a dataset.

This proposed study [13] employs machine learning methods, including KNN, DT, and RF, to predict and analyze the Pima Indian diabetes dataset. The findings indicated that the RF model had superior performance, with an accuracy rate of 0.84. Tareq [14] recommends using the RF classifier technique on the sleep health and lifestyle (SHL) dataset, which consists of three classes: insomnia, sleep apnea, and none, and this classifier achieves 88% accuracy.

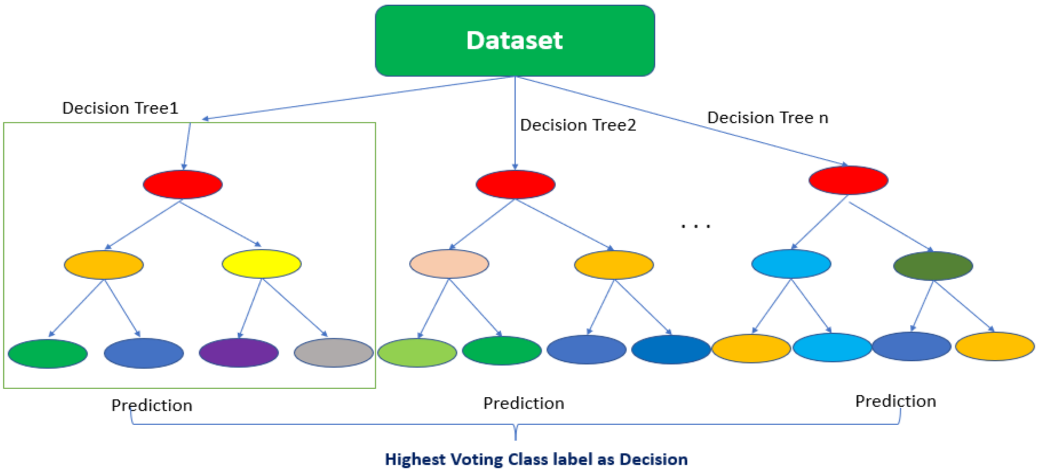

Figure 5. RF classifier

2.2. Soft-computing classification algorithms

2.2.1. MLP

It is a fully connected feed-forward neural network contains at least three layers one is input layer; second hidden layer and last layer is output layer. This model accepts more than one hidden layer also. Each unit within a layer is connected to units in adjacent levels, allowing for homogeneous computing from the input layer to the output layer shown in Figure 6. Safar *et al.* [15] MLP network to identify trends in various healthcare datasets like Diabetes, Heart Disease, Breast Cancer, Liver, and Parkinson have achieved accuracies of 97.6%, 94.73%, 98.74%, 73%, and 100% respectively using various ML algorithms such as RF, DT, extra trees, NB, GBoost, KNN, LR, MLP, EXGBoost, and SVM.

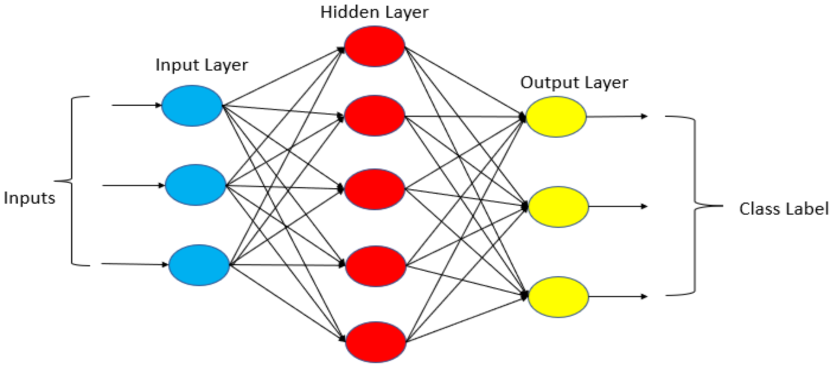

Figure 6. MLP classifier

2.2.2. ANN

In late 90’s and early 20th century majority of research work carried out in the area of ANN [16]. Now, neural networks have been used in different fields of research areas which include defense, aerospace, expert systems, recommendation systems and many. Ozyilmaz and Yildirim [17] a hybrid network conic section function neural network (CSFNN) has the best classifier for diagnosing hepatitis. Early over MLP is the good choice for the dataset but later radial basis function (RBF) produces promising results the drawback of these methods is random weight initializing in training. In Figure 7 shows the ANN classifier.

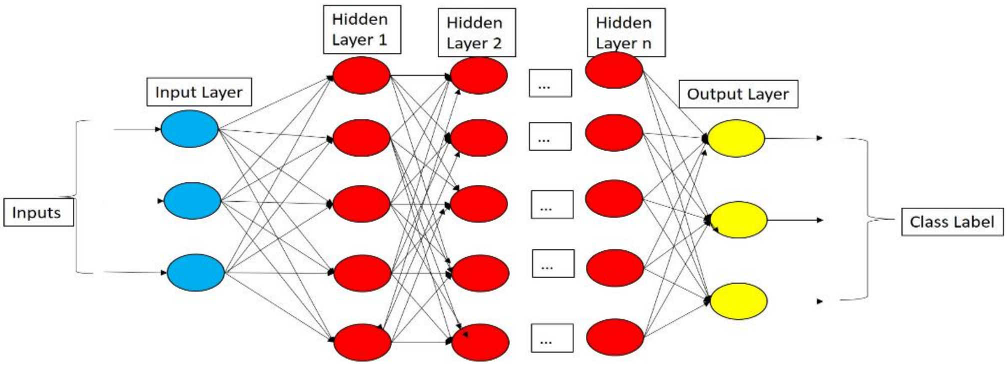

Figure 7. ANN classifier

2.2.3. DBN

DBNs are utilized to tackle the limitations of traditional neural networks in deep hierarchical networks. For instance, delayed learning, getting trapped in local minima due to inadequate parameter selection, and necessitating a substantial amount of training datasets. The DBN method is a technique used for unsupervised probabilistic deep learning shown in Figure 8.

Elkholy *et al.* [18] proposed a DBN classifier that accurately predicts CKD with an accuracy percentage of 98.5. The proposed model utilizes a DBN with softmax classifier and the categorical cross-entropy as the loss function. In Figure 8 depicts the model of DBN.

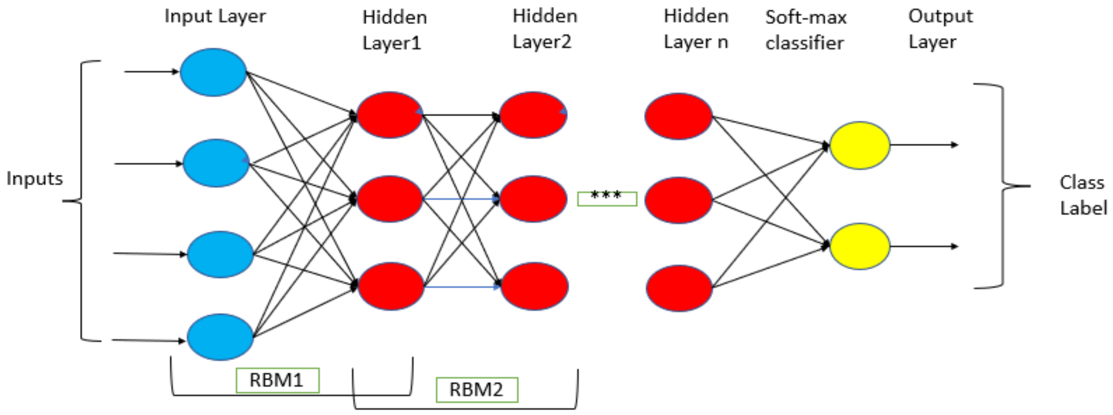

Figure 8. DBN classifier

2.2.4. PNN

It is a type of feedforward neural network that is specifically designed for the purpose of solving pattern recognition and classification tasks [19]. The PNN approach approximates the parent probability distribution function of each class by utilizing a non-parametric function and Parzen window. The probability density function (PDF) of each class is subsequently utilized to estimate the class probability of newly inputted data [20]. Bayes' rule is used to determine the class having the highest posterior probability for the new input data. This strategy reduces the likelihood of misclassification. In Figure 9 depicts the PNN classifier. PNN algorithm [21] gives the better classification accuracy percentage of 96.7%, compared to algorithms like SVM, RBF, and MLP on CKD. Femil and Jaya [22] proposed PNN classifier to accurately categorize the stages of skin lesions in a skin cancer dataset. The entire project was implemented using MATLAB, and the results demonstrate that the proposed method achieves ideal outcomes with a maximum accuracy of 97.83%.

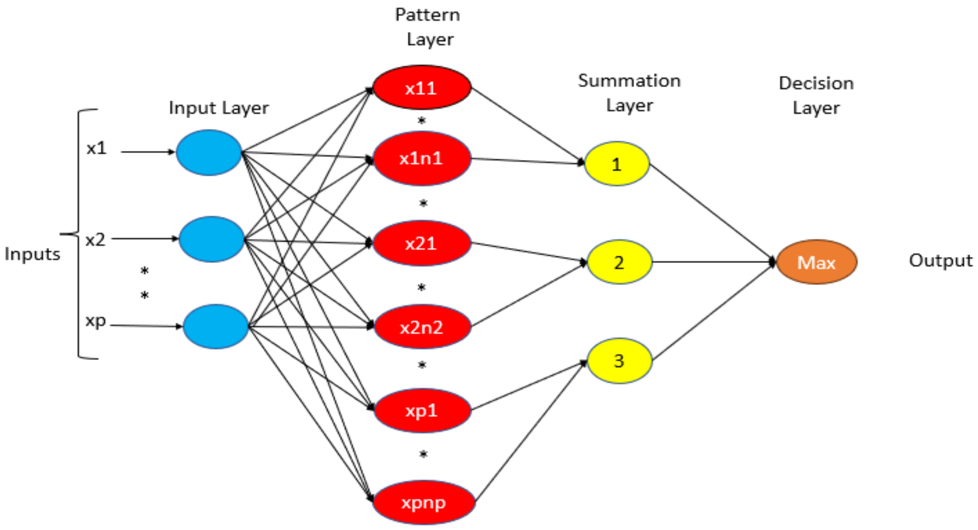

Figure 9. PNN classifier

3. PROPOSED MODEL

The proposed model contains the following stages shown in Figure 10 firstly; it will work on liver bench mark datasets.

1. BUPA dataset [23] donated by BUPA medical research Ltd. The data set contains the 7 features namely mean corpuslar value, alkphos, sgpt, sgot, gammagt, drinks and selector (decision label).
2. ILPD [24] donated by Ramana, Bendi. It has 11 features namely age, gender, tb, db, alkphos, sgpt, sgot, tp, alb, A/G Ratio, and class label.

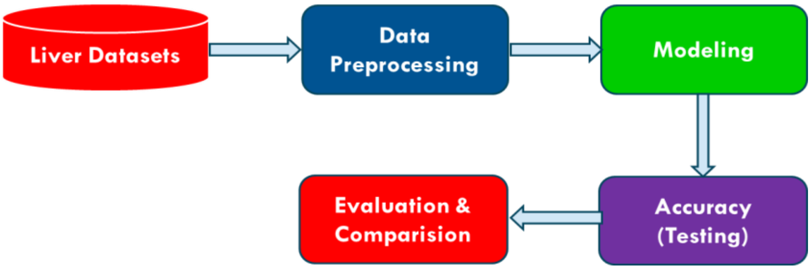

Figure 10. Proposed model architecture

The major finding observed in both the data sets are the features like alkphos, sgpt, sgot are common and these were very important for decision making process. In Table 1 depicts the summary of datasets and sample information. Secondly, in data-preprocessing stage few null values of the records in both the datasets were eliminated and in Liver patient dataset the gender coloumn is dropped. Thirdly, in modelling stage both the traditional and soft-computing based classifiers will work on both the bench mark datasets.

Table 1. Datasets and samples

S. No	Name of the dataset	No. of instances	No. of features	No. of classes
1	BUPA	345	7	2
2	Indian Liver patient dataset (ILPD)	583	11	2

Algorithm RF classifier:

- Step1: randomly sample K data points from the training set.
- Step2: create the RF with the n_estimators which represents the random trees.
- Step3: fix the parameters for RF algorithm (bootstrap=true, max_depth=3, max_features='log2', n_estimators=310).
- Step4: calculate the accuracy decision of test data based on majority voting and calculate the accuracy.

Fourthly, the evaluation metric accuracy is calculated using the standard formulae. Accuracy is a quantitative measure used to assess the performance of classification models. Informally, accuracy (3) refers to the proportion of correct predictions made by our model.

$$\text{Accuracy}=\frac{No.of\ True\ Predictions}{Total\ No.of\ Predictions} \tag{3}$$

For binary classification, accuracy (4) may also be computed by considering the number of true positives and true negatives in the following manner:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

finally, last stage is the comparison stage where we find the best classifiers and best dataset fit into the model.

4. RESULTS AND DISCUSSION

The experiment was done using Anaconda Navigator [25] Jupyter Notebook a python frame work, The necessary packages was pandas, NumPy, Matplotlib, Seaborn, Sklearn, Keras, and NeuPy. Table 2 shows the accuracy of supervised algorithms and for all the algorithms the train and test set split were 80% and 20%. RF has achieved best accuracy of 78.26% on BUPA dataset and 78.53% on ILPD compared to other classification approaches.

Our study suggests that soft computing-based classifiers such a ANN, DBN and PNN exhibits poor accuracy in BUPA dataset and on the other side. The ILPD is best suitable dataset for both traditional classifiers like NB, SVM, KNN, DT, and RF and soft computing-based classifiers. In Figure 11 show accuracy of classifiers where X-axis represents the list of both traditional and soft-computing based classifiers and Y-axis represents accuracy percentage on both BUPA and ILPD datasets.

Table 2. Accuracy of supervised learning algorithms on BUPA and ILPD dataset

S. No	Name of the classifier	BUPA dataset	ILPD dataset
1	NB	52%	57%
2	SVM	76.81%	73.06%
3	KNN	69.57%	65.28%
4	DT	65.22%	73.06%
5	RF	78.26%	78.53%
6	MLP	73.91%	72.02%
7	ANN	39.13%	73.06%
8	DBN	39.13%	72.02%
9	PNN	65.22%	75.21%

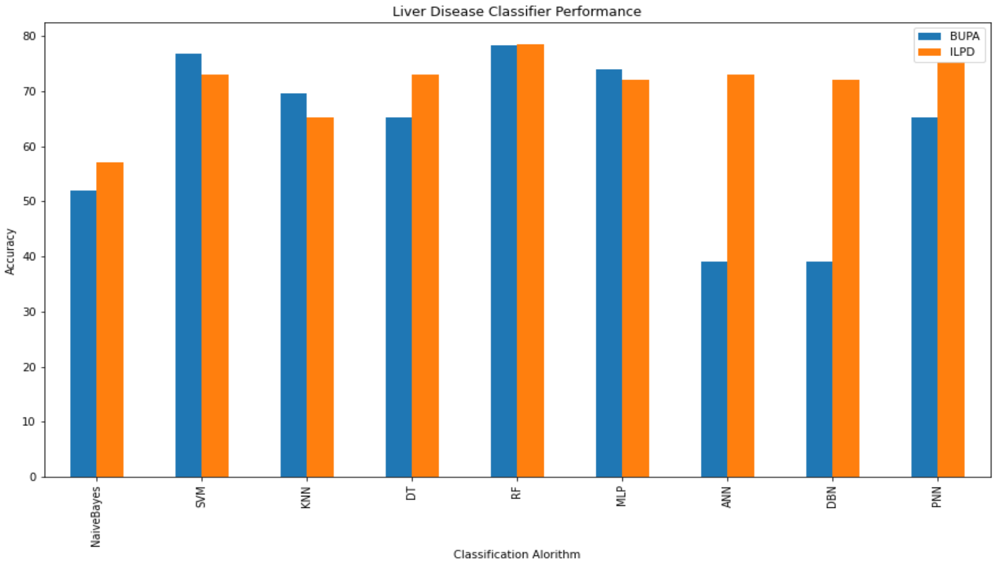

Figure 11. Comparison graph

5. CONCLUSION

Based on the results we found that ILPD correlates with BUPA Dataset. The top classifiers both traditional based and soft-computing like NB, SVM, KNN, DT, RF, MLP, ANN, DBN, and PNN in this study tended to have an inordinately gives best accuracy results to Indian LPD in compared with BUPA dataset. However, further and in-depth studies may be needed to improve its classifier performance.

REFERENCES

[1] D. Haritha, "Comparative study on brain tumor detection techniques," in *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, Oct. 2017, pp. 1387–1392, doi: 10.1109/SCOPES.2016.7955668.

[2] R. Battur and J. Narayana, "Classification of medical X-ray images using supervised and unsupervised learning approaches," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1713–1721, Jun. 2023, doi: 10.11591/ijeecs.v30.i3.pp1713-1721.

[3] O. Dahmane, M. Khelifi, M. Beladgham, and I. Kadri, "Pneumonia detection based on transfer learning and a combination of VGG19 and a CNN built from scratch," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1469–1480, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1469-1480.

[4] J. R. Annam, S. Kalyanapu, S. Ch, J. Somala, and S. B. Raju, "Classification of ECG heartbeat arrhythmia: a review," *Procedia Computer Science*, vol. 171, pp. 679–688, 2020, doi: 10.1016/j.procs.2020.04.074.

[5] J. K. Alwan, D. S. Jaafar, and I. R. Ali, "Diabetes diagnosis system using modified Naive Bayes classifier," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1766–1774, Dec. 2022, doi: 10.11591/ijeecs.v28.i3.pp1766-1774.

[6] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: performance calculation of dementia prediction by support vector machines (SVM)," *Informatics in Medicine Unlocked*, vol. 16, p. 100200, 2019, doi: 10.1016/j.imu.2019.100200.

[7] T. B. Krishna, N. Vimala, P. Vinay, N. Siddhardha, and P. M. Manohar, "Early heart disease prediction using support vector machine," in *Lecture Notes in Networks and Systems*, vol. 719 LNNS, 2023, pp. 471–479.

[8] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013, doi: 10.1016/j.protcy.2013.12.340.

[9] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinformatics*, vol. 21, no. 1, p. 278, Dec. 2020, doi: 10.1186/s12859-020-03626-y.

[10] T. Balakrishna *et al.*, "Diagnosis of thyroid disorders using decision tree classification technique," *CiiT International Journal of Artificial Intelligent Systems and Machine Learning*, vol. 8, no. 4, 2016.

[11] T. Balakrishna, B. Narendra, M. H. Reddy, and D. Jayasri, “Diagnosis of chronic kidney disease using random forest classification technique,” *Helix*, vol. 7, no. 1, pp. 873–877, 2017, doi: 10.29042/2017-873-877.

[12] R. Chandra, M. Kapil, and A. Sharma, “Comparative analysis of machine learning techniques with principal component analysis on kidney and heart disease,” in *Proceedings of the 2nd International Conference on Electronics and Sustainable Communication Systems, ICESC 2021*, Aug. 2021, pp. 1965–1973, doi: 10.1109/ICESC51422.2021.9533011.

[13] N. Hu and J. Gao, “Research on diabetes prediction model based on machine learning algorithms,” in *Proceedings - 2023 International Conference on Computers, Information Processing and Advanced Education, CIPAE 2023*, Aug. 2023, pp. 200–203, doi: 10.1109/CIPAE60493.2023.00044.

[14] W. Z. T. Tareq, “Sleep disorders detection and classification using random forests algorithm,” in *Studies in Systems, Decision and Control*, vol. 513, 2024, pp. 257–266.

[15] A. A. Safar, Dhiadeen M. Salih, and A. M. Murshid, “Pattern recognition using the multi-layer perceptron (MLP) for medical disease: A survey,” *International Journal of Nonlinear Analysis and Applications*, vol. 14, no. 1, 2023, doi: 10.22075/IJNAA.2022.7114.

[16] A. Remaida *et al.*, “Application of artificial neural networks for personality traits prediction based on handwriting,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 3, pp. 1534–1544, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1534-1544.

[17] L. Ozyilmaz and T. Yildirim, “Artificial neural networks for diagnosis of hepatitis disease,” in *Proceedings of the International Joint Conference on Neural Networks*, 2003, vol. 1, pp. 586–589, doi: 10.1109/ijcnn.2003.1223422.

[18] S. M. M. Elkholy, A. Rezk, and A. A. E. F. Saleh, “Early prediction of chronic kidney disease using deep belief network,” *IEEE Access*, vol. 9, pp. 135542–135549, 2021, doi: 10.1109/ACCESS.2021.3114306.

[19] K. R. Kodepogu *et al.*, “A novel deep convolutional neural network for diagnosis of skin disease,” *Traitement du Signal*, vol. 39, no. 5, pp. 1873–1877, Nov. 2022, doi: 10.18280/ts.390548.

[20] J. Gada, A. Savla, S. Chheda, and P. Bhogale, “Brain tumor segmentation,” *International Journal of Computer Applications*, |vol. 138, no. 13, pp. 6–8, 2016, doi: 10.5120/ijca2016908975.

[21] E. H. A. Rady and A. S. Anwar, “Prediction of kidney disease stages using data mining algorithms,” *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019, doi: 10.1016/j.imu.2019.100178.

[22] J. J. Femil and T. Jaya, “An efficient hybrid optimization for skin cancer detection using PNN classifier,” *Computer Systems Science and Engineering*, vol. 45, no. 3, pp. 2919–2934, 2023, doi: 10.32604/csse.2023.032935.

[23] “Liver disorders Medicine,” *UCI Machine Learning Repository*, 1990, doi: 10.24432/C54G67.

[24] “Indian liver patient dataset (ILPD),” *UCI Machine Learning Repository*, 2012.

[25] “Anaconda Navigator | Anaconda,” 2024. https://www.anaconda.com/anaconda-navigator.

BIOGRAPHIES OF AUTHORS

Mr. Tilakachuri Balakrishna Research scholar (19022P0529), University College of Engineering Kakinada (A), JNTUK, Kakinada and he did his M.Tech in Computer Science and Technology with Specilization of Artificial Intelligence and Robotics in Andhra University, Visakhapatnam. He did his B.Tech from JNTUK Kakinda. His area of research is medical datamining and machine learning. Presently he is working as Assistant Professor in Department of Computer Science and Engineering in Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India. He can be contacted at email: balakrishnagec@gmail.com.

Prof. Dr. Jagadeeswara Rao Annam working as professor CVR College of Engineering (A), Hyderabad and did his Ph.D. from University of Hyderabad. He is Experienced Professor with a demonstrated history of working in the research industry. Strong education professional skilled in computer science, C++, datamining, pattern recognition, and machine learning. He guided 15 M.Tech and 35 B.Tech student projects and he published 18 international journals. He is Member of Institute of Electronics Engineers (IEEE). He can be contacted at email: ajagarao@gmail.com.

Prof. Dr. Dasari Haritha working as professor of CSE, University College of Engineering Kakinada (A), Jawaharlal Nehru Technological University, Kakinada. She has 20+ years of experience. She guided 2 Ph.D. and her research interest is on image processing, pattern recognition, software engineering and networking. She published 23 research papers in international journals and 19 conferences and she had 2 patents. She is peer reviewer to the Institution of Engineering and Technology for image processing journal. She can be contacted at email: harithadasari9@yahoo.com.