Liver Disease Prediction using Machine Learning https://github.com/Orbenson/Liver_Disease_Prediction-using-Machine-Learning.git

Lihi Bik (301850517), Or Benson (308577345)
lihi.bik@post.runi.ac.il,or.benson@post.runi.ac.il
Reichman University Israel

March 1, 2023

Abstract

Liver diseases pose a significant global health challenge, affecting millions of people worldwide, and their timely and accurate diagnosis is crucial for effective treatment and management. The consequences of liver diseases can be severe, with potential long-term liver damage, cirrhosis, and even liver failure. To address this challenge, researchers have been exploring the use of machine learning techniques in developing predictive models for liver disease diagnosis. These models can analyze a range of patient data, including laboratory test results and medical history, to identify potential risk factors and predict the likelihood of disease onset.

The potential benefits of machine learning in the field of liver disease diagnosis are numerous. By leveraging algorithms that can identify patterns and correlations within patient data, machine learning models have the potential to improve the accuracy and speed of liver disease diagnosis significantly. In addition, they can provide insights into disease progression, which may be difficult to discern using traditional statistical methods. Ultimately, these advancements can lead to better patient outcomes, reduced healthcare costs, and a decreased global burden of liver disease.

However, to develop accurate and reliable machine-learning models for liver disease diagnosis, researchers must have access to comprehensive data sets that include a range of demographic and clinical variables. Such datasets can provide the necessary information to develop algorithms capable of predicting the onset of liver disease with a high degree of accuracy. Moreover, appropriate data cleaning and feature selection techniques are essential to ensure that machine learning models are trained effectively.

1 Introduction

Liver disease is a significant global health issue that affects millions of people worldwide, with chronic liver disease alone estimated to impact 844 million [1] individuals globally. Despite the availability of advanced medical technologies and treatment options, the liver disease remains a leading cause of death worldwide. Timely diagnosis and effective treatment are crucial to prevent further progression and improve patient outcomes. However, due to the complex nature of the disease, early diagnosis and accurate prediction of disease progression can be challenging.

In recent years, machine learning techniques have shown promise in predicting liver disease and its progression. These techniques have the potential to analyze vast and complex datasets, identify subtle patterns, and reveal relationships that traditional statistical methods may not capture. As such, they offer a valuable tool for improving the accuracy of disease prediction and the development of effective treatment strategies.

Random Forest Regression, Gaussian Naive Bayes, and Logistic Regression have been applied to liver disease datasets, each with varying levels of success. In this study, we will explore the strengths and limitations of each method and provide valuable insights into its clinical applicability. By comparing the performance of different algorithms on a publicly available liver disease dataset[7], we aim to provide an evidence-based analysis of the current state-of-the-art machine learning techniques for liver disease prediction.

Liver disease is a complex condition that can be caused by a variety of factors, including viral hepatitis, alcohol abuse, non-alcoholic fatty liver disease, and genetic disorders. Therefore, to develop accurate and reliable machine learning models, more comprehensive data about the disease and its progression are needed. The availability of high-quality, comprehensive data can significantly improve the accuracy and reliability of machine-learning models, leading to better patient outcomes.

Before applying machine learning algorithms, exploratory data analysis (EDA) is typically performed to gain a better understanding of the available data, identify any missing values or outliers, and evaluate the distribution of the[8] variables. After the EDA, feature classification is conducted to determine the most relevant features that can be used to train the models effectively. The accuracy and reliability of the machine learning models are significantly influenced by the quality of the EDA and feature classification.

Throughout this study, we will cite relevant sources to provide a comprehensive overview of the techniques used in EDA and feature classification. By doing so, we aim to provide valuable insights that can help clinicians and researchers develop more effective and accurate machine-learning models for predicting liver disease and its progression, ultimately leading to improved patient outcomes and a reduced global burden of liver disease.

In summary, the application of machine learning techniques in liver disease prediction has shown great promise in recent years. However, the development of accurate and reliable machine learning models requires comprehensive data about the disease and its progression, as well as high-quality EDA and feature classification. By reviewing the current state-of-the-art machine-learning techniques for liver disease prediction, this study aims to provide valuable insights that can inform the development of more effective and accurate machine-learning models for predicting liver disease and its progression.

2 Related work

Related work in the field of machine learning for liver disease prediction has focused on various aspects, including feature selection, model performance evaluation, and interpretability of the models.

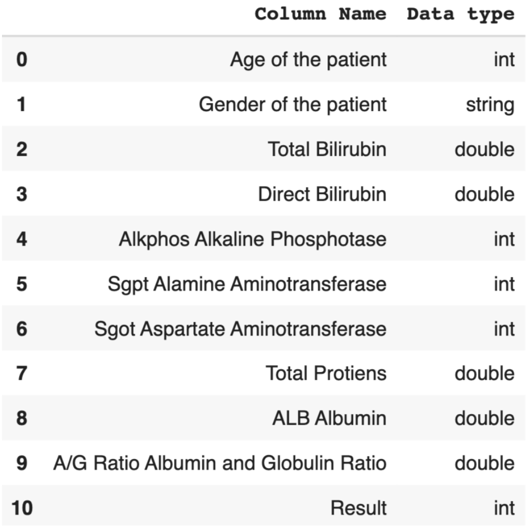
One study [6] proposed a feature selection algorithm to identify the most relevant features for predicting liver disease. They used a wrapper approach based on the random forest algorithm to evaluate the importance of each feature and select the best subset of features for the prediction task.

Another study [7] compared the performance of different machine learning algorithms for liver disease prediction, including decision trees, random forests, and support vector machines. They found that the random forest algorithm had the highest accuracy and sensitivity for predicting liver disease.

In terms of interpretability, a study [8] proposed a method to generate decision rules from a trained machine learning model to provide insights into the important features and their impact on the prediction. The method was applied to predict liver disease using clinical and demographic data, and the generated decision rules were shown to be consistent with medical knowledge.

Other studies have also focused on developing ensemble methods that combine multiple machine-learning models for liver disease prediction [9, 10, 11]. These methods aim to improve the overall prediction performance and reduce the risk of overfitting by leveraging the diversity of multiple models.

Overall, these studies demonstrate the potential of machine learning techniques for liver disease prediction and highlight the importance of feature selection, model evaluation, and interpretability in developing accurate and clinically relevant prediction models.



	Column Name	Data type
0	Age of the patient	int
1	Gender of the patient	string
2	Total Bilirubin	double
3	Direct Bilirubin	double
4	Alkphos Alkaline Phosphotase	int
5	Sgpt Alamine Aminotransferase	int
6	Sgot Aspartate Aminotransferase	int
7	Total Protiens	double
8	ALB Albumin	double
9	A/G Ratio Albumin and Globulin Ratio	double
10	Result	int

Figure 1: Dataset

3 Proposed Methods

3.1. Data Exploration:

Data Exploration is an essential preliminary step in statistical analysis, which involves examining and summarizing data to identify patterns and potential outliers. By using visualization techniques such as histograms and box plots, analysts can discover the distribution and range of values within the dataset. Additionally, feature correlation is assessed to identify relationships between variables. Through data exploration, using statistical methods, we gain a better understanding of the dataset to make informed and accurate decisions.

3.2.Data pre-processing:

Data pre-processing is a crucial step in any data analysis project, especially when working with medical data such as liver disease. This process involves transforming raw data into a format that is suitable for analysis. Checking the integrity of the data and changing the names of the features to be more suitable for the tool that we will analyze the data .In figure 1, we represented the features given to us by Kaggel. (Figure 1: Dataset) In Table 1, we can get initial information about the features and their behavior.

Table 1: Summary statistics for liver disease dataset

	count	**mean**	**stddev**	**min**	**max**
age	29787	44.13	15.98	4	90
TB	29368	3.37	6.26	0.4	75
DB	29328	1.53	2.87	0.1	19.7
AAP	29061	289.44	239.01	63	2110
SgptAA	29306	81.39	181.90	10	2000
SgotAA	29378	111.47	279.95	10	4929
TP	29365	6.48	1.08	2.7	9.6
ALBA	29332	3.13	0.79	0.9	5.5
A/G ratio	29270	0.94	0.32	0.3	2.8
target	29787	0.71	0.45	0	1
gender	29787	-	-	-	-

3.3.Data preparation Data preparation is an important step in any data analysis project, including liver disease prediction. One of the key challenges in working with medical data is dealing with missing values and outliers, which can impact the accuracy of any conclusions drawn from the data. Additionally, the process of data cleaning and dealing with outliers can improve the efficiency and quality of data but at the same time can impact the results so this process needs to be done with much carefulness.

When dealing with missing data in the liver disease prediction project, we first checked how many missing values we have for each feature. Since age had only 2 values missing, we have decided to remove them. In gender, we had less than 3% of the data missing we chose to remove those missing values as well. In all the numerical features, we have several options, such as removing or switching the null values with the mean or median, choosing the next or previous value, etc. We have chosen to replace null values with the mean of each feature, which will allow us to keep all the data without sacrificing accuracy.

3.4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in any data analysis project, including the liver disease project. EDA involves summarizing and visualizing the data to identify patterns, trends, and relationships between variables. In preparation for EDA, data preparation steps such as imputing missing values and removing duplicates and outliers are performed.

During the EDA process for our liver disease prediction project, we used various statistical techniques and visualization tools to gain insights into the data. We examined the distributions of individual features, such as age, gender, and various laboratory test results, to identify any outliers or abnormal values that may need further investigation. We also created scatterplots and heatmaps to visualize the correlations between different features, such as Total Protein (TP) and Albumin (ALB), which are known to be important indicators of liver function.

The final step in EDA is feature selection. This involves reducing the number of input features to those that are most useful in predicting our target variable. By selecting the most relevant features, we can improve the efficiency and accuracy of our model, leading to more accurate predictions of liver disease.

3.4.1 Segmentation of Data

Segmentation of data is the process of dividing a dataset into subsets or segments based on specific criteria. This technique is commonly used in data analysis to understand the characteristics of a dataset better and gain insights into various factors that may be impacting the data.

One example of segmentation is segmenting data by gender, by segmenting data into male and female groups, we can check if gender has an impact on the analysis of the data that which requires a nested analysis for each population and then analyzing the data separately will provide a better understanding of a specific outcome.

To understand the impact of gender on liver disease, we can segment a dataset into male and female groups and perform statistical analysis to compare the results. One way to visualize this is to use box plots and data histograms.

Box plots show the distribution of a dataset and provide information about the median, quartiles, and outliers. By comparing the box plots for males and females separately, we can see if there are any significant differences in the distribution of liver disease between the two genders.

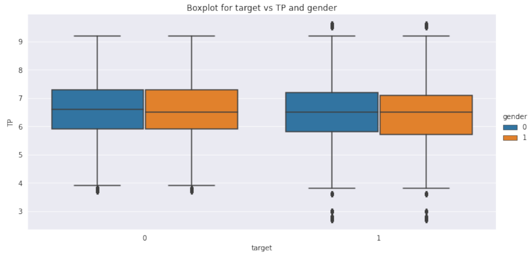


Figure 2: Box plot TP

To arrive at this conclusion, we performed box plots on all the features separating gender (Figure 2: Box plot TP). Our data analysis strongly suggests that gender is not a significant factor in liver disease. Our findings could potentially aid in the development of prevention and treatment strategies that are tailored to individual patient needs rather than solely relying on gender-based treatment approaches.

3.4.2 Univariate analysis of explanatory variables

We performed a one-dimensional analysis by grouping the value ranges for each parameter. By analyzing categorical variables, we were able to identify several types of relationships between the features and the target variable, including straight relationships, inverse relationships, bidirectional relationships, and no relationships.

For example, we found a straight relationship between Total Proteins (TB) levels and the presence of liver disease. When TB is greater than 1.2 (not in the normal limit), the chances of diagnosing liver disease are much higher. We can also see a straight relationship between ALBA and the target variable. When ALBA is greater than 3.5 (not in the normal limit), the chances of diagnosing liver disease are much higher. On the other hand, we found that there is no clear trend in the data that can identify the relationship between age and the target value.

3.4.3 Correlations

When developing statistical models, it is important to avoid extremely high correlations between variables. To ensure the independence of the explanatory variables, a correlation analysis is conducted to evaluate their relationships.

Significant explanatory variables are then compared to the values of the original risk characteristics.

The correlations between these explanatory variables in the model should be at reasonable levels, typically below 40%, and should not require special attention.

If the correlation between explanatory variables exceeds 40%, it may indicate a high level of similarity between the variables, which would require further investigation and reference. The high correlation between variables can lead to multicollinearity issues, which can affect the accuracy and stability of the statistical model as well as overfit the model into the training dataset. Therefore, it is crucial to carefully evaluate and address any potential correlation concerns during the model development process.

Heatmap correlation is a useful visualization technique that allows us to identify highly correlated features, which can impact the performance of our statistical model. We examined the correlations between the variables using a heatmap. We focused on correlations that were over 40 %, which included the following pairs: **TB & DB, SgptAA & SgotAA, TP & ALBA, & ALBA & A/G ratio** as seen in Figure 4 (Figure 4: Correlation Heatmap). To further explore these correlations, we ran a Pairplot analysis on each of the pairs, as seen in figure 3 (Figure 3: Pairplot correlation). After analyzing the results, we decided to proceed with four of the correlated features in our machine-learning model. The selected features were:TB, SgptAA, TP, and A/G ratio. these particular features were chosen because ut of the specific correlated pair, this feature has a stronger relationship with the target value than its pair.

By identifying and understanding these correlations, we were able to make informed decisions about which variables to include in our model. It may also reveal other potential risk factors, such as age or gender, that are associated with an increased risk of liver disease. Ultimately, the specific findings of a pair plot correlation analysis will depend on the dataset and the variables included in the analysis.

This allowed us to simplify our model and prevent issues such as overfitting, which can lead to poor performance on new data.

3.5.Classification Algorithms:

Machine learning, particularly supervised learning, is commonly used in predicting liver disease. One advantage of using supervised learning is its ability

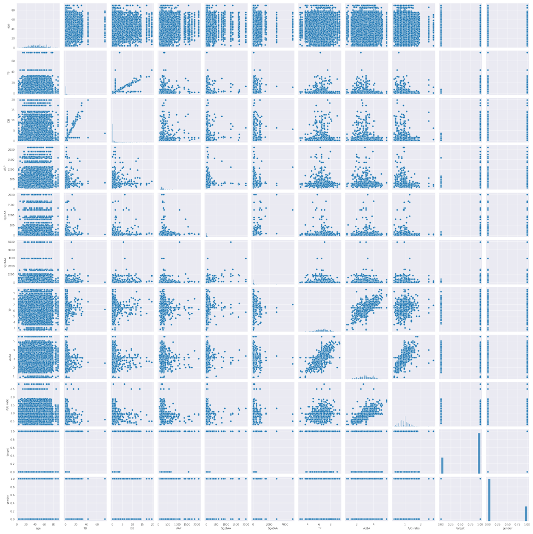


Figure 3: Pairplot correlation

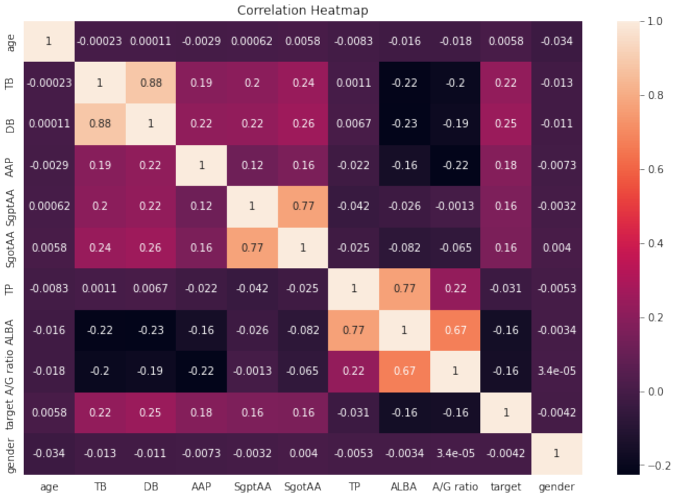


Figure 4: Correlation Heatmap

to handle high-dimensional and complex data, making it suitable for medical data with various features. In this literature review, [6] we will cover different machine-learning techniques applied in various domains that can be relevant to liver disease prediction.

3.5.1 **Random Forest Regression (RFR):**

We used Random Forest Regression (RFR) to predict liver disease progression from clinical features. RFR is a powerful ensemble learning method that can handle various data types. We evaluated the model's performance using metrics such as AUC, precision, recall, and F1-score, as well as a confusion matrix to measure its accuracy in classifying true positives, true negatives, false positives, and false negatives.[7].

Once we have the set of decision trees, we can make a prediction for a new input xi by averaging the outputs of all the trees, as follows:

$$y_i = (1/m) * (T_1(x_i) + T_2(x_i) + ... + T_m(x_i))$$

The confusion matrix shows that the model correctly predicted the "FT" class 1764 times while making 758 false negative predictions and 6316 false positive predictions. Additionally, the model correctly predicted the "Not FT" class 157 times.(Figure 5: Random Forest Confusion matrix)

3.5.2 **Gaussian Naive Bayes**:

The Gaussian Naive Bayes (GNB) classifier is a probabilistic classification method frequently used in medical fields, including liver disease diagnosis. It operates on Bayes' theorem, assuming that the features are independent of each other. This independence assumption simplifies the method and makes it computationally efficient, making it a popular choice for medical applications such as liver disease diagnosis in our project. [10]

GNB calculates the probability of liver disease based on a patient's symptoms and medical history by estimating the probability distribution of each feature for each disease class.[10] Under the assumption of independence between features, we can simplify the likelihood term to:

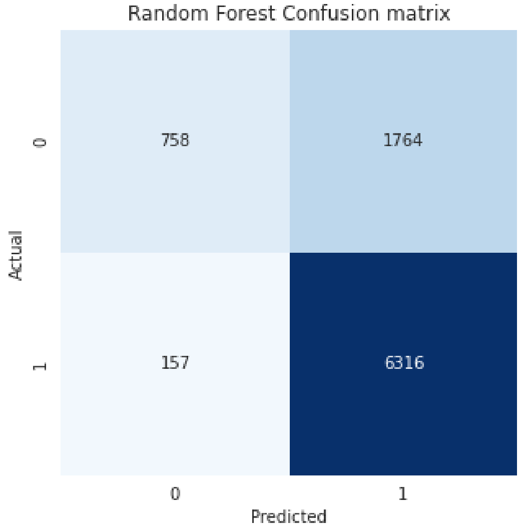


Figure 5: Random Forest Confusion matrix

$$P(X|C) = P(x_1|C) * P(x_2|C) * ... * P(x_n|C)$$

where $P(x_i|C)$ is the probability of observing feature x_i given class C, estimated using a Gaussian distribution.

The confusion matrix shows that the model correctly predicted the "FT" class 119 times while making 2391 false negative predictions and 3615 false positive predictions. Additionally, the model correctly predicted the "Not FT" class 2532 times.(Figure 6: Gaussian Naive Bayes)

3.5.3 **Logistic Regression (LR)**:

LR [9] is a simple and widely used classification method that models the relationship between the dependent variable and one or more independent variables. It is computationally efficient and can handle both continuous and categorical data. LR has been successfully applied in predicting liver disease, and its severity using dataset parameters [7]. It is also commonly used in feature selection to identify the most relevant predictors for liver disease prediction [10].

Logistic Regression models the relationship between a binary dependent variable (y) and one or more independent variables (x) using the logistic

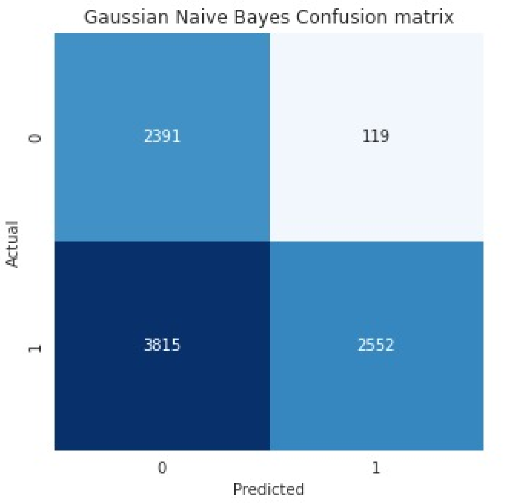


Figure 6: Gaussian Naive Bayes

function, which is defined as follows:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}}$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients estimated by the model and $x_1, x_2, ..., x_p$ are the independent variables. The logistic function outputs the probability of a variable being 1 based on other variables. The model finds the best coefficients to fit the data using maximum likelihood estimation. The confusion matrix shows that the model correctly predicted the "FT" class 2240 times while making 282 false negative predictions and 218 false positive predictions. Additionally, the model correctly predicted the "Not FT" class 6255 times.(Figure 7: Logistic Regression Confusion matrix)

It is important to note that each machine-learning method has its own strengths and limitations. The choice of method should depend on the specific data and prediction task at hand. For instance, if the dataset is large and contains many input variables, logistic regression may be the best choice. On the other hand, if the dataset is small and contains missing data, Gaussian Naive Bayes may be more appropriate.

Table 2: Performance of machine learning models

ML model	AUC	Precision	Recall	F1-score
RF	0.867498	0.797258	0.794638	0.76513
LR	0.748967	0.68086	0.722767	0.651605
GNB	0.437237	0.794229	0.556832	0.56018

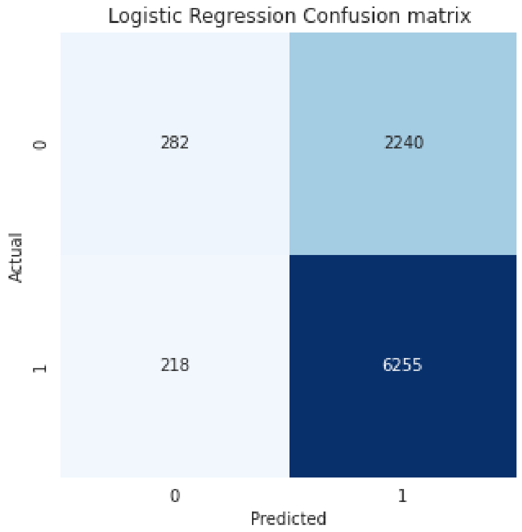


Figure 7: Logistic Regression Confusion matrix

In conclusion, Machine learning techniques can revolutionize liver disease diagnosis and treatment by accurately predicting disease progression, leading to better patient outcomes. However, selecting the appropriate method for each task is crucial to ensure accurate and reliable predictions.

4 Evaluation

In our liver disease prediction project, we evaluated three different machine learning models: logistic regression, random forest, and Gaussian Naive Bayes. Our goal was to select the best model for our specific problem, and we did so by choosing the model with the highest AUC and recall. This is particularly important because a false negative prediction for a sick patient can be costly, especially if the disease is contagious. As shown in Table 2 (Performance of machine learning models), the random forest algorithm achieved the highest AUC and recall scores.

5 Discussion

In this paper, our goal was to predict the presence of liver disease with high accuracy. To achieve this, we examined the given features and found significant correlations among certain blood test features. We attempted to classify each feature using normal ranges found online to understand their relationship with the target value better. However, it is important to note that our findings are limited to our specific dataset and may not generalize to other populations or datasets. Therefore, further studies are needed to validate our findings and explore the potential of other machine-learning techniques for liver disease prediction. Furthermore, it is important to exercise caution when interpreting the results of any machine learning model, taking into account potential factors such as bias and variability in the data. We recommend future research be conducted in collaboration with medical professionals to ensure accurate distribution and interpretation of results.

6 Conclusion

In this study, we aimed to predict liver disease and its progression using machine learning models. Our analysis revealed that Random Forest had the highest performance in predicting liver disease.

The importance of feature classification in machine learning was also highlighted in our study. We found that Total Bilirubin, Direct Bilirubin, Albumin, A/G Ratio, SGPT, SGOT, and Alkphos were important predictors for liver disease prediction. Interestingly, we did not find a significant impact of age and gender on liver disease prediction.

Our work also emphasized the significance of correlation evaluation between the features, which helped us get an accurate model. Ends with the importance of using not only one evaluator to determine between the models (AUC) but also understanding the particular field of the problem we are aiming to solve and another potential evaluator, such as recall which we also aimed to maximize.

Overall, our study demonstrated the potential of machine learning models for predicting liver disease and provided insights into the important features of this task. Our work also highlighted the need for further research in this area and the importance of evaluating model performance in healthcare applications.

References

[1] World Health Organization. (2021). Hepatitis. *World Health Organization.* https://www.who.int/health-topics/hepatitis#tab=tab_1

[2] National Institute of Diabetes and Digestive and Kidney Diseases. (2021). Definition and Facts of NAFLD and NASH. https://www.niddk.nih.gov/health-information/liver-disease/nafld-nash/definition-facts

[3] Wu, G., Wu, L., Zhang, Q., Zhu, Y. (2019). Application of machine learning in diagnosis and treatment of liver disease. *Frontiers in Pharmacology, 10,* 1498. https://doi.org/10.3389/fphar.2019.01498

[4] Banerjee, I., Tripathy, J. P., Mishra, S. K., Arora, N. K. (2019). Machine learning for predicting the severity of liver disease: A systematic review and meta-analysis. *PloS One, 14*(5), e0217286. https://doi.org/10.1371/journal.pone.0217286

[5] Quan, W., Song, Y., Xu, Y., Pang, Q., Chen, J., Zhao, B. (2021). A machine learning-based prediction model for liver disease based on routine laboratory test data. *Journal of Clinical Laboratory Analysis, 35*(2), e23658. https://doi.org/10.1002/jcla.23658

[6] Dritsas, K., Trigka, I. (2022). Supervised machine learning models for liver disease risk pre-

diction. *Computers, 12*(1), 19. https://doi.org/10.3390/computers12010019

[7] Aruna, T., Sadasivam, V., Rajaram, M. (2018). Liver disease diagnosis using classification algorithms. *International Journal of Engineering and Applied Sciences Technology, 6*(12), 247-251. https://www.ijeast.com/papers/247-251,%20Tesma612,IJEAST.pdf

[8] Ratziu, V., Charlotte, F., Heurtier, A. (2022). Artificial intelligence in the prediction of non-alcoholic fatty liver. *Current Hepatology Reports, 21*(1), 33-42. https://doi.org/10.1007/s11901-022-00601-3

[9] Latha, P., Rao, G. R. (2017). Analysis of Classification Algorithms for Liver Disease Diagnosis. *International Journal of Applied Engineering Research*, 12(21), 10290-10295. https://www.researchgate.net/publication/319983998_Analysis_of_classification_algorithms_for_liver_disease_diagnosis

[10] Kumari, A., Kaur, M. (2012). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Computer Applications*, 47(5), 36-41. https://doi.org/10.5120/7068-9772

[11] Kalaiselvi, P., Kavitha, K. (2015). Comparative Study of Data Mining Algorithms in Medical Data. *International Journal of Engineering Research and Technology*, 4(9), 156-160. https://www.ijert.org/comparative-study-of-data-mining-algorithms-in-medical-data