Journal of Scientific & Industrial Research
Vol. 82, February 2023, pp. 263-268
DOI: 10.56042/jsir.v82i2.70215

Health Care Automation in Compliance to Industry 4.0 Standards: A Case Study of Liver Disease Prediction

Manjula Devarakonda Venkata[1]*, Sumalatha Lingamgunta[2] & K Murali[3]

[1]Department of CSE, Pragati Engineering College, Surampalem 533 437, Andhra Pradesh, India

[2]University College of Engineering, JNTUK, Kakinada 533 001, Andhra Pradesh, India

[3]Dept of ECE, Vijaya Institute of technology for women, Enikepadu, Vijayawada 521 108, Andhra Pradesh, India

Received 04 May 2022; revised 04 October 2022; accepted 07 October 2022

The industrial internet contributes to the standards of Industry 4.0, which involve handling large volumes of data using advanced soft-computing techniques. Machine Learning (ML) is an advanced soft-computing technique that plays a critical role in predicting and detecting serial chronic diseases, thereby automating the diagnosis. The process constitutes and uses several data mining algorithms and methods for efficient medical data analysis. Recent studies on several chronic diseases, liver disorders and diseases associated with the organ have been fatal. In this paper, the liver patient dataset from India is considered and investigated for developing a classification model. Liver disease is a dangerous, life-threatening disease often diagnosed false positive. Mild liver enlargement, improper or ambiguous functionality over a brief period, is prominent even in healthy people, which has become the main reason for ignoring the same at the early stage. It is essential to predict liver disease through the parameters and their values from the liver functionality test sensing the behavior of similar patients who were ignored in the initial stage. In this paper, the machine learning technique is demonstrated to predict liver disease using the liver function test data of the 580 patients as training data. The model has been developed with an accuracy of approximately 75%. The simulation-based experiment is based on the publicly available dataset and can be extended to any native set to predict the patients' health quickly. The Random Forest Algorithm is used to develop the model in Matlab, and the analysis is carried out using parameters like total bilirubin, alkaline phosphotase, alamine aminotransferase, total proteins, and A/G ratio.

Keywords: Advanced soft-computing techniques, Indian dataset, Industrial internet, Machine learning, RF algorithm

Introduction

Physical appointments and manual examinations are usual practices in traditional health care. The health care emerged as a significant sector, mainly referring to the recent pandemic outbreak. The novel COVID-19 demanded rapid medication with no physical intervention from the doctor or practitioner. This needs an efficient automation framework that can be inducted from Industry 4.0. Hence it is possible to mention that the automation process can be imparted to detect or cure several other diseases. One such is liver disease. Liver Diseases have been listed as one of the major diseases. Sometimes, it may end up resulting in disability. According to the available data, by 2011 nearly 508000 deaths have occurred due to this.[1]

Around 17.7 million people succumbed to death in 2015.(2) It is estimated that approximately 23.6 million people would be suffering with liver concerned diseases by the end of 2030.(3) Some have undergone appropriate treatment; however, several scenarios, like the expensive medication and the complex process, are considered a big hindrance.[4] Other parameters, like the tenure of the treatment and delayed decisions are significant. The total expenditure for treating and diagnosing the liver disease is extremely high and usually out of reach to the commoner. It is evident from the recent reports that the expenditure of about $79 billion across the world has been utilized to treat patients suffering from the ultimate stage of the disease, costing around $35 billion.[5] It is also significant to mention that the condition associated with the liver is severe and takes a long time to cure. Hence general citizens cannot afford to long time hospitalization costs. Prediction of the disease helps doctors and practitioners to take all the preliminary measures and suggest accurate and appropriate medication along with efficient treatment to nullify the severity. At this moment, machine

*Author for Correspondence
E-mail: dv.manjula@pragati.ac.in

learning has a role in solving the Liver Disease problem by predicting it early and detecting it, leading to an efficient diagnosis.

The prime objective of this work is to provide a feasible solution to efficient lever disease detection for a further effective diagnosis. The feasibility is in terms of low-cost and less-time for analysis. This certainly requires sophisticated techniques with intelligence. Hence, machine learning methods are considered for their performance in providing solutions to several engineering problems.

Data related to the patients suffering with the disease plays a major role for training the model and finally to provide a good prediction of the disease. The preliminary survey mentioned that the similar data is recorded while it is also inferred that the data is exponentially growing in volumes through the recent decade of period.[6] The ML algorithms are known for their excellence in handling non-linear problems by providing solutions which are practicable.[7] Interestingly medical image and data analysis is a complex non-linear problem with patterns which are hard to model. These ML techniques have proved to produced accurate results of prediction and forecasting in complex non-linear problems related to medical diseases. Among them, the supervised models have the ease of handling in both product development and analysis in the field of medicine.[8] Based on this, the main objective of the study has been the early detection and efficient diagnosis of chronic disease in people from developing and under developed nations. Most of the medical disease diagnosis problems are considered as either classification, modelling, or prediction type. The ML techniques can be easily employed for the classification of medical images and model the features data set obtained from these images of chronic diseases.[9–21] These classification techniques are typically supervised models. The SVM is hybridized with the Popular Particle Swarm Optimization (PSO) to predict several significant features for liver disease detection.[15] Compared to SVM, RF, Bayesian network, and an MLP-neural network, excellent accuracy has been reported.[16] Further the SVM can accurately predict drug-induced hepatotoxicity better than the Bayesian and several other existing models.[17] A CNN model can predict liver cancer in patients suffering with hepatitis accurately with 98%. Taking the above discussion with respect to the abilities of ML in to consideration prominent, ML techniques are employed to predict in order to treat chronic patients. While the prime objective is to develop the prediction model using ML techniques, the other interest is to analyse the performance of several classifier using the image metrics. The supervised models which are developed are further used to diagnosis and prescribe dialysis. Techniques like KNN, SVM, DT, RF, NB, and Logistics Regression are employed in the prediction modelling. Confusion matrix and other methods are used for the performance evaluation. The developed system serves as decision support system.

Methodology

Random Forest

The RF is an excellent method for classifying the multi-dimensional data. This can also perform regression, and other data modelling related operations. In the methodology, each DT initially classifies the (test) sample. This yields a class which has maximum number of occurrences. The sample is referred as Out-of-Bag (OOB) data. As a result, every tree will have an OOB data which is further used to evaluate the tree's error. In this process, the corresponding error is computed using the following expressions.

$$PE = P_{x,y}\big(mg(X,Y)\big) < 0 \quad \ldots (1)$$

Here, the function ‘mg’ can be represented as

$$mg(X,Y) = av_k I(h_k(X) = j) \quad \ldots (2)$$

The term I(*) refers to an indicator function. Similarly, the terms $h_1(x)$, $h_2(x)$….$h_k(x)$ are ensemble of classifiers.

The function mg(X, Y) computes the extent to which the average of votes typical at X & Y whenever the right class exceeds.

Especially for larger datasets with multiple dimensions in terms of input variables. The usual process of parameter suppression or minimization is also not mandatory taking the robustness of the algorithm. These variables are used to yield the prediction model in the structure of the algorithm, which is presented in Fig. 1 for better idea.

Machine Learning for Liver

Machine learning is an artificial intelligence branch in which a computer program learns from the

experience and its performance measure improves with the experience. It is possible to train the ML system to diagnose whether the person suffers from liver disease or not. However, there are certain gaps that the ML has to be taken care of, which are listed below.

a. Some tasks can be defined well only by examples.
b. Machine learning helps us to find the hidden correlations and relationships from the large amount of data.
c. Gives better results for prediction and model generation.
d. Environments may change sometimes.
e. Some problems with a huge amount of knowledge too hard for humans to be described.

The block diagram of the adopted classification methodology using the ML is shown in Fig. 2. Classification is a process that predicts a specific result in the presence of given input information. There is a requirement of one training set consisting of attributes and the respective outcomes called the goal attribute to determine the specified output. The algorithm tries to find connections between the attributes that would make it conceivable to anticipate the result. Then an unknown dataset is provided to the algorithm which consists of same set of attributes except for the attribute which is not known. The algorithm analyzes the input information and produces the respective outcome. There is a database of medical information related to liver disease where the prediction attribute is whether the patient suffers from liver disease or not. Classification is of two types: Supervised Classification (SC) and Unsupervised Classification (USC). In the SC, the corresponding main method extracts knowledge from database. Here, the database refers to set of training examples. These are known previously while in USC, training examples. Typically, the classification has two phases namely training and testing phases. In the training phase, the dataset trains the classifier. The other is Testing phase where testing of classifier is done to analyze its performance using different samples of the test set. Prediction accuracy is a criterion to evaluate the performance of classifier. Classification accuracy describes the percentage of instances which are correctly classified.

Algorithm 1 Random Forest

Precondition: A training set $S := (x_1, y_1), \ldots, (x_n, y_n)$, features F, and number of trees in forest B.

```
function RandomForest(S, F)
    H ← ∅
    for i ∈ 1, ..., B do
        S^(i) ← A bootstrap sample from S
        h_i ← RandomizedTreeLearn(S^(i), F)
        H ← H ∪ {h_i}
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    At each node:
        f ← very small subset of F
        Split on best feature in f
    return The learned tree
end function
```

Fig. 1 — RF pseudo-code

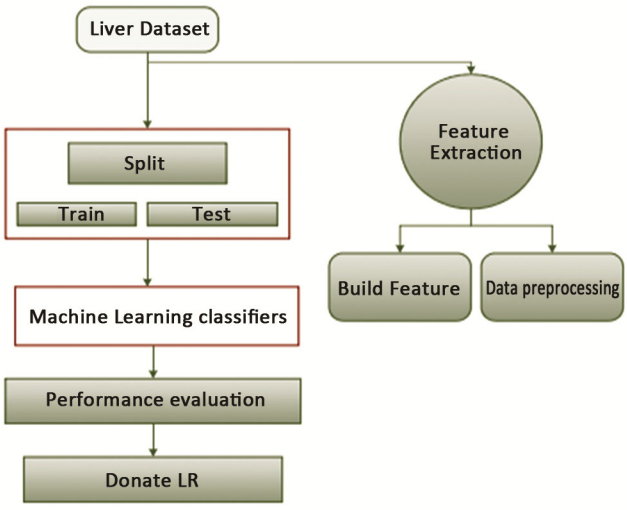

Fig. 2 — Block diagram of ML based method

Results and Discussion

The proposed method uses a dataset with several headers related to the non-clinical metrics of several tests pertaining to liver disease. The initial observation of the dataset finds that the data type is uniform with all the data. In the line to prepare the data for processing, data cleaning is the next step after careful analysis of the dataset. Data cleaning step involves careful observation visible data quality issues and rectify them. This follows a check for Nan data, duplicate data and convert certain columns to required data types. Some features of the dataset are total bilirubin, alkaline phosphatase, alamine aminotransferase, total protein, albumin-globulin ratio, gender, and finally whether the subject has liver disease or not.

The simulation results about the method incorporated using the ML and the outcomes are presented in terms of various observations in this Section. Initially, the objective function behavior over the progress in trials and the corresponding least square is studied. This is illustrated in its response to

fitness as shown in Fig. 3. The figure is watershed plot of the observed and predicted points pertaining to the model using the RF. The response of the objective function to the number of trials (numT), and the corresponding minimum least square (minLS) value are plotted to form a three-dimensional representation. The non-linear diminishing of the error is noticeable with progress in trial.

Further, the evaluation is extended using the plot presented in Fig. 4. The minimum estimated and observed objective magnitudes dependency can be inferred to be swinging only to reach to the best optimal minimum value while the number of trees grown is close to 120. Further, the performance of the RF algorithm can be analysed using this plot. Out of 580 data, about 165 are tested negative while the remaining are positive for the disease. The out-of-bag classification error is calculated for every increment in the tree grown. Hence, it is possible to accurately predict the patient status based on the data set employed.

A true positive is an outcome where the model correctly predicts liver disease. Similarly, a true negative is an outcome where the model correctly predicts the no disease case. Further, the false positive is an outcome where the model incorrectly predicts the non-disease case as disease. The plot in Fig. 5 is a Receiver Operating Curve (ROC) graph with the false positive rate on the X axis and the true positive rate on the Y axis. The point (0, 1) is the perfect classifier that classifies all positive cases and negative cases correctly. It is (0, 1) because the false positive rate is 0 (none), and the true positive rate is 1. The ROC plot is essential to understand the diagnostic ability of the classifier. The Classification of the data is pictorially presented in Fig. 5 & Fig. 6.

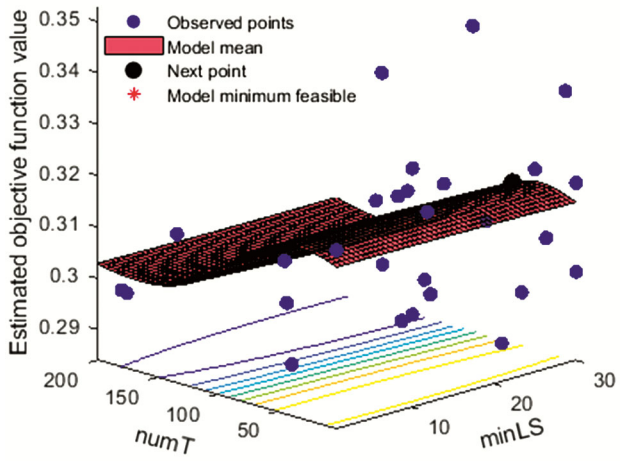

Fig. 3 — Response of objective function

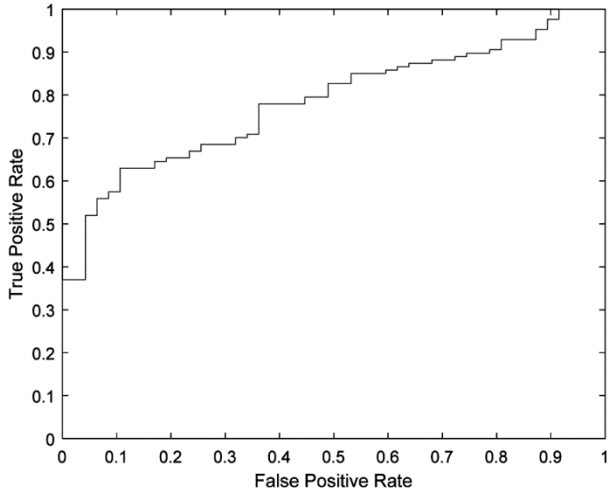

Fig. 5 — ROC description

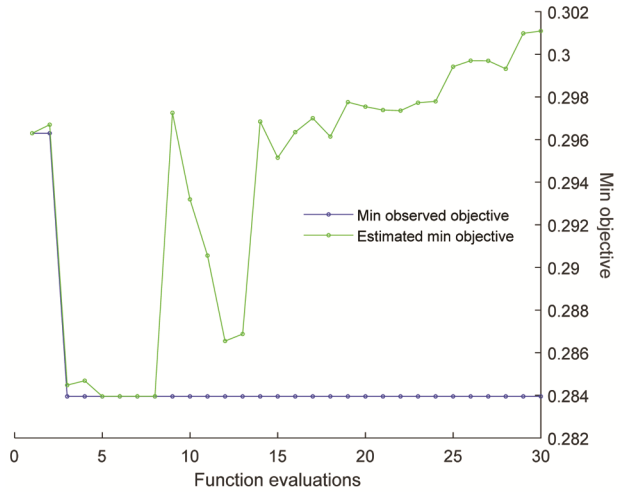

Fig. 4 — Observed and estimated objective values

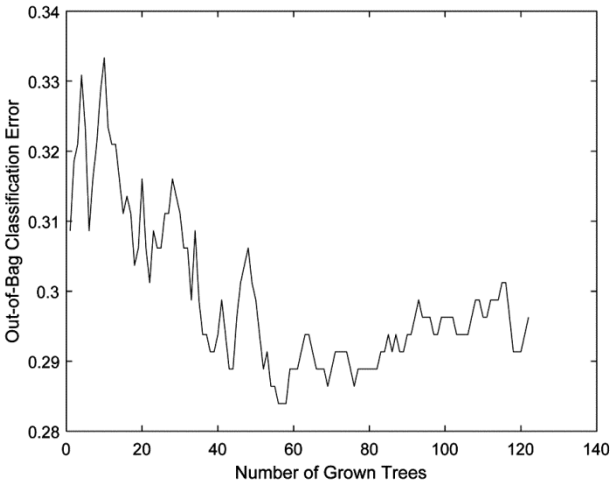

Fig. 6 — Performance of RF algorithm

Table 1 — Performance evaluation

Metric	Value
Accuracy model	75.8391
Recall	0.9449
Precision	0.7407
Specificity	0.1064

Table 2 — Confusion matrix

True class		
	120	7
	42	5
	Predicted class	

Metrics like accuracy model, recall, precision, and specificity are computed to evaluate the performance of the algorithm in numerical terms. The accuracy is approximately 79% which is convincingly good as a prediction model of the liver disease. The recall measured as the ability of a classification model to identify all data points in a relevant class. In this case it is computed as 0.9449 which is good value. Similarly, the precision is reported to be 0.7407 and is the ability of a classification model to return only the data points in a class. Similarly, the corresponding specificity computed as 0.1064. The metrics are tabulated in Table. 1 and the confusion matrix is presented in Table. 2. The total number of truly predicted cases can be read as 125 with true positive counting to 120 and true negative is 5. The false predictions are much less than true prediction thereby, claiming the model accuracy to be acceptable.

Conclusions

The ML-based system for the early prediction of liver disease based on the Indian dataset has been successfully developed using the RF algorithms. The performance analysis of the technique is presented in terms of metrics for evaluation. The performance of predicting the positive cases from the dataset is approximately 95% which is evident from the recall metrics. Further, the precision reported to be approximately 74% emphasizing that the performance of predicting the positive class which are actually positive. The core objective of the work is to create an advanced tool that can precisely provide means of treatment along with superior decisions in complex situations. The technique can be utilized to automate the process and early detection to get rid of chronic disease. The algorithm could successfully conceive a model successfully on the available Indian liver disease dataset. However, one critical concern is the volume of the dataset which is comparatively small. This challenges the robustness of the model. Enhancing the dataset and developing more complex model would be a best scope of future work.

References

1 Purushottam K S & Sharma R, Efficient heart disease prediction system, *Procedia Comput Sci*, **85** (2016) 962–969.

2 Singh P, Singh S & Pandi-Jain G S, Effective heart disease prediction system using data mining techniques, *Int J Nanomed*, **13** (2018) 121–124.

3 *Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC*. [Online]. [https://www.cdc.gov/ kidneydisease/ basics.html. (Accessed: 12 Dec 2018)].

4 Ahmed M R, Khatun M A, Ali A & Sundaraj K, A literature review on NoSQL database for big data processing, *Int J Eng Technol*, **7(2)** (2018) 902–906.

5 Hossain R, Mahmud S M H, Hossin M A, Noori S R H & Jahan H, PRMT: Predicting risk factor of obesity among middle-aged people using data mining techniques, *Procedia Comput Sci*, **132(1)** (2018) 1068–1076.

6 Dwivedi A K, Analysis of computational intelligence techniques for diabetes mellitus prediction, *Neural Comput Appl*, **30** (2017) 1–9.

7 Mahmud S M H, Hossin Md A, Ahmed R, Noori S R H & Sarkar Md N I, Machine learning based unified framework for diabetes prediction, *Int Conf Big Data Eng Technol* (Chengdu, China) 2018, 46–50. https://doi.org/10.1145/ 3297730.3297737

8 Heydari M, Teimouri M, Heshmati Z & Alavinia S M, Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran, *Int J Diabetes Dev Ctries*, **36(2)** (2016) 167–173.

9 Kukar M, Kononenko I & Grošelj C K K A, Intelligence in, and undefined 1999, Analyzing and improving the diagnosis of is chaemic heart disease with machine learning (Elsevier), doi: 10.1016/s0933-3657(98)00063-3

10 Jain D & Singh V, Feature selection and classification systems for chronic disease prediction: A review, *Egypt Informatics J*, **19(3)** (2018) 179–189, https://doi.org/10.1016 /j.eij.2018.03.002

11 Carvalho D, Pinheiro P R & Pinheiro M C D, A hybrid model to support the early diagnosis of breast cancer, *Procedia Comput Sci*, **91** (2016) 927–934.

12 Kumari M, Breast cancer prediction system, *Procedia Comput Sci*, **132** (2018) 371–376.

13 Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O & Poorolajal J, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, *Clin Epidemiology Glob Health*, **7** (2019) 293–299.

14 Asri H, Mousannif H, Moatassime H Al & Noel T, Using machine learning algorithms for breast cancer risk prediction and diagnosis, *Procedia Comput Sci*, **83** (2016) 1064–1069.

15 Aganathan, K, Tayara H & Chong K T, Prediction of drug-induced liver toxicity using SVM and optimal descriptor sets, *Int J Mol Sci*, **22** (2021) 8073.

16 Phan D V, Chan C L, Li A A, Chien T Y & Nguyen V C, Liver cancer prediction in a viral hepatitis cohort: A deep learning approach, *Int J Cancer*, **147** (2020) 2871–2878.

17 Joloudari J H, Saadatfar H, Dehzangi, A & Shamshirband S, Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection, *Inform Med*, **17** (2019) 100255.
18 Amrane M, Oukid S, Gagaoua I & Ensari T, Breast cancer classification using machine learning, *Electric Electronics, Computer Science, Biomedical Engineerings Meeting (EBBT)* (IEEE) 2018, 1–4.
19 Al-Hadidi M R, Alarabeyyat A & Alhanahnah M, Breast cancer detection using knearest neighbor machine learning algorithm, *9th Int Conf Dev eSyst Eng (DeSE)* (IEEE) 2016, 35–39. DOI 10.1109/DeSE.2016.8
20 Liu X, Wang X, Su Q, Zhang M, Zhu Y, Wang Q & Wang Q A, A hybrid classification system for heart disease diagnosis based on the RFRS method, *Comput Math Methods Med*, **2017** (2017) 8272091.
21 Ramana B V, Babu M S P & Venkateswarlu N B, A critical comparative study of liver patients from USA and INDIA: an exploratory analysis, *Int J Comput Sci Eng*, **9(3)** (2012) 694–0814.