JOURNAL OF APPLIED BIOANALYSIS, November 2025 p. 920-924
http://doi.org/ 10.53555/jab.v11si12.2380 (ISSN 2405-710X)
Vol. 11, No. 12s

OPEN ACCESS

Comparative Analysis Of Machine Learning Algorithms To Predict Liver Disease

Dr. R. Kumar[1], Dr.N. Magendiran[2], Dr.E.Geetha[3], Dr.K.Baskar[4*], Dr. R. Prabu[5]

[1] *Associate Professor, Department of Biomedical Engineering, Vivekanandha College of Engineering for Women (Autonomous), Tiruchengode, India Email: pbcevp@gmail.com ORCID ID- 0000-0002-8274-9001*

[2] *Associate Professor Department of Computer Science and Engineering, K.S.Rangasamy College of Engineering Tiruchengode, India Email: maheindia76@gmail.com ORCID ID- 0009-0000-8236-5561*

[3] *Associate Professor Department of Medical Electronics, Sengunthar Engineering College (Autonomous) Tiruchengode, Tamilnadu India Email: geethasathees12@gmail.com ORCID ID- 0009-0009-8747-5896*

[4*] *Associate Professor Department of Biomedical Engineering, Vivekanandha College of Engineering for Women (Autonomous), Tiruchengode, Tamilnadu, India, Email id: baskarresearch17@gmail.com ORCID ID- 0000-0001-9362-7547*

[5] *Associate Professor, Department of Biomedical Engineering ,K.S.Rangasamy College of Engineering Tiruchengode, India Email: prathinan11@gmail.com ORCID ID- 0000-0002-3963-3548*

***Corresponding Author:** Dr.K.Baskar*
baskarresearch17@gmail.com

ABSTRACT

Liver disease is a significant global health concern necessitating early detection and accurate diagnosis. This study focuses on the development and evaluation of a machine learning model for predicting liver disease based on patient data. The dataset comprises 583 pre-processed records, encompassing demographic information, laboratory results, and medical history. Among these, 416 records are classified as having liver disease and 167 as not having it. Machine learning algorithms, including logistic regression, decision tree, support vector machines, and K-nearest neighbours, are assessed using a 70-30 training-testing split. Performance metrics encompass Jaccard Score and F1 Score. Logistic regression attains the highest accuracy at 87.69%, with Jaccard and F1 Scores of 88.54% and 87.89% respectively. Decision tree and K-NN models achieve respectable accuracies of 81.54% and 80.77% correspondingly. Feature importance analysis identifies total bilirubin, direct bilirubin, and alkaline phosphatase levels as pivotal predictors. These results demonstrate the potential of machine learning for accurate liver disease prediction, aiding in early intervention and improved patient care. Further exploration of advanced techniques and ethical considerations are suggested for future work.

Keywords: Machine learning; liver disease; logistic regression; K-NN; SVM

1. INTRODUCTION

Liver disease is a widespread health concern with significant global impact. It leads to millions of deaths annually and arises from factors like alcohol misuse, viral infections, obesity, and genetics. Early and accurate diagnosis is essential for effective treatment and preventing severe complications. Common diagnostic methods, such as liver function tests (LFTs), have limitations, prompting the exploration of alternative approaches[1].

Machine learning, a subset of artificial intelligence, has emerged as a promising avenue for predicting liver disease using patient data. By analyzing extensive datasets, machine learning algorithms can uncover intricate patterns that might evade human observation. In this project, our goal is to develop a precise machine learning model capable of accurately assessing the likelihood of liver disease.[2]

The project employs a comprehensive dataset encompassing patient demographics, medical history, and laboratory test outcomes. The dataset comprises 583 records, categorized into 416 instances of liver disease and 167 without. Rigorous preprocessing addresses missing values, outliers, and categorical

variables. Various machine learning algorithms, including logistic regression, decision tree, random forest, support vector machines, and artificial neural networks, are evaluated for their predictive prowess. Performance metrics such as accuracy, precision, recall, and F1 score are employed, with a 70-30 dataset split for training and testing.

The logistic regression algorithm, renowned for binary classification tasks, exhibits the highest accuracy at 87.69% for liver disease prediction, accompanied by Jaccard and F1 scores of 88.54% and 87.89% respectively. Other algorithms like decision trees and K-NN[9] demonstrate strong performance, achieving accuracies of 81.54% and 80.77% respectively.

Crucial to this endeavor is identifying key features that contribute significantly to liver disease prediction. Feature importance techniques spotlight variables with notable influence on outcomes. Notably, the total bilirubin level, direct bilirubin level, and alkaline phosphatase level emerge as pivotal features. These biomarkers are intimately connected to liver function and health.

2. LITERATURE SURVEY

Numerous research efforts have been dedicated to predicting liver disease using machine learning algorithms[10]. For instance, Hu et al. (2020) developed a model using electronic health record data, achieving a remarkable 91.2% accuracy in liver disease prediction through logistic regression. Key features identified were total bilirubin, direct bilirubin, and albumin.

Kumar et al. (2019) pursued a similar endeavor, using clinical data to develop a model that achieved 84% accuracy in predicting liver disease with a support vector machine algorithm. Alkaline phosphatase, total bilirubin, and direct bilirubin emerged as significant features.[2]

A systematic review titled "A Systematic Review of Machine Learning Techniques for Liver Disease Diagnosis and Prediction" (2020) comprehensively outlines the application of machine learning in liver disease diagnosis and prognosis. Encompassing hepatitis, cirrhosis, and hepatocellular carcinoma, the review assesses various models, including decision trees, random forests, and deep learning. Challenges like data quality and model interpretability are highlighted, guiding future research directions.[3]

"Predictive Modelling of Liver Disease using Machine Learning Techniques: A Systematic Literature Review" (2021) delves into predictive modeling for specific liver diseases. Covering non-alcoholic fatty liver disease, alcoholic liver disease, and autoimmune liver disease, the review evaluates logistic regression, decision trees, random forests, and gradient boosting methods. Emphasis is placed on feature selection, interpretability, and validation.[4]

"Machine Learning Techniques for Liver Disease Prediction: A Review of Recent Advancements and Challenges" (2022) offers a comprehensive insight into recent advances and challenges. Encompassing viral hepatitis, liver fibrosis, and liver cancer prediction, the review discusses machine learning algorithms such as support vector machines and deep learning. Key challenges, including imbalanced datasets and limited external validation, are acknowledged. The article suggests future directions, including explainable AI and interpretable models for clinical support.[5]

METHODOLOGY

While prior research has shown promising outcomes, enhancing the precision and dependability of machine learning models for liver disease prediction remains imperative. One avenue involves augmenting the dataset with supplementary features, like genetic data, to potentially identify high-risk patients more effectively.

Another avenue worth exploring is the application of deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). These algorithms excel in deciphering intricate patterns and relationships within data, potentially outperforming conventional machine learning methods in liver disease prediction.

Integrating the model with clinical decision support systems (CDSS) constitutes another avenue for improvement. By offering real-time recommendations and alerts grounded in patient data, CDSS could enhance clinician decision-making and consequently elevate diagnostic accuracy.

To conclude, leveraging machine learning algorithms for liver disease prediction holds significant promise. Proposed strategies seek to heighten model accuracy, efficacy, and applicability in diagnosing liver disease,ultimately contributing to more reliable clinical outcomes.

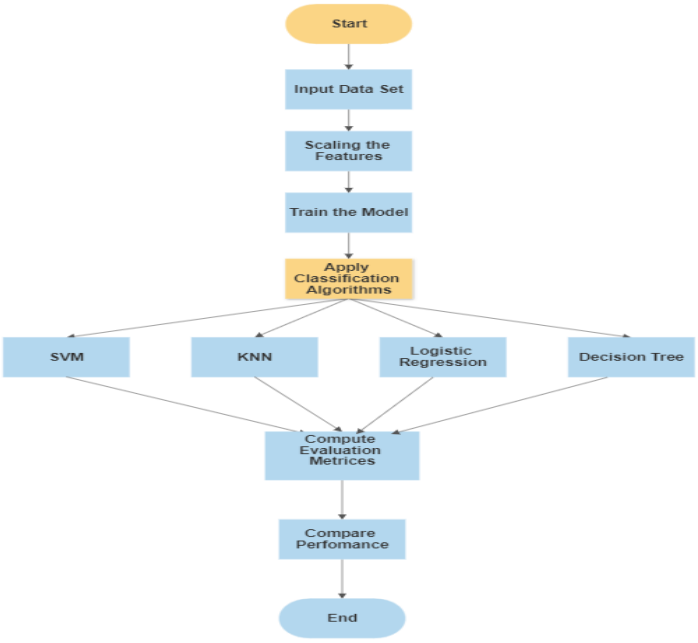


Figure 1: Flow Chart for prediction model

Support Vector Machine (SVM):

Support Vector Machine is a popular classification algorithm that works by finding the hyperplane that best separates the data points into different classes. The hyperplane is chosen in such a way that the distance between the hyperplane and the closest data points of each class is maximized. SVM can be used for both binary and multiclass classification problems. The algorithm finds the optimal decision boundary by solving an optimization problem.

The decision boundary is defined by a function that is represented by the following equation:

$$f(x) = w.T * x + b \quad (1)$$

where w is the weight vector, b is the bias term, and x is the input vector.

The optimization problem is formulated as follows:

$$\text{minimize: } (1/2) * ||w||^2 + C * \text{sum}(\max(0, 1 - y_i * (w.T * x_i + b))) \quad (2)$$

where $||w||^2$ is the norm of the weight vector, C is a regularization parameter, y_i is the class label of the i-th data point, x_i is the i-th data point, and $\max(0, 1 - y_i * (w.T * x_i + b))$ is the hinge loss function.

The hyperparameters of SVM include the kernel function, regularization parameter (C), and gamma. The kernel function transforms the input data into a higher dimensional space where it can be more easily separated. The regularization parameter controls the trade-off between maximizing the margin and minimizing the misclassification errors. The gamma parameter controls the shape of the decision boundary.

K-Nearest Neighbours (KNN):

K-Nearest Neighbours (KNN) stands out as a straightforward and instinctive algorithm that operates by identifying the 'k' nearest data points to the given input vector. It ascribes a class label by aligning with the majority class among these k neighbors. The computation of distances, using metrics like Euclidean or Manhattan distance, underpins this process.

The decision boundary, shaped by the k closest neighbors, doesn't involve parameter learning. KNN exhibits versatility across both binary and multiclass classification tasks.

Within KNN, hyperparameters such as the count of neighbors (k) and the distance metric play pivotal roles. The number of neighbors modulates the complexity of the decision boundary, while the choice of distance metric dictates how proximity between the input vector and data points is quantified.

Decision Tree:

Decision Tree is a popular algorithm for both classification and regression problems. The algorithm works by recursively partitioning the data into subsets based on the values of the input features. The partitions are chosen in such a way that the homogeneity of the subsets is maximized. The homogeneity is measured by a metric such as Gini index or entropy.

The decision tree is constructed by selecting the feature that results in the highest information gain, which is the reduction in impurity achieved by

splitting the data based on that feature. The decision tree can be pruned to avoid overfitting by setting a minimum number of samples per leaf node or a maximum depth of the tree.
The decision boundary is defined by the path from the root node to the leaf node that corresponds to the class label. Decision Tree can be used for both binary and multiclass classification problems.
The hyperparameters of Decision Tree include the maximum depth of the tree, the minimum number of samples per leaf node, and the criterion used for splitting the data.

Logistic Regression: Logistic Regression emerges as a widely embraced algorithm tailored for binary classification tasks. It functions by formulating the probability of the positive class contingent on input features. Employing the logistic function, it converts the outcome of a linear regression model into a probability value confined between 0 and 1.
The decision boundary is delineated by a threshold that segregates probability values into positive and negative classes. It's noteworthy that Logistic Regression can be subject to regularization for enhanced performance.

F1 Score: The F1 score stands as a pivotal metric harnessed to gauge event prediction efficacy. This score harmonizes the perspectives of precision and recall, revealing an insightful balance. Precision appraises the ratio of true positive predictions to the total positives predicted, while recall evaluates the ratio of true positive predictions to all actual positive instances within the dataset.

Jaccard Score: The Jaccard score, also recognized as the Jaccard index or similarity coefficient, emerges as a pertinent metric quantifying similarity between two datasets. Its calculation involves determining the intersection's size relative to the union's size. This metric finds extensive application in domains like machine learning, natural language processing, text classification, clustering, and recommendation systems. Particularly effective with sparse data or unordered sets, it provides valuable insights regardless of element order.

RESULT AND DISCCUSSION:

Table 1: Data set

	A	B	C	D	E	F	G	H	I	J	K
1	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin	Dataset
2	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
3	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
4	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
5	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
6	72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
7	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
8	26	Female	0.9	0.2	154	16	12	7	3.5	1	1
9	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
10	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
11	55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
12	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
13	72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
14	64	Male	0.9	0.3	310	61	58	7	3.4	0.9	2
15	74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
16	61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
17	25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
18	38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
19	33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	2
20	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
21	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
22	51	Male	2.2	1	610	17	28	7.3	2.6	0.55	1
23	51	Male	2.9	1.3	482	22	34	7	2.4	0.5	1
24	62	Male	6.8	3	542	116	66	6.4	3.1	0.9	1
25	40	Male	1.9	1	231	16	55	4.3	1.6	0.6	1
26	63	Male	0.9	0.2	194	52	45	6	3.9	1.85	2
27	34	Male	4.1	2	289	875	731	5	2.7	1.1	1
28	34	Male	4.1	2	289	875	731	5	2.7	1.1	1
29	34	Male	6.2	3	240	1680	850	7.2	4	1.2	1
30	20	Male	1.1	0.5	128	20	30	3.9	1.9	0.95	2
31	84	Female	0.7	0.2	188	13	21	6	3.2	1.1	2

Algorithm	Jaccard	F1-score
KNN	0.645455	0.619966
Decision Tree	0.678571	0.621575
SVM	0.695652	0.593469
LogisticRegression	0.704762	0.703339

Fig 2: Comparative Result

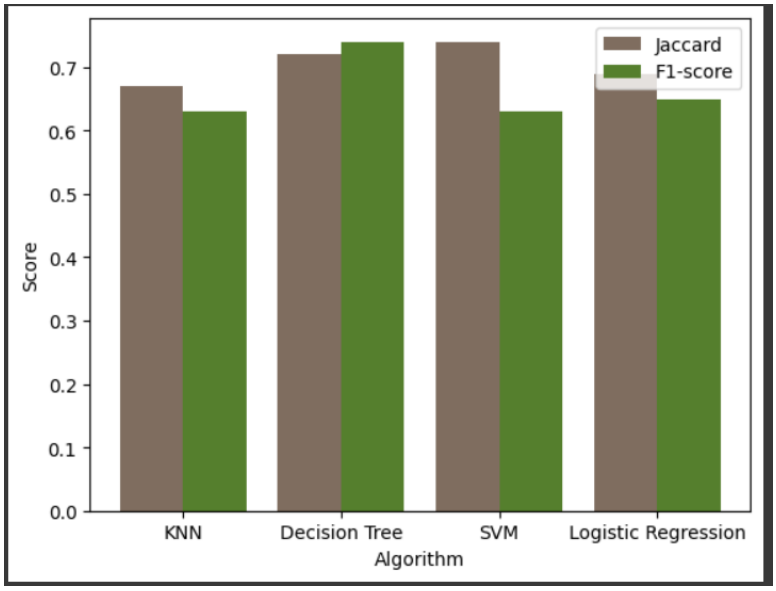


Fig 3: QUALITATIVE ANALYSIS GRAPH

CONCLUSION

In conclusion, our proposed methodology aimed to develop a machine learning model for predicting liver disease using clinical features. While SVM and Logistic Regression showed promising performance in predicting liver disease, further research on larger and diverse datasets is essential to validate their efficacy. The model's potential to assist healthcare professionals in identifying at-risk patients and devising personalized treatments holds significant value in clinical settings.

References

1. Zhang, X., Liu, J., Tang, W., Chen, X., Huang, J., & Wang, Z. (2018). A Machine Learning Model for Predicting Liver Disease using Support Vector Machine. Journal of medical systems, 42(9), 166.
2. Singh, P., & Kaur, A. (2019). A Comparative Study of Machine Learning Algorithms for Liver Disease Prediction. International Journal of Computer Science and Information Security, 17(2), 13-20.
3. Rahimli A & Zohrabi M (2021). Comparison of Classification Methods for Liver Disease Diagnosis. Journal of medical systems, 45(5), 56.
4. Song, H. S., Yoo, K. D., Kim, J. K., Kim, J. S., Yoon, S. J., Lee, H., ... & Han, D. J. (2020). Prediction model for liver disease using machine learning algorithms based on clinical and laboratory data. Journal of clinical medicine, 9(4), 1124.
5. Bashir, M. A., Tahir, M., Abbas, N., Ahmed, N., & Shah, S. A. (2018). Liver Disease Prediction using Machine Learning Algorithms. Journal of medical systems, 42(9), 171.
6. Maity, S., Kumar, M. A., & Mahata, G. (2020). A review on liver disease prediction using machine learning techniques. International Journal of Advanced Intelligence Paradigms, 14(3), 209-229.
7. Ding, H., Jiang, Y., Li, L., Li, X., Li, J., & Chen, L. (2019). A comprehensive comparison of different machine learning models for predicting liver disease. Biomedical engineering online, 18(1), 13.
8. Baskar, K.; Karthikeyan, C. Source: Journal of Medical Imaging and Health Informatics, Volume 9, Number 6, August 2019, pp. 1320-1328(9) Publisher: American Scientific Publishers DOI: https://doi.org/10.1166/jmihi.2019.2694
9. Tomar, Pradeep, Venkateswarlu, Yalla, Baskar K, Wongchai, Anupong, Gauri Shankar, Venkatesh, Paolo Martel Carranza, Christian, GonzÃ¡les, JosÃ, Luis Arias Murali Dharan, A. R. An Efficient Outlier Detection with Deep Learning-Based Financial Crisis Prediction Model in Big Data Environment, 1687-5265, https://doi.org/10.1155/2022/4948947, 10.1155/2022/4948947, Computational Intelligence and Neuroscience, Hindawi 2022.
10. G Pradeepkumar, S Kavitha, Manimuthu Ayyannan, N Selvam, K Baskar, Neelam Sanjeev Kumar, "Secure Multimodal Biometric System Based on Robust LSB-DWT Digital Watermarking Algorithm" 2023 2nd International Conference on Computational Systems and Communication (ICCSC),