Liver disease detection using machine learning techniques

Deepika Bhupathi[1], Christine Nya-Ling Tan[1], Sreenivas Sremath Tirumula[1] and Sayan Kumar Ray[1]
[1] Manukau Institute of Technology
bhup20@manukaumail.com, christine.tan@manukau.ac.nz, sreenivas.tirumala@manukau.ac.nz, sayan.ray@manukau.ac.nz

Abstract

Around a million deaths occur due to liver diseases globally. There are several traditional methods to diagnose liver diseases, but they are expensive. Early prediction of liver disease would benefit all individuals prone to liver diseases by providing early treatment. As technology is growing in health care, machine learning significantly affects health care for predicting conditions at early stages. This study finds how accurate machine learning is in predicting liver disease. This present study introduces the liver disease prediction (LDP) method in predicting liver disease that can be utilised by health professionals, stakeholders, students and researchers. Five algorithms, namely Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (K-NN), Linear Discriminant Analysis (LDA), and Classification and Regression Trees (CART), are selected. The accuracy is compared to uncover the best classification method for predicting liver disease using R and Python. From the results, K-NN obtains the best accuracy with 91.7%, and the autoencoder network achieved 92.1% accuracy, which is above the acceptable level of accuracy and can be considered for liver disease prediction.

Keywords: Liver disease, Machine learning, Prediction, Data analytics, Healthcare, Autoencoders

1 Introduction

The liver is one of the most critical organs of the human body. It plays an essential role in the body's function. Primary purposes include removing toxins from the body, fighting against infections, and balancing the hormones and secretion of bile juice (Devikanniga et al., 2020). If these functions are not performed by the liver correctly, it will result in several complications and liver diseases. Therefore if a virus infects the liver or chemicals that injure the liver are consumed, or the immune system's

dysfunction occurs, severe damage to the liver or malfunctioning may happen, which ultimately might cause death (Nahar & Ara, 2018).

Liver disease is one of the most chronic and threatening diseases globally that can cause various side effects if not treated early (Dutta et al., 2022). According to World Health Organization (WHO) report in 2018, the number of deaths due to liver diseases is around one million and ranked 11th in the world with a critical number of fatalities (World Total Deaths, n.d.). As the symptoms of liver diseases cannot be visible until the condition becomes chronic, it is challenging and daunting for medical health professionals to identify liver disease at its early stages (Devikanniga et al., 2020). In addition, the traditional testing methods like sonography, MRI scans and CT scans that are available for detecting liver diseases are expensive and harmful with numerous side effects (Joloudari et al., 2019). Thus, a significant constraint found by health care workers is to predict liver diseases at an early stage, at minimal cost and at the same time provide a better health care system to treat liver diseases. Severe liver diseases include problems with indigestion, dry mouth, pain in the abdomen, skin colour turning yellow, numbness, memory loss and fainting problems (Shaheamlung et al., 2020). Unnoticed at the initial stages, these symptoms are only visible when the disease turns chronic. However, even though the liver is partially infected, it can still function (Devikanniga et al., 2020).

Diagnosis of liver diseases can be divided into three stages i.e., the first stage is liver inflammation, the second is liver scarring (cirrhosis), and the final stage is liver cancer or failure. Since these scenarios are present in liver disease, early prediction is significant to provide better health for New Zealanders. If liver disease is diagnosed early, there will be a chance of early treatment and control of deaths due to liver diseases (Arbain & Balakrishnan, 2019). But when the liver fails to function, few treatments are available except liver transplantation (Shaheamlung et al., 2020), which is very expensive, particularly in New Zealand (Hepatitis C, 2021). Apparently, in New Zealand, 35 - 40% of the population are not diagnosed with Hepatitis C at the early stages because of the asymptomatic behaviour of liver disease. Unfortunately, most of these individuals do not know the risks linked to liver disease. Due to the asymptomatic behaviour and higher costs of liver disease treatment, it is essential to prevent or diagnose early for better treatment.

With advancements in biomedical sciences, the health care system has significantly improved by predicting disease using machine learning techniques (El-Shafeiy et al., 2018). Machine Learning algorithms are one of the potential solutions to this problem due to their handling large amounts of data and employing different approaches like classification, association and clustering, which benefits in realistic arbitration of disease prediction (Naseem et al., 2020).

There are different learning techniques in ML methods, one of which is supervised learning. Supervised learning techniques use labelled data and map the input and output data. These supervised learning methods are widely used for prediction and classification (Osisanwo et al., 2017). Supervised learning techniques would be appropriate as this research predicts whether the patient has liver disease or has no liver disease. The supervised learning methods used in this study are Support Vector Machine (SVM) (Boser et al., 1992), Naïve Bayes (McCallum & Nigam, 1998), K-Nearest Neighbors (K-NN) (Fix & Hodges, 1951), Classification and Regression Trees (CART) (Breiman et al., 1984), and Linear Discriminant Analysis (LDA) (Kemp, 2003). The main objective of this research is to compare the accuracies using five supervised learning algorithms, i.e., SVM, Naïve Bayes, K-NN, CART, LDA and Autoencoders, for predicting whether the patient has liver disease or not. This study also proposes the liver disease prediction (LDP) method to help relevant stakeholders pursue an effective healthcare strategy.

Moreover, this paper examines the techniques that indicate liver diseases at an acceptable level of accuracy and determines the methods that produce the best accuracy. This study selects a single data set of liver patients with five supervised learning techniques that are applied to that data set in R. The accuracy results from other learning techniques are also used to compare the best algorithm for predicting liver diseases. The stakeholders, including doctors, researchers, lab technicians, or

companies dealing with healthcare improvements, can use these results to predict liver diseases at a lower cost and provide better health care in liver treatment.

2 Literature Review

In a study conducted by Vijayarani and Dhayanand (2015), the liver disease prediction applied the SVM and Naïve Bayes (using MATLAB 2013 software) on the Indian Liver Patient Records dataset having 583 instances and 11 attributes, with accuracies of 79.66% (SVM) and 61.28% (Naïve Bayes). In their findings, the time taken to execute SVM was 3210ms, almost two times the time taken by Naïve Bayes (i.e., 1670ms), without preprocessing missing values. In addition to the accuracies, they found that SVM had better performance than Naïve Bayes.

Auxilia (2018) made an accurate prediction for liver disease using different ML methods, including SVM, Random Forest, Decision Trees, Artificial Intelligence and Naïve Bayes. The research was conducted using R on the Indian Liver Patient Records dataset, with 583 instances and 11 attributes. The accuracies were obtained from SVM (77%), Random Forest (77%), Decision Trees (81%), Artificial Intelligence (71%), and Naïve Bayes (37%), with the highest accuracy from the Decision Trees algorithm, and least with Naïve Bayes.

Wu et al. (2019) did a prediction analysis on patients having Fatty Liver Disease (FLD). The research collected 700 patient records from New Taipei Hospital, which had screening tests for fatty liver disease; out of 700 patients, 577 records were considered depending on the patient's age and sufficient data. Of those 577 patients, 377 had fatty liver disease, and the remaining had no fatty liver disease. The dataset contains patient health details of age, gender, systolic and diastolic blood pressure, abdominal girth, glucose level, triglyceride, HDL-C, SGOT-AST, and SGPT-ALT. Synthetic Minority Over-Sampling Technique (SMOTE) was applied at the data preprocessing stage, and normalisation was done. Four ML algorithms, namely Random Forest, Naïve Bayes, Artificial Neural Network and Logistic Regression with 3, 5, and 10-fold cross-validation, were applied in the next step. In addition to the accuracies, the area under the receiver operating curve for all the algorithms was observed. Random Forest had given the best accuracy with all the cross-validations from all the results.

Singh et al. (2020) focused their research on predicting liver disease using different classification methods with feature selection and implementing software for easy prediction. The study was conducted on the Indian Liver Patient Records dataset. Some attributes were removed during the feature selection phase using the Correlation-based Feature Selection Subset Evaluator with the Greedy Stepwise search method in WEKA. Only five (5) attributes were selected through this method: Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, and Aspartate Aminotransferase. With this, six different classification methods were applied: Logistic Regression, Naïve Bayes, Sequential Minimal Optimization (SMO), Random Forest, Instant based Classification (IBk), and Logistic Regression has provided the highest accuracy with 74.36%. The least accuracy was produced by Naïve Bayes (55.9%).

Most of the past research concentrated on just the analysis but not the preprocessing part for this Indian Liver Patient Records dataset. So, this research bridges the gap by considering preprocessing as a significant stage in data analysis. Moreover, several other algorithms are also applied in this research.

3 Research Methodology

The proposed liver disease prediction (LDP) method used in this research is based on SEMMA (Santos & Azevedo, 2005), which stands for Sample, Explore, Modify, Model, and Assess (Azevedo & Santos, 2008).

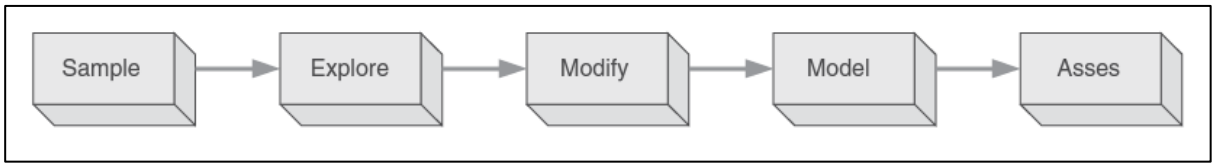

Figure 1: SEMMA lifecycle (Mariscal et al., 2010)

SEMMA lifecycle (see Figure 1) is a simple process to understand, aiming to get the solutions quickly for data mining problems and determine business goals. This methodology has developed by an institute named SAS Institute.

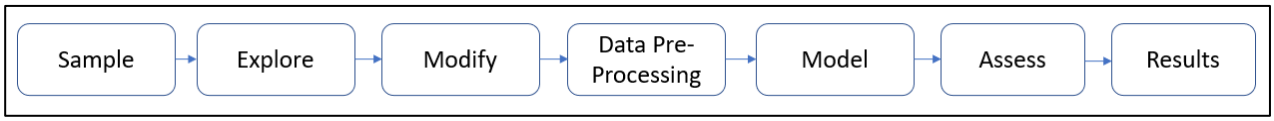

Figure 2. The liver disease prediction (LDP) method
Source: Developed for this study

The LDP method involved in this research are Sample, Explore, Modify, Data preprocessing, Model, Assess and Results. Along with these steps from the SEMMA lifecycle, two more steps, Data preprocessing and Results, are added to this research process. These steps (see Figure 2) include:

3.1 Sample

The first stage in the LDP method proposed in this study is 'Sample'. After deciding on the topic for the study, the first step is data collection. It is referred to as data collection and considering the part of data useful for the study (Azevedo & Santos, 2008). So, the data sets related to liver diseases are searched on different platforms named UCI repository and Kaggle. The suitable dataset is found from the platform Kaggle, a binary classification dataset that determines whether the patient has liver disease. After observing the credibility of the dataset, this dataset named 'Indian Liver Patient Records' is selected.

3.2 Explore

The second stage is 'Explore'. Exploring the data stage involves data understanding. This exploration stage also comprises finding the surprising trends and patterns present in the data to generate new ideas (Azevedo & Santos, 2008). In this study, exploring the data is at two stages. One is the data exploration on the background of liver disease. The other stage is exploring the dataset, which shows the details regarding the attributes present and how these attributes are correlated with each other and how these input attributes are correlated with the output attribute are studied. In addition, missing values are also identified. This analysis is performed using R.

3.3 Modify

The third stage is 'Modify'. Modify refers to data transformation (Azevedo & Santos, 2008). In this study, the attributes in the dataset are not in the same format, and the attribute's data type restricts the analysis to be done on the attribute. So, some of the features having the data type integers are converted into numerical, which makes all the attributes have the same numerical data type and makes the analysis be done efficiently.

3.4 Data Preprocessing

The fourth stage is 'Data preprocessing'. This data preprocessing refers to cleaning and preparing the data for modelling (Azevedo & Santos, 2008). This data preprocessing involves replacing the missing values and balancing the dataset as the class distribution of the dataset is imbalanced. This balancing is done using the Random Over Sampling Example (ROSE) (Menardi & Torelli, 2014).

3.5 Model

The fifth stage is the 'Model'. The modelling stage means applying the selected techniques or the algorithms to the data (Azevedo & Santos, 2008). The five algorithms, SVM, Naïve Bayes, LDA, CART and K-NN, are applied.

3.6 Assess

Assess stage, which is the sixth stage, involves assessing the data by deciding whether the data produced from modelling techniques are reliable and accurate. This stage also evaluates how well the algorithms performed on the data (Azevedo & Santos, 2008).

3.7 Results

The seventh stage of the proposed LDP method is 'Results'. The results stage involves presenting the results after assessing the data. All the results of accuracies and confusion matrix metrics will be described.

4 Performance Analysis

4.1 Descriptive Analysis

The dataset selected for this study is the liver disease dataset. This dataset, named 'Indian Liver Patient Records', is obtained from Kaggle. The data from the dataset is collected from the North-East part of Andhra Pradesh, India (*Indian Liver Patient Records*, n.d.). It is a binary classification dataset predicting whether the patient has liver disease or not. As stated in Table 1, the dataset contains 583 instances and 11 attributes. Of those 11 attributes, one of the attributes is class which denotes whether the patient has liver disease.

Dataset Name	# Of Instances	#Of Attributes	#Of Class
Indian Liver Patient Records	583	11	2

Table 1: *General details of the dataset*

Of these 583 patient records, 416 have liver disease, and 167 have no liver disease. The metadata of the dataset is indicated in Table 2. Figure 3 shows the binary classification dataset, having class two values of '1' and '2', where '1' denotes that the patients have liver disease and '2' denotes those patients do not.

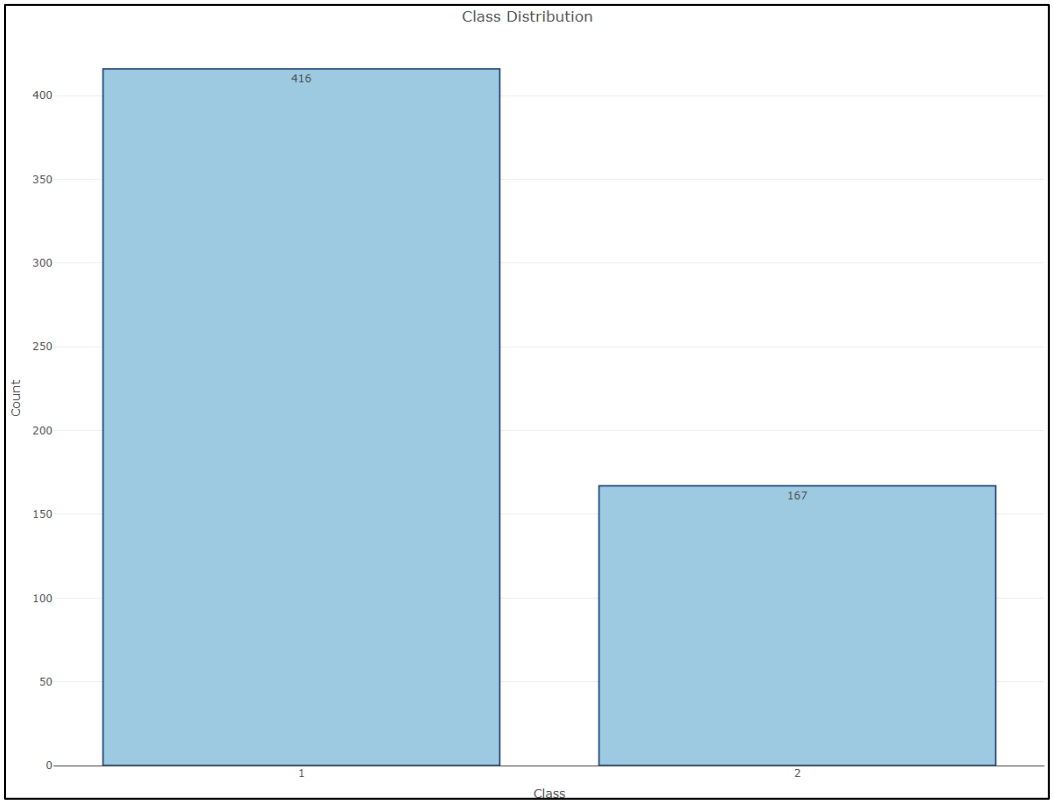

Figure 3: Class distribution

Number	Feature	Definition	Type	Explanation
1	Age	Age of the patient	Integer	Examination Results
2	Gender	Sex of the patient	Nominal	Male, Female
3	Total Bilirubin	Total Bilirubin in mg/dL	Numeric	Examination Results
4	Direct Bilirubin	Conjugated Bilirubin in mg/dL	Numeric	Examination Results
5	Alkaline Phosphotase	Alkaline Phosphotase in IU/L	Integer	Examination Results
6	Alamine Aminotransferase	Alamine Aminotransferase in IU/L	Integer	Examination Results
7	Aspartate Aminotransferase	Aspartate Aminotransferase in IU/L	Integer	Examination Results
8	Total Proteins	Total Proteins g/dL	Numeric	Examination Results
9	Albumin	Albumin in g/dL	Numeric	Examination Results
10	Albumin and Globulin Ratio	Albumin & Globulin Ratio	Numeric	Examination Results
11	Dataset	A patient has liver disease or not	Nominal	1-Has Liver Disease 2-No Liver Disease

Table 2: *Metadata of the dataset*

4.2 Data Modification

In the dataset, the data types of all the attributes are not the same; therefore, to maintain consistency and better analysis, the attributes having integer data types are converted into numerical ones. Four attributes have the integer data type: Age, Alkaline Phosphatase, Alamine Aminotransferase, and Aspartate Aminotransferase (see Table 3). These are converted to numerical.

Attribute	Definition	Data type	Converted data type
Age	Age of the patient	Integer	Numerical
Alkaline Phosphatase	Alkaline Phosphatase in IU/L	Integer	Numerical
Alamine Aminotransferase	Alamine Aminotransferase in IU/L	Integer	Numerical
Aspartate Aminotransferase	Aspartate Aminotransferase in IU/L	Integer	Numerical

Table 3: *Details of the attributes after the data modification*

4.3 Data Preprocessing

Data preprocessing is an important segment of data analysis. This study requires data processing for missing values and balancing the dataset.

4.3.1. Replacing Missing Values

This step involves replacing the missing values. If more missing values exist in the dataset, the instances or attributes corresponding to higher missing can be removed. In this dataset, only four missing values can be replaced with different values like mean and median (Hossain et al., 2021). This study takes the mean value to replace the missing values. The four missing values are in the attribute Albumin and Globulin ratio. These are replaced by taking

the mean value of the attribute. Figure 4 illustrates that the albumin and globulin ratio attribute has the missing value at 268, 328, 343 and 373 instances.

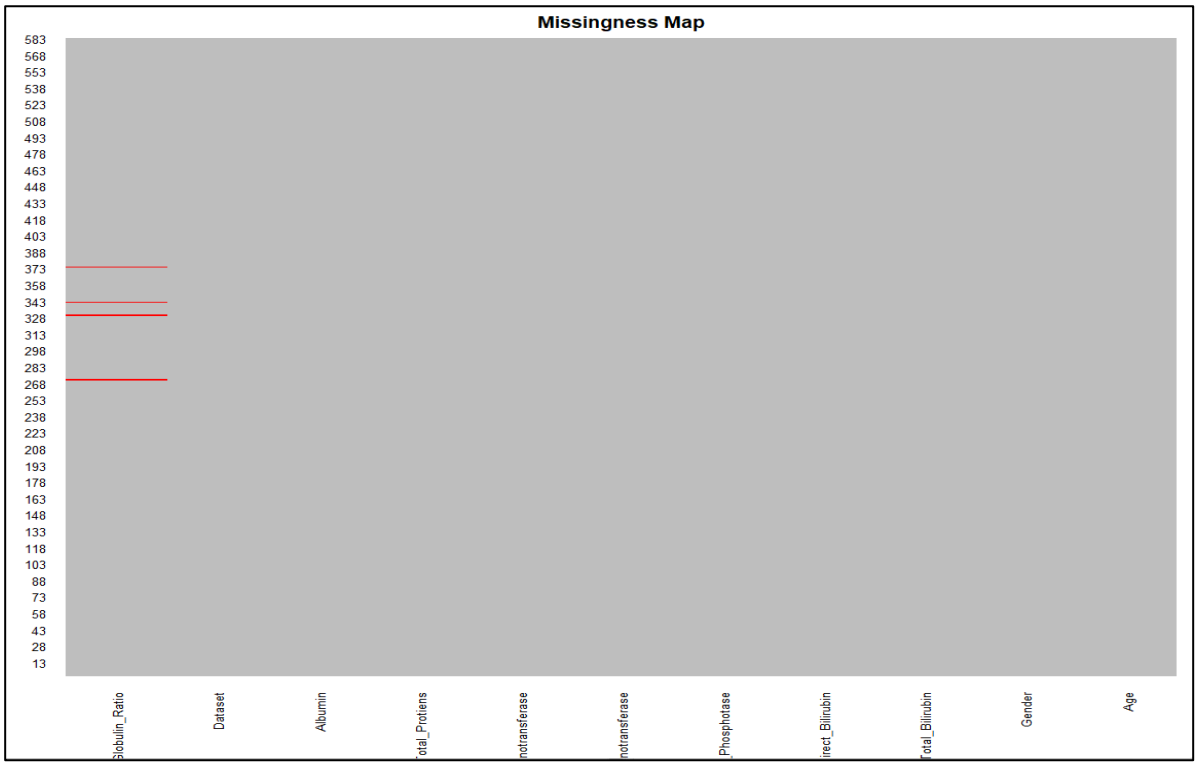

Figure 4: Missing value plot

4.3.2 Balancing the Dataset

The class distribution of the dataset is imbalanced, with 416 having liver disease and 167 without liver disease. This class distribution is imbalanced and balanced by applying ROSE using R. ROSE is a bootstrap method that produces balanced synthetic samples to balance the data (Lunardon et al., 2014). The reason for choosing ROSE is that the dataset is small, and the most reliable information might be lost if undersampling is conducted. The other reason for considering ROSE is that it generates samples similar to the rare class samples, which is also a consideration for an effective method for getting reliable accuracy from the balanced dataset as this study's main aim is the accuracy metrics of the algorithms (Lunardon et al., 2014). After applying the ROSE method on the class attribute, the sample generated is 520 having liver disease denoted by '1' and 480 without liver disease represented by '2' in Figure 5.

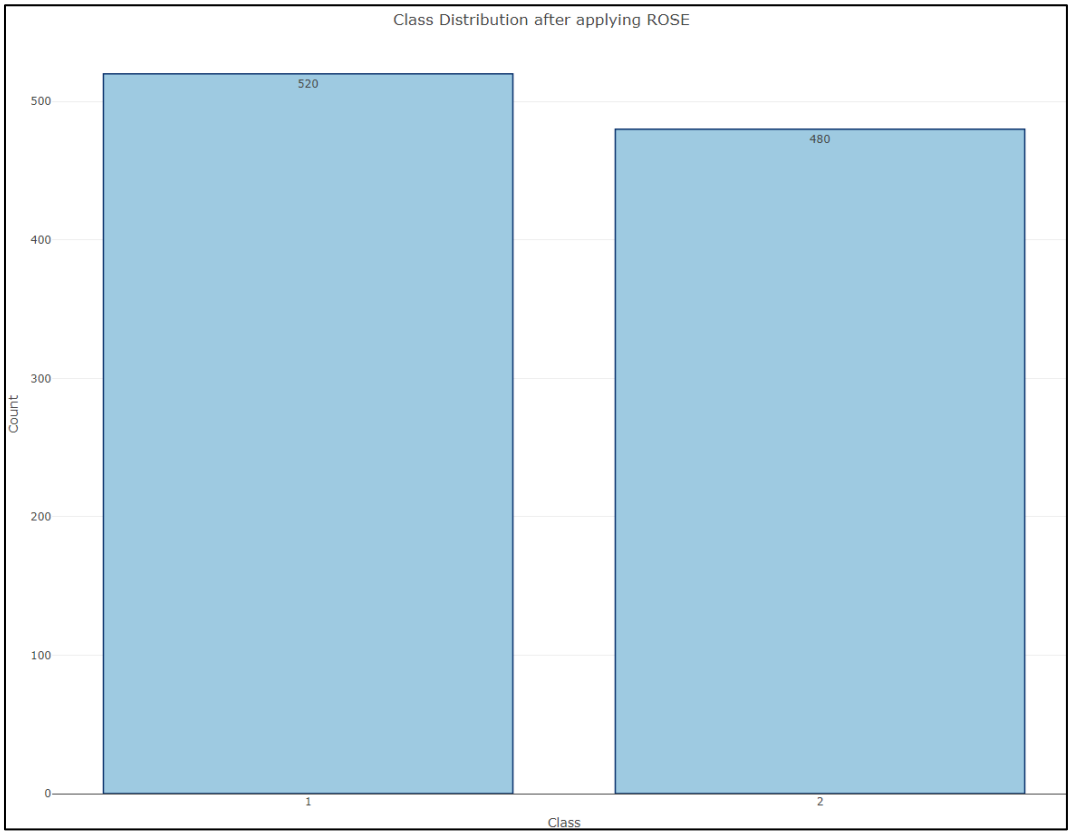

Figure 5: Class distribution after applying ROSE

5 Experimental Environment

5.1 Experimental settings and parameter settings

The data analysis of applying algorithms and finding the accuracy is done using R with version 1.4.1717. The investigation starts by loading the dataset into R and then modifying and preprocessing the data. Then five different algorithms, namely Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbours (K-NN), and Classification and Regression Trees (CART), are applied to the dataset. For all the algorithms, the seed is set to 7 and a cross-fold validation of 10. For K-NN, the k value is set to 3.

5.2 Classifiers

5.2.1. Support Vector Machine (SVM)

SVM is a supervised machine learning technique that strives to search for a hyperplane with maximum margin. Then it separates the linearly independent variables onto either side of the hyperplane and classifies the data (Devikanniga et al., 2020).

5.2.2. Linear Discriminant Analysis (LDA)

LDA reduces the dimensions at the preprocessing stage and classifies the data. LDA organises the data by mutating the attributes to lower-dimensional space, magnifying the within-class and between-class variance ratios and providing greater class separation (Tharwat et al., 2017).

5.2.3. Naïve Bayes

Naïve Bayes is one of the basic probabilistic classifiers which classifies the specific class with the given tuple. It is categorised by hypothesising that every attribute has a solitary effect on the class attribute by not depending on other attribute values (Passi & Pandey, 2018).

5.2.4 K-Nearest Neighbours (K-NN)

K-NN is one of the most straightforward and efficient classification methods. This method predicts the test data point label with the superior class of its *k* most identical points of training data (Zhang et al., 2017).

5.2.5 Classification and Regression Trees (CART)

CART is a decision tree algorithm used for classification or regression depending on the class label. If it is nominal, it classifies the dataset, or if it is numeric, it performs regression on the dataset using decision trees (Bahramirad et al., 2013).

5.2.6 Autoencoder Network

Autoencoder is a special type of Artificial Neural Network that uses an unsupervised approach for learning features, making it more efficient for small datasets with overlapping features. Autoencoders networks (deep and narrow) successfully solved complex classification problems (Tirumala, 2020).

6 Experimental Results Analysis

After applying different algorithms to the liver disease data set, accuracy, sensitivity, specificity, and confusion matrix are recorded.

Classification methods	SVM	Naïve Bayes	LDA	K-NN	CART	Autoencoders
Accuracy (%)	78.1	65.1	70.9	91.7	83.6	92.1
Sensitivity	65.58	36.54	60.58	86.54	78.85	87.65
Specificity	91.67	96.04	82.08	97.29	88.75	98.7
Correctly classified instances	781	651	709	917	836	921
Incorrectly classified instances	219	349	291	83	164	79

Table 4: *Results of different experimented algorithms*

The results obtained from the experiment, except for two algorithms, SVM and LDA, the rest three algorithms gave an acceptable level of accuracy above 75%. Autoencoders (3 layered) achieved 92.1% (921 correctly classified instance) accuracy, with K-NN achieving an almost similar level of accuracy

with correctly classified instances to 917. The lowest accuracy is for Naïve Bayes, 65.1%, with only 651 correctly classified instances.

6.1 Confusion Matrix

The confusion matrix is used to anticipate the behavioural structure of supervised learning algorithms. It is a square matrix and represents actual and predicted class values. The rows in the confusion matrix represent the actual values, and the columns represent the predicted values. In binary classification (see Figures 6a and 6b), a 2*2 matrix represents the confusion matrix consisting of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Caelen, 2017).

		Performance Measure of the Algorithm	
		Yes	**No**
Predicted	Yes	True Positive	False Positive
	No	False Negative	True Negative

Figure 6a: Confusion matrix

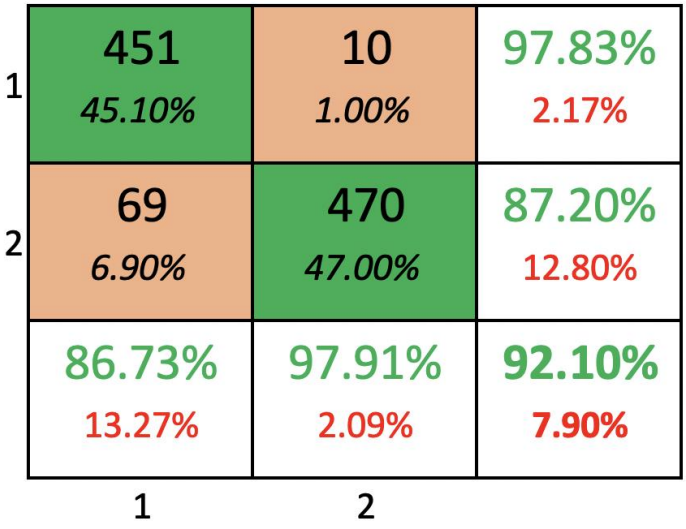

Figure 6b: Detailed confusion matrix

In this study, TP represents the correctly classified instances of liver disease patients. FP is the value of the incorrectly classified instances that the patient has no liver disease. FN represents the value of incorrectly classified instances of a patient with liver disease; TN is the value of correctly classified instances of a patient with no liver disease. In this research, five confusion matrices are generated as shown below (see Figure 7):

```
SVM                 Naive Bayes        K-NN
psvm   1   2        pnb   1   2        pknn   1   2
   1 341  40          1 190  19           1 450  13
   2 179 440          2 330 461           2  70 467

LDA                CART
plda   1   2       prpart   1   2
   1 315  86            1 410  54
   2 205 394            2 110 426
```

Figure 7: Confusion matrices for the experimented algorithms

6.2 Accuracy

Accuracy is the value of correctly classified instances in both classes (Wu et al., 2019).

Accuracy = TN+TP/(TP+FP+TN+FN)

The example calculation of accuracy for K-NN = 450+467/ (1000) = 0.917. So, the accuracy of K-NN is 91.7%. The rest of the accuracies for other algorithms is shown in Figure 8.

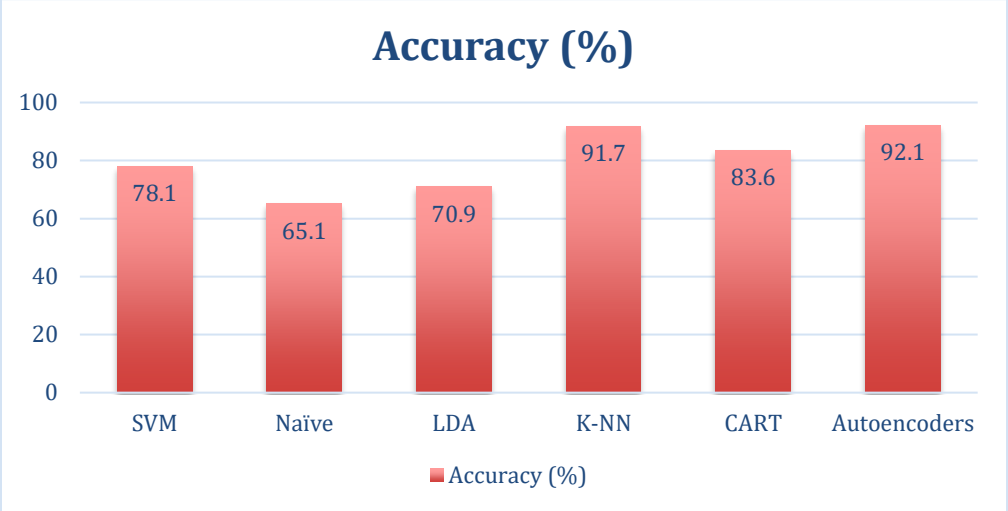

Figure 8: The accuracy results of different algorithms

6.3 Sensitivity

Sensitivity is the value of correctly classified positive instances (Coenen, 2012). It says how well the algorithm correctly classified the patient has liver disease.

Sensitivity = TP/(TP+FN)

Example calculation of Sensitivity for K-NN = 450/ (450+70) = 0.8654. The rest of the sensitivities for other algorithms is shown in Figure 9.

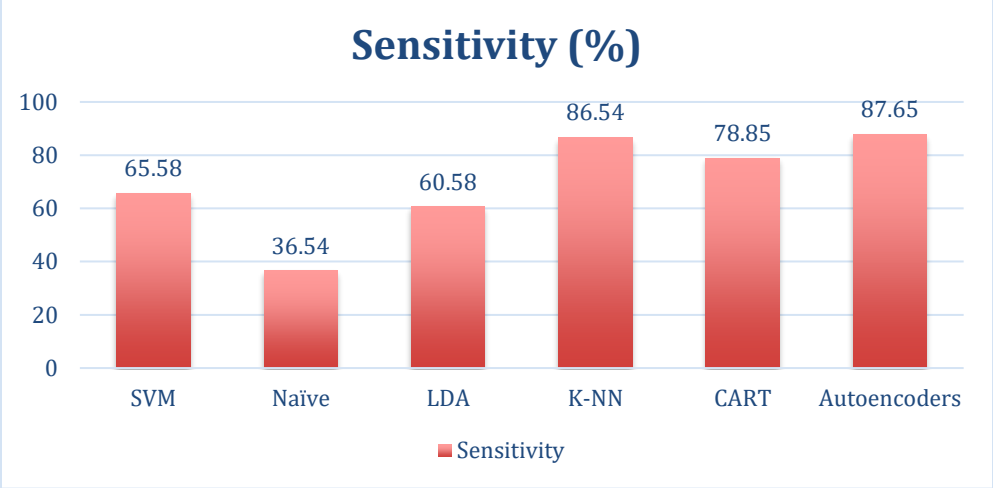

Figure 9: The sensitivity results of different algorithms

6.4 Specificity

Specificity is the value of correctly classified negative instances (Coenen, 2012). It says how well the algorithm has correctly classified that patient has no liver disease.

Specificity = TN/(TN+FP)

Specificity of K-NN= 467/ (467+13) = 0.9729. The rest of the specificities for other algorithms is shown in Figure 10.

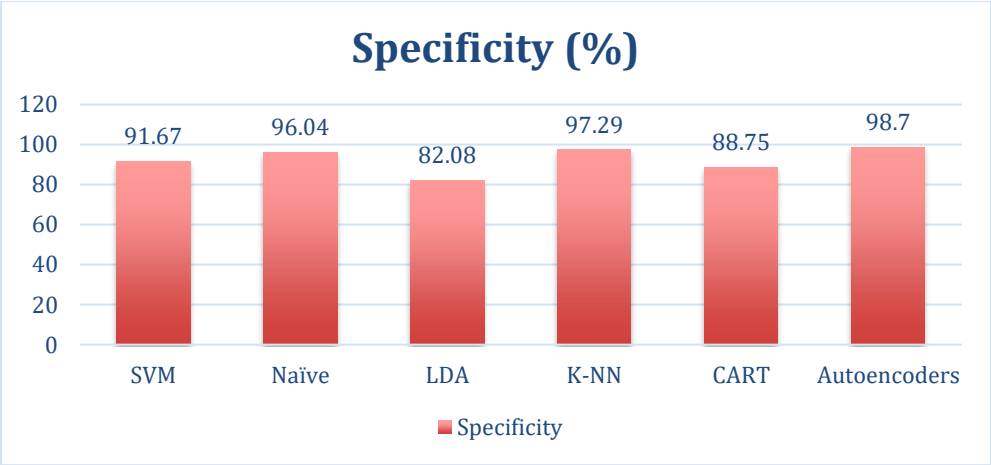

Figure 10: The sensitivity results of different algorithms

6.5 Correlation between the Attributes

In the correlation plot, it can be observed that some attributes, namely total proteins, albumin and globulin ratio, and albumin, are not much likely correlated with the class attribute, and the remaining attributes are significantly correlated with the class attribute. This correlation plot is generated using Pearson Correlation in R to know how likely the attributes are correlated.

7 Discussions and Future Recommendations

The proposed liver disease prediction (LDP) method has provided the right path for liver disease detection. From the results of this study, after balancing the dataset, SVM has 78.1%, and Naïve Bayes has 65.1%. This balancing of the dataset using ROSE significantly changes the accuracy compared to the accuracies produced by Auxilia (2018), which is 77% for SVM and 37% for Naïve Bayes.

Singh et al. (2020) also focused on the same dataset of liver patients with feature engineering done with WEKA. After feature engineering, only five attributes are selected for the analysis, and algorithms are applied. The common algorithm from this research and Singh et al. (2020) is Naïve Bayes and has an accuracy of 55.9% with only five attributes selected. By comparing that to the results of this study, Naïve Bayes has an accuracy of 65.1%. As shown in Figure 11, only three attributes are less likely correlated with class attributes, but the rest are correlated with the class attribute, affecting the accuracy. The attributes that are not correlated with class attributes can be removed, which gives the better performance of algorithms and maximised accuracy. So, from Singh et al. (2020) research, some relatable features are dissolved in feature engineering, impacting accuracy. Thus, if this research needs to be done differently, it can include some more instances for better prediction. As the given dataset has only 583 instances, they can be increased in number for a better prognosis. Along with increasing the instances, different attributes important to predict liver disease like triglycerides, urine copper, serum cholesterol, and serum glutamic-oxaloacetic transaminase (SGOT) could be added to improve the chances of liver disease prediction (Assegie et al., 2022).

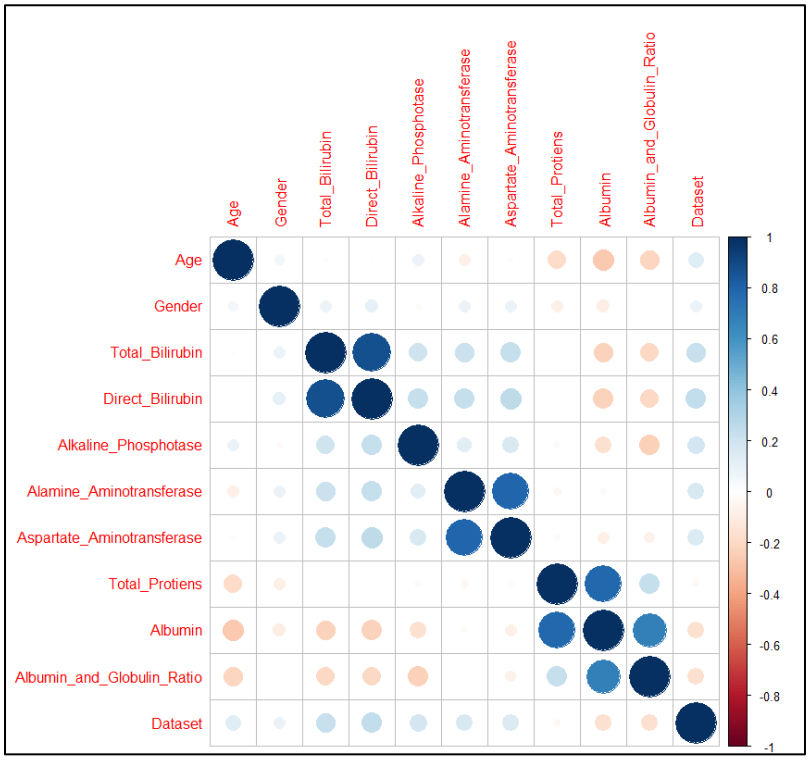

Figure 11: Correlation between the attributes

8 Conclusions

Since the liver disease is not easy to diagnose, given the delicate nature of its signs, this research is pertinent in determining the algorithms that have better accuracy in predicting this dreadful disease. The stages in the proposed LDP method provide a better alignment of each phase. Once the dataset is selected, the preprocessing step is conducted by replacing the missing values and balancing the dataset. After that, using R, five different supervised learning methods are applied (i.e., SVM, Naïve Bayes, K-NN, LDA, and CART), and the accuracy with confusion matrix metrics are recorded. The result shows that K-NN has a better accuracy of 91.7% for liver disease prediction. Autoencoders are applied in this research as a test case for understanding the classification ability of unsupervised algorithms over other traditional approaches. In this study, the autoencoder with 3-layers achieved an accuracy of 92.1%, slightly higher than K-NN due to its ability to ascertain overlapping features better than conventional K-NNs. Most of the algorithms are more than the acceptable level of accuracy, which is 75%. The results from this study would be able to assist health care professionals and relevant stakeholders in the early detection of liver disease.

References

Arbain, A. N., & Balakrishnan, B. Y. P. (2019). A comparison of data mining algorithms for liver disease prediction on imbalanced data. *International Journal of Data Science and Advanced*

Analytics (ISSN 2563-4429), *1*(1), 1–11. Retrieved from http://ijdsaa.com/index.php/welcome/article/view/2

Assegie, T. A., Subhashni, R., Kumar, N. K., Manivannan, J. P., Duraisamy, P., & Engidaye, M. F. (2022). Random forest and support vector machine based hybrid liver disease detection. *Bulletin of Electrical Engineering and Informatics*, *11*(3), 1650–1656.

Alice Auxilia, L. (2018). Accuracy prediction using machine learning techniques for Indian patient liver disease. *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 45–50. https://doi.org/10.1109/ICOEI.2018.8553682

Azevedo, A., & Santos, M. F. (2008). KDD, Semma and CRISP-DM: A parallel overview. *MCCSIS'08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*, 182–185.

Bahramirad, S., Mustapha, A., & Eshraghi, M. (2013). Classification of liver disease diagnosis: A comparative study. *2013 2nd International Conference on Informatics and Applications, ICIA 2013*, 42–46. https://doi.org/10.1109/ICoIA.2013.6650227

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429–450. https://doi.org/10.1007/s10472-017-9564-8

Coenen, F. (2012). *On the use of confusion matrixes*. University of Liverpool.

Devikanniga, D., Ramu, A., & Haldorai, A. (2020). Efficient diagnosis of liver disease using support vector machine optimised with crows search algorithm. *EAI Endorsed Transactions on Energy Web*, *7*(29). https://doi.org/10.4108/EAI.13-7-2018.164177

Dutta, K., Chandra, S., & Gourisaria, M. K. (2022). Early-Stage detection of liver disease through machine learning algorithms. *Lecture Notes in Networks and Systems*, *318*, 155–166. https://doi.org/10.1007/978-981-16-5689-7_14

El-Shafeiy, E. A., El-Desouky, A. I., & Elghamrawy, S. M. (2018). Prediction of liver diseases based on machine learning technique for big data. *International Conference on Advanced Machine Learning Technologies and Applications*, 362–374.

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties USAF School of Aviation Medicine, Randolph Field*. Texas, Tech. Report 4.

Hepatitis C. (2021, November 2). Ministry of Health NZ. https://www.health.govt.nz/your-health/conditions-and-treatments/diseases-and-illnesses/hepatitis-c

Hossain, A. I., Sikder, S., Das, A., & Dey, A. (2021). Applying machine learning classifiers on ECG dataset for predicting heart disease. *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 1–6.

Indian Liver Patient Records. (n.d.). Retrieved November 10, 2021, from https://kaggle.com/uciml/indian-liver-patient-records

Joloudari, J. H., Saadatfar, H., Dehzangi, A., & Shamshirband, S. (2019). Computer-aided decision-making for predicting liver disease using PSO-based optimised SVM with feature selection. *Informatics in Medicine Unlocked*, *17*, 100255. https://doi.org/10.1016/j.imu.2019.100255

Kemp, F. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Wiley Online Library.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for binary imbalanced learning. *R Journal*, *6*(1). https://doi.org/10.32614/RJ-2014-008

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

McCallum, A., & Nigam, K. (1998). A Comparison of event models for Naive Bayes text classification.

AAAI/ICML-98 Workshop on Learning for Text Categorisation, 41–48. https://doi.org/10.1.1.46.1529

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Nahar, N., & Ara, F. (2018). Liver disease prediction by using different decision tree techniques. *International Journal of Data Mining & Knowledge Management Process*, *8*(2), 01–09. https://doi.org/10.5121/ijdkp.2018.8201

Naseem, R., Khan, B., Shah, M. A., Wakil, K., Khan, A., Alosaimi, W., Uddin, M. I., & Alouffi, B. (2020). Performance assessment of classification algorithms on early detection of liver syndrome. *Journal of Healthcare Engineering*, *2020*.

Ortega, J. H. J. C., Lagman, A. C., Natividad, L. R. Q., Bantug, E. T., Resureccion, M. R., & Manalo, L. O. (2020). Analysis of performance of classification algorithms in mushroom poisonous detection using confusion matrix analysis. *International Journal*, *9*(1.3). https://doi.org/10.30534/ijatcse/2020/7191.32020

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138. https://doi.org/10.14445/22312803/IJCTT-V48P126

Passi, K., & Pandey, N. (2018). Increased prediction accuracy in the game of cricket using machine learning. International Journal of Data Mining & Knowledge Management Process. 8. 19-36. 10.5121/ijdkp.2018.8203.

Santos, M. F., & Azevedo, C. S. (2005). *Preâmbulo [a]" data mining: Descoberta de conhecimento em bases de dados"*. FCA-Editora de Informática, Lda.

Shaheamlung, G., Kaur, H., & Kaur, M. (2020). A survey on machine learning techniques for the diagnosis of liver disease. *Proceedings of International Conference on Intelligent Engineering and Management, ICIEM 2020*, 337–341. https://doi.org/10.1109/ICIEM48762.2020.9160097

Singh, J., Bagga, S., & Kaur, R. (2020). Software-based prediction of liver disease with feature selection and classification techniques. *Procedia Computer Science*, *167*, 1970–1980. https://doi.org/10.1016/j.procs.2020.03.226

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169–190. http://dx.doi.org/10.3233/AIC-170729

Tirumala, S. S. (2020). *A Component Based Knowledge Transfer Model for Deep Neural Networks*. Diss. Auckland University of Technology.

Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(4), 816–820.

World Total Deaths. (n.d.). World life expectancy. Retrieved October 28, 2021, from https://www.worldlifeexpectancy.com/world-rankings-total-deaths

Wu, C.-C., Yeh, W.-C., Hsu, W.-D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., Wang, Y.-C., Yang, H.-C., & Li, Y.-C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *170*, 23–29. https://doi.org/10.1016/j.cmpb.2018.12.032

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning K for KNN classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), 1–19. https://doi.org/10.1145/2990508