Indonesian Journal of Electronics, Electromedical, and Medical Informatics (IJEEEMI)
Vol. 3, No. 1, February 2021, pp. 9-14
DOI: 10. 35882/ijeeemi.v3i1.2 **ISSN: 2656-8624**

Support Vector Machine And K-Nearest Neighbor Based Liver Disease Classification Model

Tsehay Admassu Assegie
Department of Computer Science, Faculty of Computing Technology, Aksum Institute of Technology, Aksum University, Aksum, 1010, Ethiopia

Article Info

Article History:
Received Feb 5, 2021
Revised Feb 19, 2021
Accepted Feb 21, 2021

Keywords*:*
Machine learning
Support Vector Machine
K-Nearest Neighbor
Liver disease prediction
Liver disease classification

Abstract

Machine-learning approaches have become greatly applicable in disease diagnosis and prediction process. This is because of the accuracy and better precision of the machine learning models in disease prediction. However, different machine learning models have different accuracy and precision on disease prediction. Selecting the better model that would result in better disease prediction accuracy and precision is an open research problem. In this study, we have proposed machine learning model for liver disease prediction using Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) learning algorithms and we have evaluated the accuracy and precision of the models on liver disease prediction using the Indian liver disease data repository. The analysis of result showed 82.90% accuracy for SVM and 72.64% accuracy for the KNN algorithm. Based on the accuracy score of SVM and KNN on experimental test results, the SVM is better in performance on the liver disease prediction than the KNN algorithm.

Corresponding Autor:

tsehayadmassu2006@gmail.com,
Department of Computer Science,
Aksum University, Aksum, Ethiopia

I. Introduction

Liver disease is one of the life threating diseases in the world. In some cases, the disease can be threated but, if left untreated [1-6] and when the treatment is not in early stages of the disease event the liver disease causes death. This disease is a cause of deaths in most of the developing countries. The reason behind is lack of proper treatment and diagnosis and the severity of the disease. In some cases, the disease can be threated if detected early.

The detection or identification of the disease requires specialist who have a thorough understanding or experience on threating the disease and at the same time the patient's capacity in paying the amount that is required for getting treatment. In most of the cases, the physician faces complexity in identifying the disease precisely due to the overlapping symptoms with other diseases. The decision making process during the identification of the disease can be aided with machine learning models. The use of machine learning algorithms such as SVM and KNN in prediction of the liver disease is prevalent in reducing the consequence of the disease. Machine learning is gaining greater importance in object recognition, recommendation systems, handwritten digits recognition [1, 2, 3, 5, 6] and disease prediction and diagnosis due to the improved accuracy and precision results. Now days, machines are playing greater role in assisting physicians to make the right decision in identification and diagnosis of diseases.

Different machine learning algorithms namely, Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayes (NB) Convolutional neural Network (CNN), Decision Tree algorithms and K-Nearest Neighbor (KNN) [2-12] are used in building the models that are being used in prediction of liver disease. The most important concern of the prediction models is to minimize the death rate due to the liver disease hereby, providing better result in disease prediction. The machine learning models assist in the early identification of the disease. The problems in liver disease diagnosis are the following: Difficulty in identifying the liver disease patient accurately due to the over lapping symptoms of the liver disease with other diseases. 2) The severity of the disease increases if not identified early and leads to complications and even deaths. 3) The time taken to identify the liver disease using laboratory test manually is more and the patient may suffer in the meantime. To solve the problems stated in 1), 2) and 3) this paper proposes a machine-learning model to simplify the liver disease identification and treatment process. Hence, this study explores the answer to the following questions: 1. To develop a liver disease prediction model based on SVM and KNN learning algorithms for early detection of a liver disease and moreover, assist the physicians in the decision making process. 2. To evaluate and validate the performance and the accuracy of the proposed liver disease prediction model. 3. To compare the performance of SVM and KNN using accuracy, precision and confusion matrix as performance metrics? 4. To explore the behavior of cross validation and training score for varying samples of training set.

II. LITERATURE REVIEW

This section will focus on the literature related to machine learning algorithms for liver disease prediction. In [3] the authors proposed a machine-learning model for predicting liver disease by using SVM classification. In their study, the authors considered texture attribute as the primary attribute contributing to the occurrence of the liver disease. The SVM classification algorithm has provided better accuracy for prediction of the liver disease with 68.75% accuracy. In another study [4], artificial neural network (ANN) is applied to build a machine-learning model for diagnosis of liver disease. The SVM classification algorithm provided an accuracy of 71% in the prediction of the liver disease. In [5] a liver diagnosis model is proposed using the decision tree classification algorithm. The authors used the random forest algorithm to build the proposed liver disease prediction or classification model. The UCI machine learning data repository is used in building the model. Liver disease attributes like the age, gender, total bilirubin, direct bilirubin and total bilirubin of the liver disease patients is used in training and testing. The proposed liver disease prediction model based on the decision tree classification-learning algorithm provided an accuracy of 69.30% in predicting the liver disease.

A comparative study conducted on various classification algorithms used to build liver disease prediction model in [6] showcased that, SVM provided better accuracy in classification of liver disease when compared to Naïve Bayes (NB), K-Nearest Neighbor (KNN) and the Artificial Neural Network (ANN) algorithms. Based on the literature review we have selected the SVM for building our proposed liver disease prediction model.

In [7] a Multi-layer perception (MLP) based liver disease prediction model is proposed. The MLP algorithm is applied to the 239 sample manually collected form clinical data repositories. The experimental result of this study showed that the MLP provided better performance compared to NB algorithm. In [7] a Multi-layer perception (MLP) based liver disease prediction model is proposed. The MLP algorithm is applied to the 239 sample manually collected form clinical data repositories. The experimental result of this study showed that the MLP provided better performance compared to NB algorithm. In another study on liver disease prediction using machine-learning algorithms, logistic regression, the KNN and SVM is used to build the proposed liver disease prediction model. And the performance of the three classification algorithms namely, logistic regression, KNN and SVM on liver disease prediction is compared [9]. The accuracy score and confusion matrix is used to analyze the performance of each classification algorithm and the result of comparative analysis showed that three of the algorithms provided a good prediction result with an accuracy falling in range 71%-73%.

In [10] SVM and back propagation algorithms are applied to develop a learning model for prediction of liver disease. The comparative analysis result on the accuracy score of the prediction models showed that back propagation model provided better performance than SVM with an accuracy score of 73.2% and SVM scored an accuracy of 71%. The accuracy shows there is still wider scope for improvement for better performance using SVM for liver disease classification.

The literature reveals that many different attributes contribute to the occurrence of liver disease. Among the contributing factors to the liver disease some are personal habits, such as smoking, consumption of alcohol and family history. To reduce the mortality rate caused by liver disease, SVM and NB based model for prediction of liver disease is proposed in [11-18] for assisting physicians' in the diagnosis of the liver disease. Various metrics such as accuracy score, precision and f-score measures is used to compare the performance of SVM and NB and the experimental analysis result showed that SVM provided better result in prediction of liver disease. In [12] SVM algorithm is used to develop machine-learning model for liver disease diagnosis. In the study, the performance of the algorithm is shown to be 73 %. The authors used the Indian liver disease data repository. The performance of the model varies for different number of features applied in training the SVM learning algorithm. The support vector machine algorithm is also proved to be a powerful machine-learning model for building liver disease prediction model as the model has a reasonable accuracy.

An artificial neural network based liver disease prediction model is proposed in [13-14]. The authors have also compared the proposed ANN based model with the existing models developed using learning algorithms such as SVM and NB. The result of comparative analysis showed that ANN provided an accuracy score of 70%. In [20], the authors proposed a classification model for liver disease prediction with Naive Bayes and support vector machine. The authors evaluated the performance of the Naïve Bayes and support vector machine with prediction accuracy. An experimental test reveals that the support vector machine performed better compared to the Naïve Bayes for liver disease classification. However, the support vector machine has slower execution time compared to the Naïve Bayes. In [21], the authors compared the performance of three classification algorithms, namely logistic regression, support vector machine and K-nearest neighbor. The experimental result of the study revealed that logistic regression has better performance in terms of accuracy. However, the study used only accuracy as performance metrics in the evaluation and other metrics such as receiver operating characteristics curve, (ROC curve) confusion matrix is not used. In [22], logistic regression-based liver disease classification model is proposed. The authors evaluated the performance of the model on liver disease test set and result shows that the model has 74% accuracy on liver disease classification. In their study, the authors also compared the logistic regression accuracy with other machine learning algorithms, namely the support vector machine and artificial neural network. Comparative analysis of the performance of these models reveals that the logistic regression has better accuracy on classification of liver disease as compared to the support vector machine and the artificial neural network. The artificial neural network has accuracy of 71% and the support vector machine has 58.26% as showcased by the authors. In [23], the authors compared support vector machine and K-nearest neighbor-based model on liver disease dataset classification. The authors suggested that support vector machine and K-nearest neighbor has acceptable performance for liver disease classification. Based on the literature survey [1-26], we chose the support vector machine and K-nearest neighbor for liver disease prediction in this study. In [27-31], the authors compared six machine learning algorithms namely, Naïve Bayes, KNN, SVM, decision tree, random forest and logistic regression for liver disease prediction. The experimental analysis of the result of the accuracy shows that the highest accuracy achieved by these algorithms is 75% with logistic regression.

III. Materials and Methods

In this section, the data repository being used in training and testing a liver disease prediction model and the machine learning algorithm as well as the programming language used in experimental tests and training the model is discussed. The data repository used in training and testing the proposed liver disease prediction model is the Indian liver disease data repository available online. The machine learning methodology used in building the proposed model for solving the liver disease classification problem is the SVM and KNN. To implement the proposed machine-learning model for prediction of liver disease and perform experimental analysis of performance of the SVM and KNN algorithms, a python programming language is used. The logical steps for liver disease prediction using support vector machine and KNN is shown in Fig. 1.

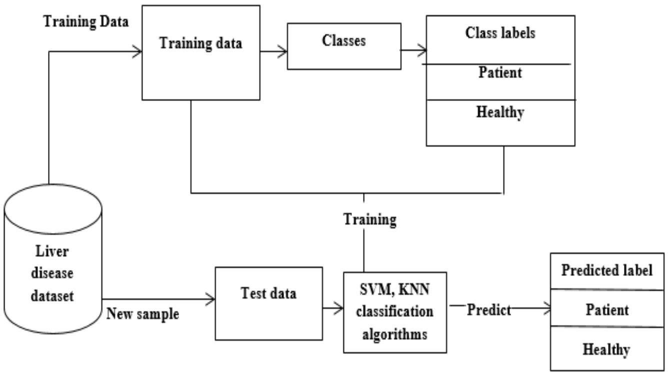

Fig. 1. SVM and KNN model for liver disease prediction

A. Dataset description

The detail of liver disease data repository used in building the liver disease prediction model is described in this section. The actual Indian liver disease data repository available online and the description of this data repository is summarized in table 1. The data repository consists of 583 data instances with 11 attributes and have two classes or labels, the liver disease positive (liver disease patients represented by 1) and liver disease negative (represented by 2). The attributes of liver disease are age (in years), sex (Male/Female), Albumin and so on described in table 2. In the data repository, 416 samples are liver disease patient instances and 167 liver disease positive or liver disease patient records. Form the total 583 samples in the data repository, the 20% (roughly 117 samples) are used in testing and the 80% (466 samples) of the samples are used in training the model in the data repository. The dataset features are demonstrated in table 1.

TABLE I. SAMPLE LIVER DATSSET FEATURE USED IN CLASSIFICATION.

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

TABLE II. LIVER DATSSET FEATURE DESCRPTION .

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

IV. RESULTS AND DISCUSSIONS

In the experimental analysis, we have considered three performance metrics to compare the support vector machine and the K-nearest neighbor and the trained model is tested on the liver disease test set. The performance metrics used in the experimental test include accuracy score, confusion matrix and receiver operating characteristic curve (ROC) curve. Overall, the result shows that the support vector machine performed well as compared to the K-nearest neighbor.

1. *Accuracy of the model*

Accuracy score is the most important metric to evaluate the performance of learning models. This metric is a measure of how a model fits unknown class label based on the known class labels given to the model in the training set. The performance of the SVM and KNN learning algorithms based liver disease prediction model was evaluated using accuracy metric. The model is tested randomly on unknown samples and the analysis result showed an average accuracy of 74.52% on five random tests performed on the model using SVM and an average accuracy of 70.93% for KNN algorithm. The accuracy plot of the two models on random test is illustrated in figure 3. As shown in figure 3 the Support Vector Machine (SVM) has better performance than the K-nearest Neighbor (KNN) algorithm on the random tests on the models. The accuracy for the SVM falls in the range 70% to 82% whereas that of the KNN falls in the range 68% to 74%.

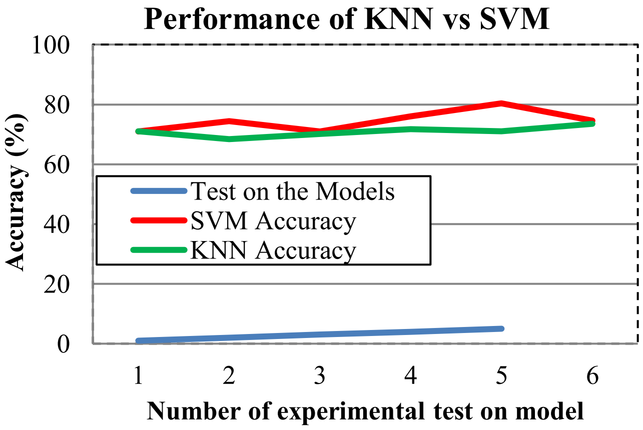

Fig. 2. Accuacry of SVM and KNN for liver disease preidction

2. *Reciver operating characteristics curve (ROC)*

The receiver operating characteristic curve is used to compare the KNN and SVM model on specific class namely the true positive (TP) or the liver disease patient class. the ROC metric is important to compare how the SVM and KNN models perform on predicting the true positive class or liver disease patient as this this the important class which we require the model to perform well.

The receiver operating characteristics or ROC curve for the support vector machine is demonstrated in figure 4. The ROC curve metric is used to evaluate the output quality or the performance of support vector machine with cross validation. This metrics validates the performance of the support vector machine using true positive rate (TPR) and false positive rate (FPR). As demonstrated in Fig.4, the TPR is y-axis and the FPR is the x-axis, which means the more the ROC, curve area the better the performance of the model. The area under curve is 0.66 for liver disease patient and non-patient classes. This value is higher than the ROC curve area for the K-Nearest Neighbor model demonstrated in Fig.4. Hence, we can conclude that the support vector machine better in predicting liver disease TP cases meaning the observations where the class is liver disease patient as compared to the K-Nearest Neighbor model.

In Fig.4, the receiver operating characteristics curve (ROC) curve of K-Nearest Neighbor (KNN) is illustrated. As illustrated in figure 6, the ROC curve is steep as compared to the support vector machine (ROC curve) which is vertical as compared to the KNN ROC curve. Moreover, the area under curve for the KNN model is less compared to the SVM. Hence, the KNN model performs less on predicting the True Positive (TP) or liver disease patient class compared to the SVM model

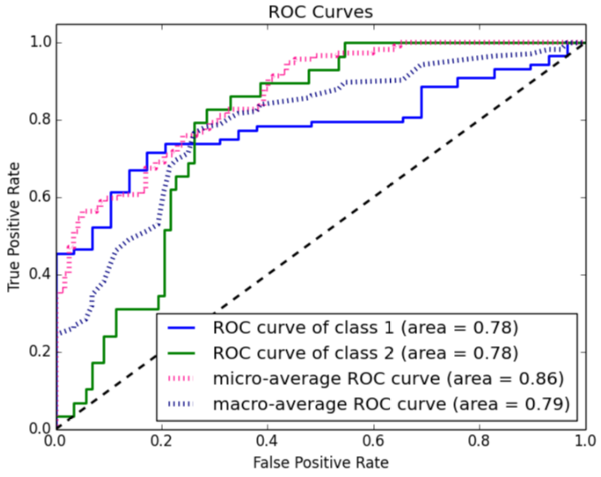

Fig. 3. ROC Curve of SVM model

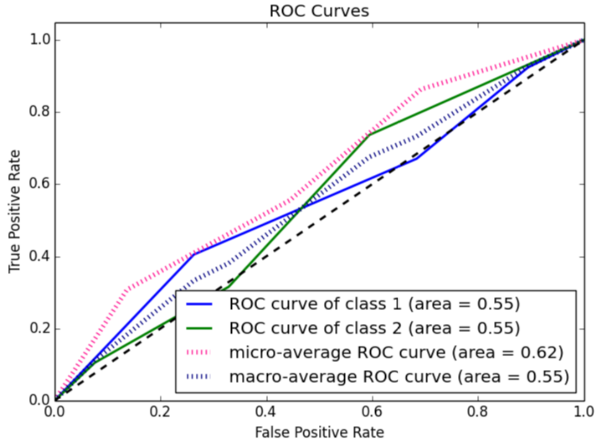

Fig. 4. ROC curve of KNN for liver disease preidction

3. *Confusion matrix for KNN and SVM on liver disease classification*

The confsuion matrix shows the real and predictive heart disease dataset observations by the KNN and SVM model as demonstrated in figure 5 and 6 respectively. The confusion matrix of KNN and SVM demonstarted in figure 5 and 6 respectively shows that the SVM model has better classification accuracy with total of 86 observations and 31 miss-classifications. However, the KNN model has 84 truly classified observations and 34 mis-classification or classification errors. Hence, the performance of SVM is better as comapred to the KNN model which produced more number of mis-classification or classification errors than SVM model for liver disease classifcation.

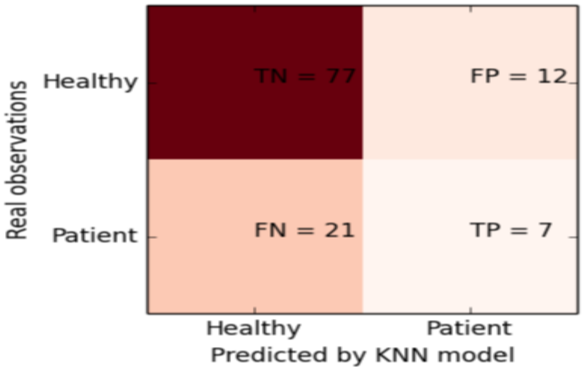

Fig. 5 Confusion matrix for KNN model on liver disease classification

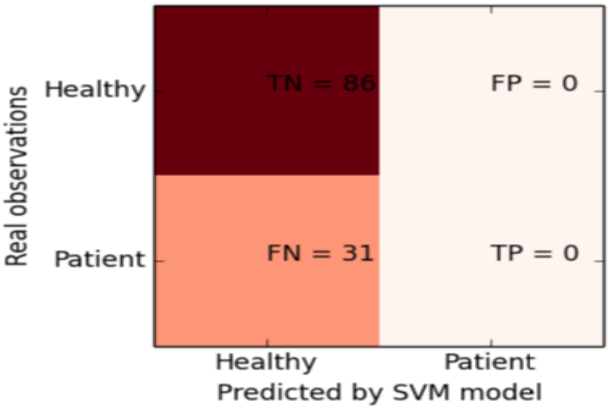

Fig. 6. Confusion matrix for SVM model on liver disease classification

V. Conclusion

This study compared two machine learning algorithms namely, support vector machine (SVM) and K-nearest Neighbor (KNN) for liver disease prediction. We have compared them with accuracy and confusion matrix and ROC curve on test dataset. Result shows that the accuracy of the models on prediction of liver disease is 75.21% for SVM and 70.93% for KNN. Overall, SVM algorithm is better for predicting liver disease compared to KNN. The comparative result between SVM and KNN with confusion matrix shows that the number of true labels is greater for SVM than the number of true label for KNN. An experimental result analysis with ROC curve reveals that SVM is better in predicting the liver patient class as compared to the KNN. Overall, SVM is better algorithm for heart disease classification as compared to the KNN algorithm.

VI. Recommendations and Future Work

In this study, we have proposed a model for liver disease classification with support vector machine and k-nearest neighbor algorithm. Moreover, the study compared the performance of support vector machine on Indian liver dataset collected from online data repository. The dataset consists of 11 features and 583 observations, which is limited for machine learning algorithm to perform effectively on liver disease classification. Overall, we recommend the design and implementation of a more accurate liver disease classification model with more dataset and features.

References

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

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

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

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

[5] Shambel Kefelegn, Pooja Kamat, Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey, International Journal of Pure and Applied Mathematics Volume 118 No. 9 2018.

[6] Yuan Cao, Zhi-De Hu, Xiao-Fei Liu, An-Mei Deng, Cheng-Jin Hu, An MLP Classifier for Prediction of HBV-Induced Liver Cirrhosis Using Routinely Available Clinical Parameters, Hindawi Publishing Corporation Disease Markers Volume 35, 2013.

[7] A Novel Computer-Aided Diagnosis Framework Using Deep Learning for Classification of Fatty Liver Disease in Ultrasound Imaging, 20th International Conference on e-Health Networking, Applications and Services (Healthcom), IEEE, 2018.

[8] Kanza Hamid, Amina Asif, Machine Learning with Abstention for Automated Liver Disease Diagnosis. International Conference on Frontiers of Information Technology, IEEE. 2017.

[9] Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury, Prediction of Liver Disease using Classification Algorithms, 4th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018.

[10] Sumedh Sontakke, Jay Lohokare, Reshul Dani, Diagnosis of Liver Diseases using Machine Learning, International Conference on Emerging Trends & Innovation in ICT (ICEI) Pune Institute of Computer Technology, Pune, India, Feb 3-5, IEEE, 2017.

[11] Dr. S. Vijayaran, Mr.S.Dhayanand, Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

[12] Esraa M. Hashem, Mai S. Mabrouk, A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis, American Journal of Intelligent Systems 2014.

[13] Ebenezer Obaloluwa Olaniyi, Khasman Adan, Liver Disease Diagnosis Based on Neural Networks, Advances in Computational Intelligence, 2017.

[14] Assegie Tsehay Admassu, A support vector based heart disease prediction, journal of software engineering and intelligent systems December 2019.

[15] Assegie Tsehay Admassu, Sushma S. J, A Support Vector Machine and Decision Tree Based Breast Cancer Prediction, International Journal of Engineering and Advanced Technology, February 2020.

[16] Assegie Tsehay Admassu, Pramod Sekharan Nair, The Performance Of Different Machine Learning Models On Diabetes Prediction, Inernationa Journal of Scientifc and Technology Research, January, 2020.

[17] Assegie Tsehay Admassu, Sushma S J, Dr. Prasanna Kumar S C, Weighted Decision Tree Model for Breast Cancer Detection, Technology Reports of Kansai University, Volume 62, Issue 03, January, 2020.

[18] S. Vijayarani, S. Dhayanand, Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

[19] Thirunavukkarasu K, Ajay S. Singh, Md Irfan , Abhishek Chowdhury, Prediction of Liver Disease using Classification Algorithms, 4th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2018.

[20] Syed Hasan Adil1, Mansoor Ebrahim, Kamran Raza, Liver Patient Classification using Logistic Regression, 4th International Conference on Computer and Information Science, IEEE, 2018.

[21] Mehtaj Banu H, Liver Disease Prediction using Machine-Learning Algorithms, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6, August 2019.

[22] Esraa M.Hashem, Mai S. Mabrouk, A Study of support vector machine algorithm for liver disease diagnosis, Biomedical Engineering, Misr University for Science and Technology (MUST University), 6th of October, Egypt, 2019.

[23] Ashfaq Ahmed K, Sultan Aljahdali, Syed Naimatullah Hussain, Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques, International Journal of Computer Applications (0975 – 8887) Volume 69– No.11, May 2013.

[24] Nazmun Nahar, Ferdous Ara, Liver disease prediction by using different decision tree techniques, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018.

[25] A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossain, A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms, International Journal of Scientific and Technology Research, 2019

[26] H. A. Dian Permana, Mada Sanjaya W.S., “Desain dan Implementasi Perancangan Elektrokardiograf (EKG) berbasis Bluetooth,” vol. 2, no. 1, p. 407, 2015.

[27] C. F. C. S. C. Tai,* C. W. Chang, “Designing Better Adaptive Sampling Algorithms for ECG Holter Systems,” *Pan Am. Heal. Care Exch. PAHCE 2011 - Conf. Work. Exhib.*, vol. 44, 1997.

[28] D. Lucani, G. Cataldo, J. Cruz, G. Villegas, and S. Wong, “A portable ECG monitoring device with Bluetooth and Holter capabilities for telemedicine applications,” *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 5244–5247, 2006.

[29] H. Jin and B. Miao, “Design of Holter ECG system based on MSP430 and USB technology,” *2007 1st Int. Conf. Bioinforma. Biomed. Eng. ICBBE*, pp. 976–979, 2007.

[30] A. Juarez-Carrasco and J. E. Chong-Quero, “Design and development of a holter prototype with Bluetooth transmission,” *Pan Am. Heal. Care Exch. PAHCE 2011 - Conf. Work. Exhib. Coop. / Linkages An Indep. Forum Patient Care Technol. Support*, pp. 323–327, 2011.

[31] J. Lee, D. D. McManus, S. Merchant, and K. H. Chon, “Automatic motion and noise artifact detection in holter ECG data using empirical mode decomposition and statistical approaches,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1499–1506, 2012.