International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php

Enhancing Non-Alcoholic Fatty Liver Disease Prediction with Machine Learning and Recursive Feature Elimination

Koppisetti Giridhar[1], Dr.J.Nafeesa Begum[2], Bajjuri Usha Rani[3], Manasa Adusumilli[4], M.Padmavathi[5], Boddu L.V.Sivarama Krishaa[6], M. Sitharam[7], Pamula Udayarau[8]
[1]Department of CST, MITS Deemed to be University, Madanapalle, AP- India.
[2]Department of CSE, Government College of Engineering, Bargur, Krishnagiri, Tamilnadu-India.
[3]Department of CSE, Lakireddy Bali Reddy College of Engineering (A), Mylavaram, AP – India.
[4]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP – India
[5]Department of CSE, Swarna Bharathi Institute of Science and Technology(SBIT), Khammam - Telangana
[6,7,8]Department of CSE, School of SEAS, SRM University – AP, AP- India.
Email: giridhark@mits.ac.in[1] , nafeesa.research@gmail.com[2] , bajjuri.usharani2022@gmail.com[3] , padmaja.manasa@gmail.com[4], macherlapadmavathi@gmail.com[5], sivaramakrishna.b@srmap.edu.in[6], sitharam.m@srmap.edu.in[7] , udayaraju.p@srmap.edu.in[8]

ABSTRACT
Non-alcoholic disease detection is one of the leading research works in recent days. Modern life has changed the food and environmental culture, making them overweight, stressed, unhealthy conditions always and which causes various diseases due to overweight and diabetes. Commonly, an alcoholic addict can be affected by Fatty Liver Diseases (FLD), whereas identifying fatty liver diseases for a non-alcoholic person is a challenging task. It is not so easy even suspecting that a patient has FLD at the earlier stage of the symptoms since the symptoms of FLD are very similar to other diseases, and it may lead to wrong diagnosis and treatment. The severity level of 30% of FLD patients is increased suddenly and leads to heart attack, stroke, and death. Thus, based on the symptoms of weight loss, abdominal pain, and fatigue, it is essential to diagnose NAFLD, which can be identified accurately from pathological and genomic data using efficient learning methods to provide the right and better treatment immediately. This paper implements multiple machine learning algorithms for analyzing the pathological information obtained from the NAFLD and NASH DNA datasets and finding the best model concerning the performance. This paper uses 3-fold cross verification with recursive feature elimination methods to improve the original accuracy of the prediction. The performance comparison shows that the SVM model obtained 87% accuracy, which is better than the KNN and RF models. The experimental results with the performance comparison are explained in detail in the paper.
Keywords: Machine Learning Algorithm, NAFLD, NASH, Feature Based Classification.

I. INTRODUCTION

Fatty liver disease is one of the common diseases among adults, which occurs due to overweight or diabetes. The overfat in the liver may damage liver function and create liver injuries over time. Too much alcohol may also be one of the reasons for fatty liver diseases. Fatty liver diseases do not have symptoms to diagnose at the early stage. Some of the most common symptoms are fatigue, pain in the upper abdomen, and weight loss. In comparison, severe diseases have symptoms like yellow eye, dark urine, itchy skin, and blood vomiting [1]. Fatty liver diseases are classified into two types, namely non-alcoholic fatty liver disease (NAFLD) and Alcohol-related fatty liver diseases (ALD) [2].

The most severe form of NAFLD disease is called NASH. NASH-type liver diseases may lead to serious issues and even end in liver cancer. These liver diseases are diagnosed through blood tests, imaging techniques, biopsy, etc [3]. Traditionally, the liver biopsy method is widely used to detect diseases. Still, it is unsuitable for clinical practice due to its risk factors, such as invasiveness, sample error, bleeding risk, and uneven distribution of liver lesions [4]. Also, this method takes more time to diagnose the diseases and has limitations on diagnosing the early symptoms of liver diseases. And the current diagnosing system is inefficient in distinguishing different stages of FLD diseases. Therefore, an accurate, non-invasive, high-speed technique is required for quick diagnosis and treatment. So, AI-based techniques have become popular to diagnose FLD diseases [5].

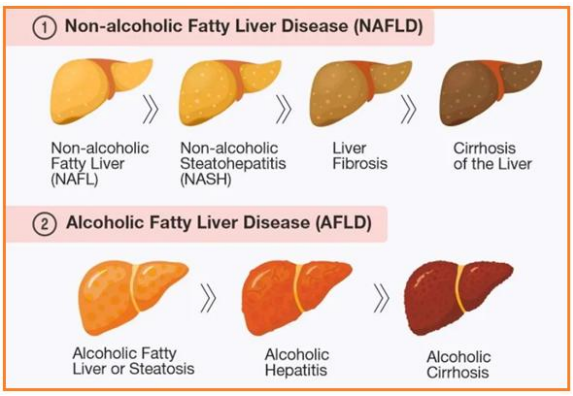


Figure 1. difference between NAFLD and AFLD

Figure 1 compares the differences between the Non-Alcoholic Fatty Liver Disease (NAFLD) and Alcoholic Fatty Liver Disease (AFLD). It shows that NAFLD has four stages such as Non-Alcoholic Fatty Liver (NAFL), Non-Alcoholic Steatohepatitis (NASH), Liver Fibrosis and Cirrhosis of the liver. Whereas, AFLD has three stages they are Alcoholic Fatty Liver of Steatosis, Alcoholic Hepatitis and Alcoholic Cirrhosis. It also shows that AFLD is more affected the liver compared to the NAFLD.

The AI-based non-invasive imaging techniques have produced more accurate results than traditional methods. The AI-based model mimics the human brain's activities to perform problem-solving and data-learning skills. This AI-based technique is further developed into two different learning models: machine learning and deep learning [6]. Machine learning is the most popular technique in various applications to perform multiple tasks. Especially in the medical sector, ML-based techniques are widely used to manage patient records, health reports, images, etc. The ML-based imaging technique transfers the healthcare sector more efficiently and quicker to progress the medical data. ML-based models are classified into supervised and unsupervised [7]. Deep learning-based models are also widely used in recent healthcare applications. DL is the machine learning model's sub-set, automatically learning input data patterns without human intervention [8]. This learning model is inherited with various imaging techniques, such as CT, MRI, Ultrasound, etc, to accurately classify fatty liver diseases from the input data [9].

Computed Tomography (CT) Scan is the most popular imaging technique, which combines the feature of X-ray with computer technology to analyze the internal parts of the human body. The main focus of the CT technique is to identify the problems in the bone structure, abdomen, chest, liver, brain, and spinal cord. However, this technique takes longer and is highly expensive to diagnose the diseases. So, magnetic resonance imaging (MRI) was developed to detect cross-sectional images of human body parts. Compared to traditional imaging techniques, the MRI-based technique scans without emitting radiation. This technique uses an efficient magnetic field to analyze the changes, and through a high-resolution computer technique, bone and soft tissue images are generated. Similarly, ultrasound techniques, which use non-invasive light waves to produce results, have recently been widely used. An optimization model is used with imaging techniques further to enhance the accuracy of the imaging technique models. The optimization algorithm provides more optimal solutions to solve complex problems. Different optimization algorithms are used: conjugate gradient, gradient descent, simulated annealing, and Newton's method. The optimization algorithm is considered to be the best tool in the field of computer vision [10]. As mentioned above, it is mainly used to find the best solution to provide maximum values. Generally, the optimization algorithm is classified into three categories such as local, global, and hybrid search techniques. Based on the input data and problems, the type of optimization algorithm is selected. In medical imaging techniques, optimization algorithms perform various functions such as image enhancement, segmentation, feature extraction, alignment, recognition, and classification. The optimization algorithms show more efficient results in diagnosing liver diseases.

In advance of this, genetic algorithms and artificial Immune system-based fatty liver disease diagnosing systems have developed in recent years. Most recent research has suggested this method as an optimal solution to predict the severity of liver diseases. So, in this paper, a genetic algorithm and artificial Immune system-based model are designed and implemented to enhance the prediction accuracy. This model analyses the diseases through a DNA dataset collected from the patients. The following section

discusses the earlier research on fatty liver disease detection, the proposed model's performance, and the proposed approach's result. It concludes with some points for future researchers.

In this article, several approaches like SVM, KNN, and RF models are contributing to the process of predicting and classifying NAFLD disease also identify the best performing model in predicting NAFLD by comparing them. To improve the classification performance, feature selection is performed by using the recursive feature elimination (RFE) method. Prevents the overfitting problem in the model which helps to provide more reliability and robustness to the model.

This study also provides an overall analysis of ML- based model. Where, the simulation result stated that SVM model has attained a high accuracy rate of 87% in classifying the NAFLD compared to the other model. This proposition paper focuses at developing a ML algorithm for the diagnosis of NAFLD in patients without resorting to invasive liver biopsy.

II. LITERATURE REVIEW

In the study [11], the NAFLD screening model was built using four machine-learning algorithms with classifiers. This study has used physical measurement variables and 12 questionnaires to establish four ML algorithms based on 304,145 subjects for NAFLD in the national physical examination population. Of four ML algorithms, XGBoost performed best with 0.880 accuracy, 0.801 precision, 0.894 recall, 0.882 F1 score, and 0.951 AUC. Finally, XGBoost outperforms the conventional statistical technique LASSO regression used in the study. In the study [12], the XGBoost model displayed the best result among other machine learning algorithms for predicting FLD. When compared to the random forest, SVM, neural network, and logistic regression, the XGBoost model showed the highest (0.882) AUROC, accuracy (0.883), sensitivity (0.833), specificity (0.683) and F1 score (0.829). In addition, fatty liver index (FLI) is compared with ML algorithms; as a result, XGBoost, neural network, and logistic regression models displayed higher AUROC than FLI.

In the retrospective cross-sectional study [13], 15,315 Chinese participants were used, and the NAFLD among the selected participants was predicted using the developed seven machine learning-based models. Biochemical factors and clinical factors are evaluated using these seven models. At the end of NAFLD prediction, the XGBoost model proved to be the best-performed ML model by showing the highest AUROC (0.873), accuracy (0.795), specificity (0.909), AUPRC (0.810), MCC (0.557), F1 score (0.695), and positive predictive value (0.806). To detect the hazard of liver fibrosis after cholecystectomy, the extreme gradient boosting (XGB) algorithm was used as an efficient predictive model in the study [14]. The proposed method achieved higher accuracy values (93.16%) and can be an automatic diagnostic aid for MASLD patients. When comparing the performance of the XGB model with KNN, the XGB algorithm revealed the highest accuracy and AUC of 93.16% and 0.92.

The study developed a machine learning algorithm (XGBoost) with logistic regression (LR) and multi-layer perceptron (MLP) models to predict NASH and fibrosis progression over four years [15]. For this, patients' electronic health records were collected for the screening. As a result, LR and MLP models are surpassed by the XGBoost model in prediction by achieving 0.79 and 0.87 (AUROC) values for NASH and fibrosis, respectively. In the other study [16], a classification model based on ML was developed to classify the subjects as NAFLD and non-NAFLD. The subject used for this study includes 14,439 adults. Four ML algorithms are used to screen the NAFLD patients, such as decision tree, random forest (RF), extreme gradient boosting (XGBoost), and support vector machine (SVM). Among them, the SVM classifier demonstrated the best performance, exhibiting the highest accuracy rate (0.801), Kappa score (0.508), F1 score (0.795), (PPV) (0.795), and (AUROC) (0.850). The second-best performance was seen in the RF model with the maximum AUROC (0.852), F1 score (0.782), PPV (0.782), and Kappa score (0.478). Lastly, based on the physical examination and blood testing findings, the SVM classifier is the most effective method to screen NAFLD in the general population. Similarly, the study [17] SVM and RF classification model achieved the highest (99%) accuracy of NAFLD prediction by using the publicly available FLD dataset. Likewise, the study conducted in China [18] demonstrated a novel-ML-based staging model by combining the stages of hepatic steatosis in 916 patients. Among various ML models such as RF, LightGBM, XGBoost, SVM, KNN, and LR, the RF model revealed the best performance with the highest accuracy (84%), AUROC (0.91).

By using the NAFLD Activity Score (NAS), non-alcoholic steatohepatitis (NASH) from the clinical and blood data collected from 181 patients was identified in the study [19] using the machine learning method. For this, SVM, random forest, AdaBoost, LightGBM, and XGBoost machine learning

algorithms are trained using features such as sequential forward selection (SFS), chi-square, analysis of variance (ANOVA), and mutual information (MI). Among the classifiers selected in the study, random forest combined with SFS scored the highest sensitivity (86.04% ± 6.21%), accuracy (81.32% ± 6.43%), Precision (81.59% ± 6.23%), specificity (70.49% ± 8.12%) and F1-score (83.75% ± 6.23%). This study highlights that it can detect NAFLD non-invasively in the early stage. To assess the NAFLD from 1119 images, the study [20] has developed a model using the combination of ML with ultrasound method, which showed higher specificity (94.6%) and Positive predictive value (PPV) (93.1%) in the prospective trial.

Based on the above discussion and survey, it is noticed that the pathological and DNA dataset needs to be analyzed in depth to get more accuracy and reduce the false positive rate. Most of the research works have used medical imaging techniques for FLD prediction. Only a very few of them have used blood data-based FLD diagnosis. The accurate diagnosis can only be obtained from pathological, genomic, and DNA data analysis.

2.1. LIMITATION AND MOTIVATION

Thus, the study uses a specific kind of pathological and genomic data, which may not be applicable to other populations. NAFLD datasets may also have some issue while distribution, it distributed imbalanced class which leads to biased model performance. This research mainly focused on the ML-based model, but it also focused on several deep learning (DL) based models which includes LSTM and CNN model which helps to provide better classification and feature extraction. However, the study does not address the use of the system in real-time with patients during the clinical practice. Some of the features picked may not be easily understandable from the clinician's perspective and therefore the decision-making process of the model cannot be relied upon solely.

2.2. PROBLEM STATEMENT

NAFLD, which is now intensely prevalent disease in the world. It is mainly affected because of increasing prevalence of metabolic disorder and obesity. Timely diagnostic and accurate detection of NAFLD is very important to prevent the patient from severe problems which includes cirrhosis, Non-Alcoholic Steatohepatitis (NASH) and cardiovascular diseases. Current diagnostic approaches such as liver biopsies are invasive, costly and take much time in the process. Thus, it is necessary to acquire an accurate machine learning (ML) based prediction model for NAFLD categorization from other liver disorders based on biochemical, pathological and genomic characteristics. Thus, this paper seeks to employ and compare several machine learning technologies, include SVM, KNN and RF to identify which of them is most effective in the classification of NAFLD. In order to increase the predictive accuracy, the model will be selected by means of recursive feature elimination (RFE), along with 3-fold cross-validation.

III. PROPOSED METHODOLOGY

The proposed methodology involves a sequence of tasks that need to be applied to the raw data, which helps to improve the programming execution efficiency and accuracy. Figure-1 illustrates the workflow of the proposed model used in this paper and are explained below.

3.1. DATA PREPARATION

Which is one of the sticky steps in machine learning projects. Every data set is unique and highly specific to the project. Even though there are some similarities during predictive modelling projects, they can provide a general flow of actions and do a particular task. The project definition is completed before data preparation, and the assessment of the machine learning algorithm is completed after data preparation. It can deliver the unknown structure of the problem to the learning algorithm**.**

3.2. DATA PRE-PROCESSING AND FEATURE EXTRACTION

This is one of the main steps in creating a machine learning model. It involves processing data to make it suitable for the model. If the process feeds unclear or noisy data to the model, it will generate an error output. Other steps in data pre-processing include data cleaning, quality assessment, and transformation. Feature extraction extracts the essential data from the pre-processed input data. This process is mainly applied to reduce the complexity of the input data. Feature selection is also the same process but is

probably enhanced to check the prediction variable or output. That feature can create simple and easy-to-understand machine language models.

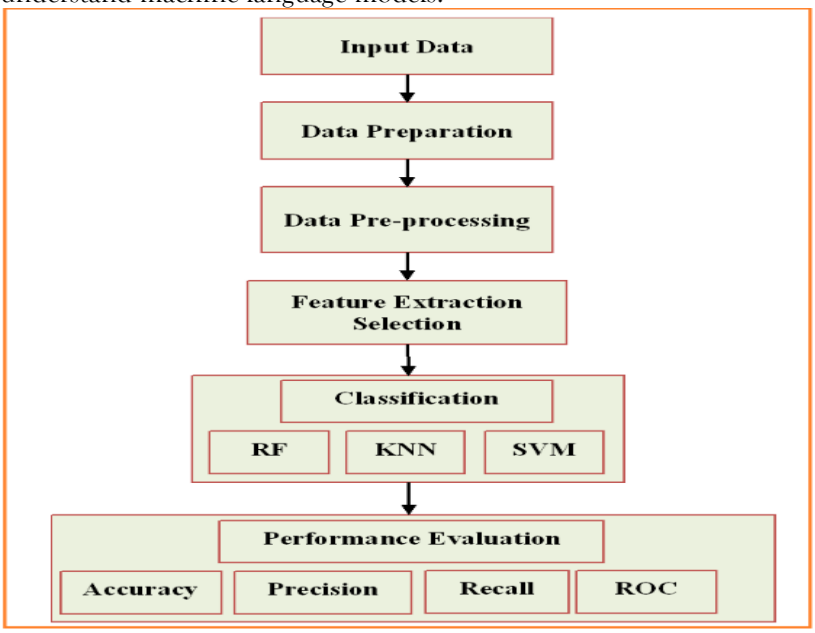


Figure-2. Proposed Work

3.3. METHODOLOGIES

Support Vector Machines (SVM): The Support Vector Machines (SVM) technique is an ML algorithm that will apply the supervised learning models to solve complex problems in classifications, detections, and regressions. The task is achieved by efficient data transmissions to define the boundaries between the data points according to the predefined labels, classes, or outcomes. There are two types of SVM: linear and non-linear. The structure of SVM is illustrated in Fgiure-2. The data is linearly divided and classified using a hyperplane line in linear SVM. The support vectors are determined by the nearest data points to the hyperplane, and those points are crucial as their changes can impact the hyperplane's position. When adding new testing data, deciding the assigned class is not dependent on which side the data reaches.

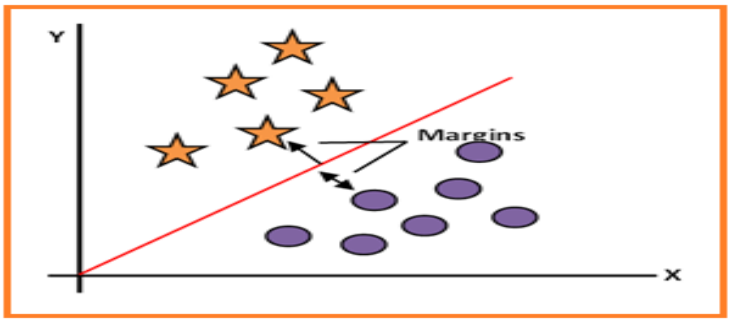


Figure-3. Support Vector Machine

SVM model mainly aims to identify the optimal hyperplane which maximally separate two classes using the following equation

$$f(x) = w^x\, x + b$$

in this above equation, where w denoted the weight vector of the model, b denoted the bias term in the model and x denoted the input feature vector.

SVM optimization was done by using the following equation: $\min_{w,b} \frac{1}{2} \|w\|^2$

$$y_i(w^T x_i + b) \geq 1, \quad \forall i$$

In this equation, $y_i \in \{-1, 1\}$ represents the class labels.

Random forest (RF): The Random Forest Algorithm (RFA) is an ML algorithm based on ensemble learning that enables the combination of the various classifiers to make an ideal model to solve complex problems, as shown in Figure-3. Various decision trees on different subsets of the given dataset create a classifier, and the average values are taken for the prediction. Henceforth, the algorithm uses every tree's prediction, and the maximum number of similar predictions is taken and analysed to determine the output.

RF model basically builds various decision trees and combines their prediction to provide more accurate result. The decision function is given below:

$$f(x) = \frac{1}{2}\sum_{t=1}^{T} h_t(x)$$

In this above equation, T represents the number of decision trees and $h_t(x)$ represents the prediction of the t^{th} tree. Finally, the class prediction result would be determined by analysing the majority voting among the decision trees.

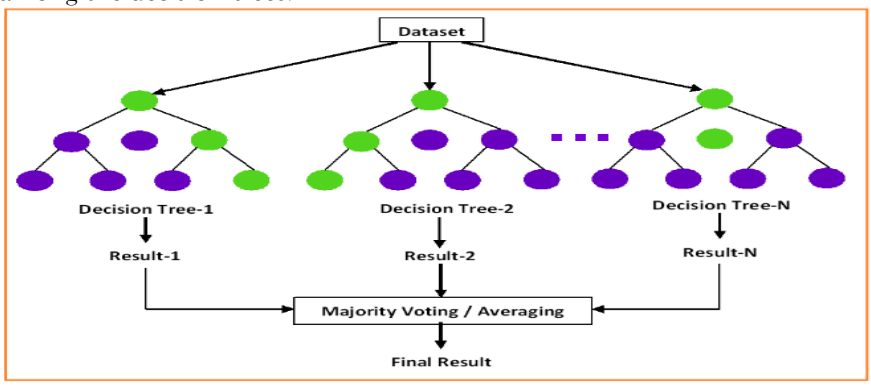


Figure -4. Random Forest Algorithm

K-nearest neighbour (KNN): The K-Nearest Neighbour Algorithm (K-NN) is the simplest ML algorithm in the supervised learning technique. This algorithm categorizes existing and new data according to their suitable similarities. The KNN algorithm quickly classifies the new data point in the input. It is used for both classification and regression processes. Compared to another method, it is a slow learner algorithm. Because it does not learn the data directly from the trained set; instead, it learns the data during classification. The identification of the new data point using the KNN algorithm is clearly shown in Figure-4(a). Figure-4(a) depicts a new data point between categories A and B. After applying the KNN algorithm, the new data point is classified similarly to category A, as shown in Figure-4(b), which is classified based on the nearest neighbours of the new data point. It is clear from the figure Category A has three neighbour points, and Category B has two neighbor points. So, the result shows the new data point is like Category A, which is classified as Category A. KNN is an approach, which classifies the new instance with relation to its closest neighbours in the feature space.

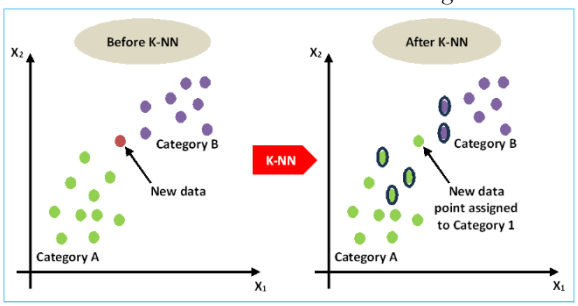


Figure-5. K-Nearest Neighbour Algorithm

It also evaluating distances of a new observation to the centroids of the grouped clusters which is normally used to make the classification decision. x All training points are usually employed, more often it uses the Euclidean distance. The equation to evaluate the Euclidean distance is given below:

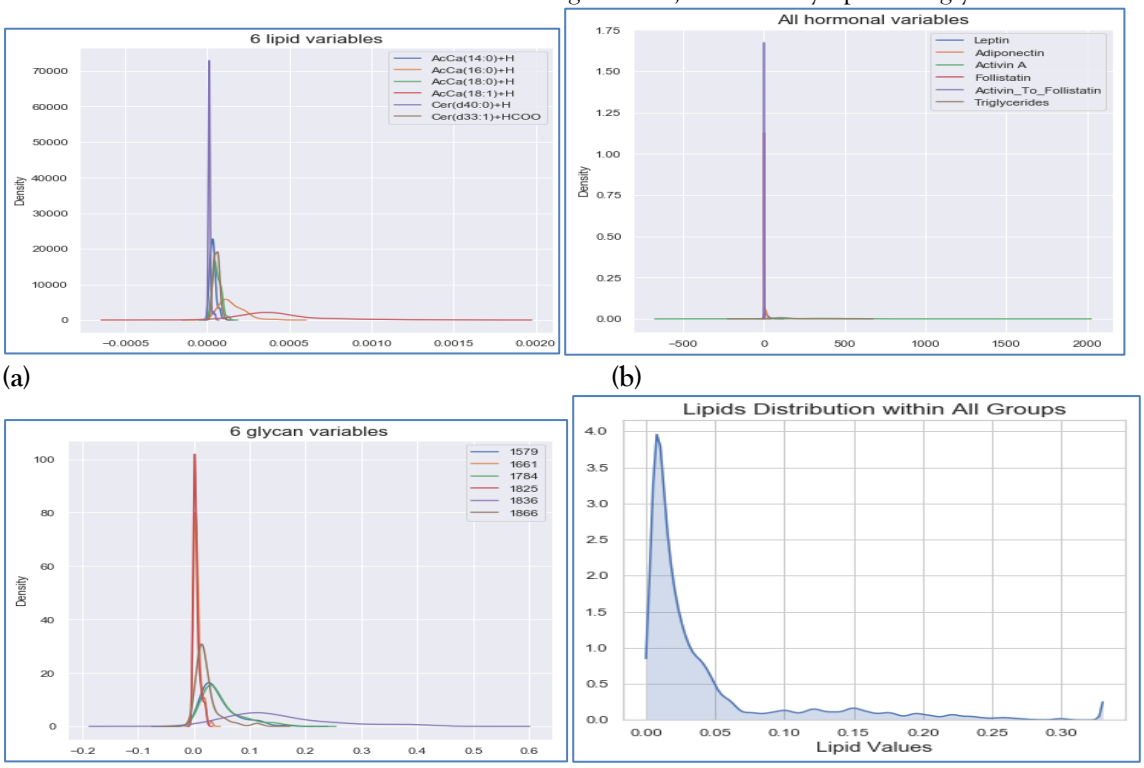
$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2}$$

In this equation, x_{ij} represents the i^{th}training instance of the j^{th} feature. Based on the majority vote of the K-nearest neighbours, the class is assigned to the model.

IV. RESULT AND DISCUSSION

In this section, various results of the proposed model for diagnosing normal and NAFLD diseases from the input medical data are discussed in detail. The presence and stages of NAFLD diseases are

classified using four types of analysis: lipid, hormonal, glycan, and free fatty acid analysis. The experiment's output is explained based on the process applied to the data. Figure-5(a), (b), and (c) depict the density of the proposed parameters lipid, hormonal, and glycan on selected 6 variables, respectively. That is, to evaluate the density of lipid (AcCa (14:0) + H, AcCa (16:0) + H, AcCa (18:0) + H, AcCa (18:1) + H, Cer(d40:0) + H, and Cer(d33:1) + HCOO) variables are selected and compared. The result shows the Cer(d40:0) + H has achieved a high-density value. Similarly, the evaluation result of the hormonal depicts Leptin, Activin_To_Follisation, Activin A, Follistatin, triglycerides, and adiponectin variables, and the Activin_To_Follisation has high density. In the analysis result of glycan with 6 variables (1579, 1661, 1784, 1825, 1836, and 1866), the variable with an 1825 value has high density. Figure-6 depicts the distribution of parameters among all data groups. That is, Figure-6(a), (b), and (C) represent the distribution values of lipids, hormones, and glycan among all groups, respectively. The overall analysis shows that hormonal data are distributed with high values, followed by lipid and glycan data.

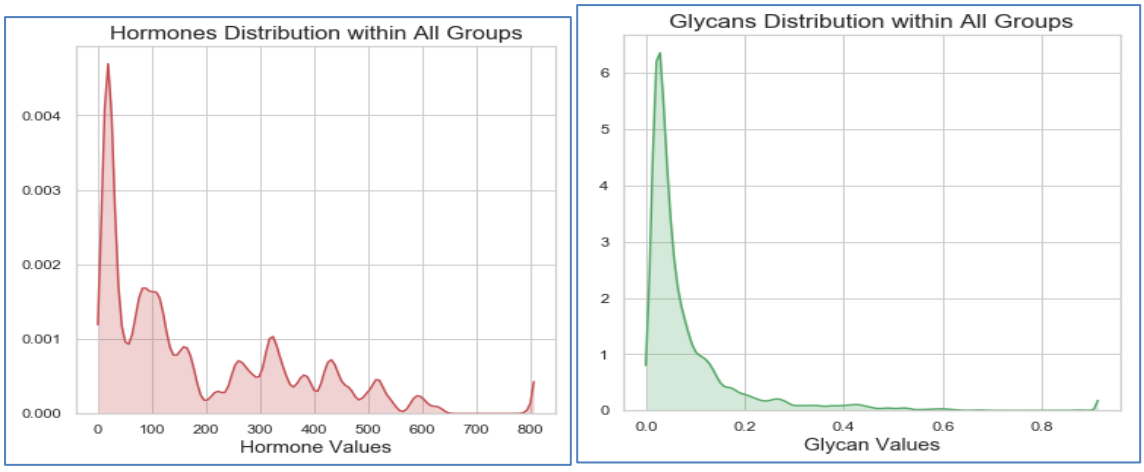


(a) (b)

(c) (D)

Figure-6. Density of the Selected Features

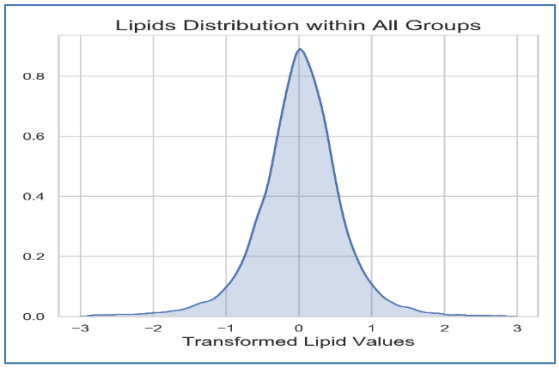
(a)

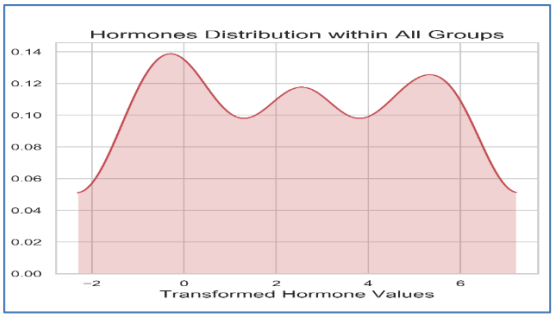


(b) (c)

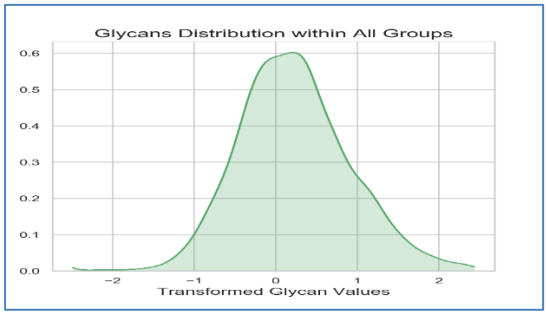
Figure-7. Parameters Distribution w.r.t Data Groups

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 8, 2025
https://theaspd.com/index.php





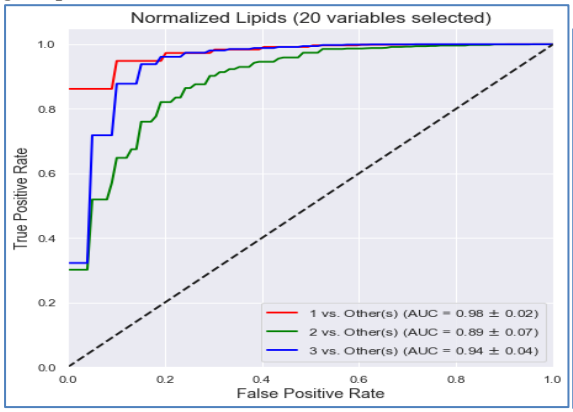
(a) . (b).



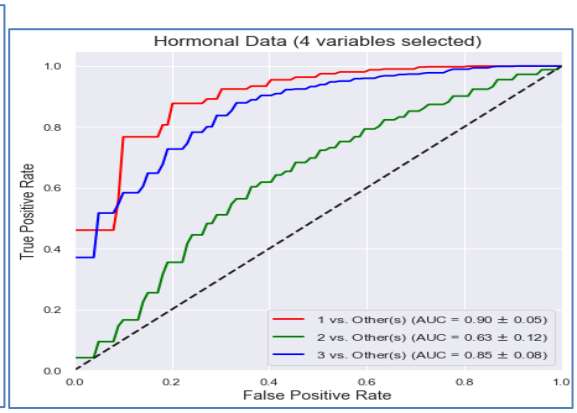
(c)

Figure-8. Feature Distribution After Transformation

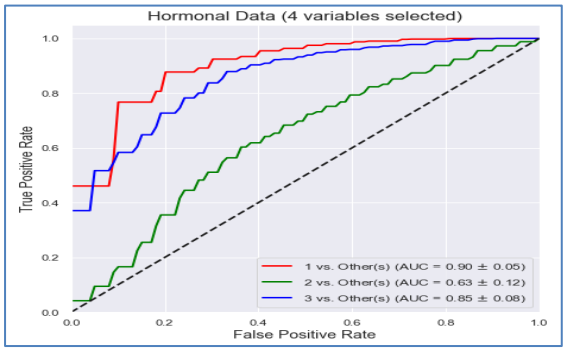
The distribution of input lipid, hormones, and glycan data among all groups after transformation is depicted in Figure-7(a), (b), and (c), respectively. The x-axis indicates the transformed values, and the Y-axis indicates the count. The analysis shows that three hormones are highly distributed within all data groups. The distribution of input lipid, hormones, and glycan data among all groups after transformation is depicted in Figure-7(a), (b), and (c), respectively. The x-axis indicates the transformed values, and the Y-axis indicates the count. The analysis shows that the hormone data is highly distributed among all data groups.



(a)



(b)



(c)

Figure-9. ROC-AUC Analysis

Figure-8(a), (b), and (c) shows the ROC curve value of the normalized lipids data, hormonal data, and glycans data with selected variables, respectively. The ROC value is estimated when the X-axis indicates the False positive rate, and the Y-axis indicates the true positive rate based on these rates. The ROC value of each feature is compared with each other, and the final output is predicted. It is achieved by splitting the selected variables into three groups (1, 2, and 3), and the ROC value of each set is evaluated. The ROC values are estimated for the combination of features by using selected variables with respective K-nearest score, F-value, and the Regressive Feature Elimination (RFE) method. The selected features are performed using 3-fold cross-validation with the iteration of 100 times. The analysis shows that the normalized lipids data achieved 0.95 AUC with the selected 20 variables, the hormonal data achieved 0.84 AUC with the selected 4 variables, and the glycan data achieved 0.75 AUC with the selected 5 variables.

4.1 Performance metrics

After evaluating the proposed parameters' ROC curve value, the proposed individual's performance and the combination of parameters are evaluated. They are evaluated using performance metrics such as accuracy, sensitivity, and specificity. The performance of the proposed model is analyzed from seven different combinations of data. The following equations calculate the proposed model's Accuracy, sensitivity, and specificity values.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100$$

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \times 100$$

Table-1. Performance Evaluation of KNN

KNN			
	Accuracy	**Sensitivity**	**Specificity**
Lipids Data	0.67	0.79	0.77
Hormonal Data	0.49	0.69	0.66
Glycans Data	0.44	0.63	0.53
Fatty Acids Data	0.42	0.59	0.58
Lipids + Glycans Data	0.69	0.81	0.80
Lipids + Hormonal Data	0.64	0.77	0.77
Lipids + Hormonal + Glycans Data	0.64	0.78	0.76

Table-2. Performance Evaluation of SVM

SVM			
	Accuracy	**Sensitivity**	**Specificity**
Lipids Data	0.87	0.90	0.92
Hormonal Data	0.54	0.67	0.81
Glycans Data	0.55	0.55	0.78
Fatty Acids Data	0.54	0.58	0.79
Lipids + Glycans Data	0.87	0.90	0.93
Lipids + Hormonal Data	0.87	0.91	0.95
Lipids + Hormonal + Glycans Data	0.87	0.91	0.95

Table-3. Performance Evaluation of RF

Random Forest			
	Accuracy	**Sensitivity**	**Specificity**
Lipids Data	0.70	0.80	0.84
Hormonal Data	0.54	0.67	0.77
Glycans Data	0.49	0.58	0.68
Fatty Acids Data	0.44	0.51	0.72
Lipids + Glycans Data	0.67	0.79	0.81
Lipids + Hormonal Data	0.64	0.77	0.80
Lipids + Hormonal + Glycans Data	0.65	0.79	0.81

Table-1 shows the trained data result of the proposed KNN model. The KNN model has achieved the highest accuracy, specificity, and sensitivity in analysing lipid data with 0.67, 0.77, and 0.79, respectively. For remaining data such as hormonal, glycan, fatty acid, lipid+glycan, lipid + hormonal, and lipid+hormonal+glycan, the KNN model predicted with 049, 0.44, 0.42, 0.69, 0.64, and 0.64 accuracy. Table-2 shows the performance metrics result of the proposed SVM model. The result of the analysis shows the proposed SVM model has predicted the Lipid, hormonal, glycan, fatty acid, lipid+glycan, lipid + hormonal, lipid+hormonal+glycan data with 0.87, 0.54, 0.55, 0.54, 0.87, 0.87, and 0.87 accuracies, 0.90, 0.67, 0.55, 0.58, 0.90, 0.91, and 0.91 of sensitivity, 0.92, 0.81, 0.78, 0.79, 0.93, 0.95, and 0.95 specificity score respectively. Table-3 shows the performance metrics result of the proposed RF model. The analysis shows that the RF model classifies the data with more than 70% accuracy, sensitivity, and specificity. The overall performance evaluation result shows that among three ML-based models, the SVM model classifies the input data with a high accuracy of 90%. All three models predict high liver fibrosis in the lipid data.

V. CONCLUSION

Nowadays, medical industries expect efficient tools to diagnose and predict various health conditions based on symptoms and causes. Related to fatty liver disease based on the symptoms of weight loss, abdominal pain, and fatigue, it is essential to diagnose NAFLD, which can be identified accurately from pathological and genomic data using efficient learning methods to provide the right and better treatment immediately. This paper implements multiple machine learning algorithms for analysing the pathological information obtained from the NAFLD and NASH DNA datasets and finding the best model concerning the performance. This paper uses 3-fold cross verification with recursive feature elimination methods to improve the original accuracy of the prediction. The performance comparison shows that the SVM model obtained 87% accuracy, which is better than the KNN and RF models. The real-time NASH dataset will be analysed using deep learning algorithms and compared with the SVM algorithm.

Future work

Develop and explore the emerging DL approaches like RNN, CNN or transformer-based architecture to enhance the NAFLD classification accuracy. It expands the dataset to Increase to include the extended population which used to provide better robustness and generalization. To interpret the decisions made by the model, it integrates the method of explainability to give doctors an easy understanding of the results. Integrating ML-based models with clinical decision support systems to help the physicians in diagnosing NAFLD in the early stage. Need to create an advanced web-based applications or well-developed diagnostic aid for predicting and monitoring the disease like NAFLD. Investigate integration of the main ideas of ML and DL models in order to improve the models' accuracy and reliability. Needs to introducing the time-series data which helps to predict the NAFLD process within time for proactive intervention.

REFERENCES

1. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/liver-fatty-liver-disease
2. Mitra, S., De, A., & Chowdhury, A. (2020). Epidemiology of non-alcoholic and alcoholic fatty liver diseases. Translational gastroenterology and hepatology, 5.
3. Cardoso, A. C., de Figueiredo-Mendes, C., & A. Villela-Nogueira, C. (2021). Current management of NAFLD/NASH. Liver International, 41, 89-94.
4. Heyens, L. J., Busschots, D., Koek, G. H., Robaeys, G., & Francque, S. (2021). Liver fibrosis in non-alcoholic fatty liver disease: from liver biopsy to non-invasive biomarkers in diagnosis and treatment. Frontiers in medicine, 8, 615978.
5. Ahn, J. C., Connell, A., Simonetto, D. A., Hughes, C., & Shah, V. H. (2021). Application of artificial intelligence for the diagnosis and treatment of liver diseases. Hepatology, 73(6), 2546-2563.
6. Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of ambient intelligence and humanized computing, 1-28.
7. Ghazal, T. M., Rehman, A. U., Saleem, M., Ahmad, M., Ahmad, S., & Mehmood, F. (2022, February). Intelligent Model to Predict Early Liver Disease using Machine Learning Technique. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-5). IEEE.
8. Xie, S., Yu, Z., & Lv, Z. (2021). Multi-Disease Prediction Based on Deep Learning: A Survey. CMES-Computer Modeling in Engineering & Sciences, 128(2).
9. https://www.complexica.com/narrow-ai-glossary/optimization-algorithms
10. Wong, G. L. H., Yuen, P. C., Ma, A. J., Chan, A. W. H., Leung, H. H. W., & Wong, V. W. S. (2021). Artificial intelligence in prediction of non-alcoholic fatty liver disease and fibrosis. Journal of gastroenterology and hepatology, 36(3), 543-550.

11. Ji, W., Xue, M., Zhang, Y., Yao, H., & Wang, Y. (2022). A Machine Learning Based Framework to Identify and Classify Non-alcoholic Fatty Liver Disease in a Large-Scale Population. Frontiers in Public Health, 10, 846118.
12. Chen, Y. Y., Lin, C. Y., Yen, H. H., Su, P. Y., Zeng, Y. H., Huang, S. P., & Liu, I. L. (2022). Machine-learning algorithm for predicting fatty liver disease in a Taiwanese population. Journal of Personalized Medicine, 12(7), 1026.
13. Liu, Y. X., Liu, X., Cen, C., Li, X., Liu, J. M., Ming, Z. Y., ... & Zheng, S. S. (2021). Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: An extended study. Hepatobiliary & Pancreatic Diseases International, 20(5), 409-415.
14. Suárez, M., Martínez, R., Torres, A. M., Ramón, A., Blasco, P., & Mateo, J. (2023). A Machine Learning-Based Method for Detecting Liver Fibrosis. Diagnostics, 13(18), 2952.
15. Ghandian, S., Thapa, R., Garikipati, A., Barnes, G., Green-Saxena, A., Calvert, J., ... & Das, R. (2022). Machine learning to predict progression of non-alcoholic fatty liver to non-alcoholic steatohepatitis or fibrosis. JGH Open, 6(3), 196-204.
16. Qin, S., Hou, X., Wen, Y., Wang, C., Tan, X., Tian, H., ... & Chu, S. (2023). Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults. Scientific Reports, 13(1), 3638.
17. Aslam, M. H., Hussain, S. F., & Ali, R. H. (2022, November). Predictive analysis on severity of Non-Alcoholic Fatty Liver Disease (NAFLD) using Machine Learning Algorithms. In 2022 17th International Conference on Emerging Technologies (ICET) (pp. 95-100). IEEE.
18. Zhang, L., Huang, Y., Huang, M., Zhao, C. H., Zhang, Y. J., & Wang, Y. (2024). Development of Cost-Effective Fatty Liver Disease Prediction Models in a Chinese Population: Statistical and Machine Learning Approaches. JMIR Formative Research, 8, e53654.
19. Naderi Yaghouti, A. R., Zamanian, H., & Shalbaf, A. (2024). Machine learning approaches for early detection of non-alcoholic steatohepatitis based on clinical and blood parameters. Scientific Reports, 14(1), 2442.
20. Tahmasebi, A., Wang, S., Wessner, C. E., Vu, T., Liu, J. B., Forsberg, F., ... & Eisenbrey, J. R. (2023). Ultrasound-Based Machine Learning Approach for Detection of Nonalcoholic Fatty Liver Disease. Journal of Ultrasound in Med
21. icine.