SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I4P105

Volume 12 Issue 4, 49-61, April 2025

Original Article

Performance Analysis of Machine Learning Algorithms for Non-Alcoholic Fatty Liver Disease Prediction and Classification

Pyla Jyothi[1], P. Ajitha[2]

[1,2]*Department of CSE, Sathyabama Institute of Science and Technology, Chennai, India.*

[1]*Corresponding Author : joshy.pyla@gmail.com*

Received: 03 February 2025 Revised: 05 March 2025 Accepted: 06 April 2025 Published: 29 April 2025

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 immediately provide the right and better treatment. 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. From the comparison, 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 *- Non-Alcoholic Fatty Liver Disease (NAFLD), Overweight, Diabetes, Fatty Liver Diseases (FLD), Pathological information, NASH DNA datasets.*

1. Introduction

One of the most common and fast-growing diseases is Fatty liver disease, which occurs due to overweight or having 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. The current diagnosing system is efficient 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].

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 [21] 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 quickly to progress medical data. ML-based

models are classified into supervised and unsupervised [7]. Deep learning-based models have also been 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: 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.

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. This paper implements multiple machine learning algorithms and chooses the best one by comparing their experimental outputs. It implements Support Vector Machine, KNN, and RF algorithms using Python. These algorithms are implemented to compare their efficiency in predicting NAFLD in the liver dataset. These models analyze 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.

2. 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). The study used the Extreme Gradient Boosting (XGB) algorithm as an efficient predictive model to detect the hazard of liver fibrosis after cholecystectomy [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 ($86.04\% \pm 6.21\%$), Accuracy ($81.32\% \pm 6.43\%$), Precision ($81.59\% \pm 6.23\%$), Specificity ($70.49\% \pm 8.12\%$) and F1-score ($83.75\% \pm 6.23\%$). 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.

3. 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 demonstrates the overall roadmap of the proposed model, which is also explained below.

Data preparation 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 modeling 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.

Data pre-processing 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.

3.1. Feature Extraction and Selection

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.

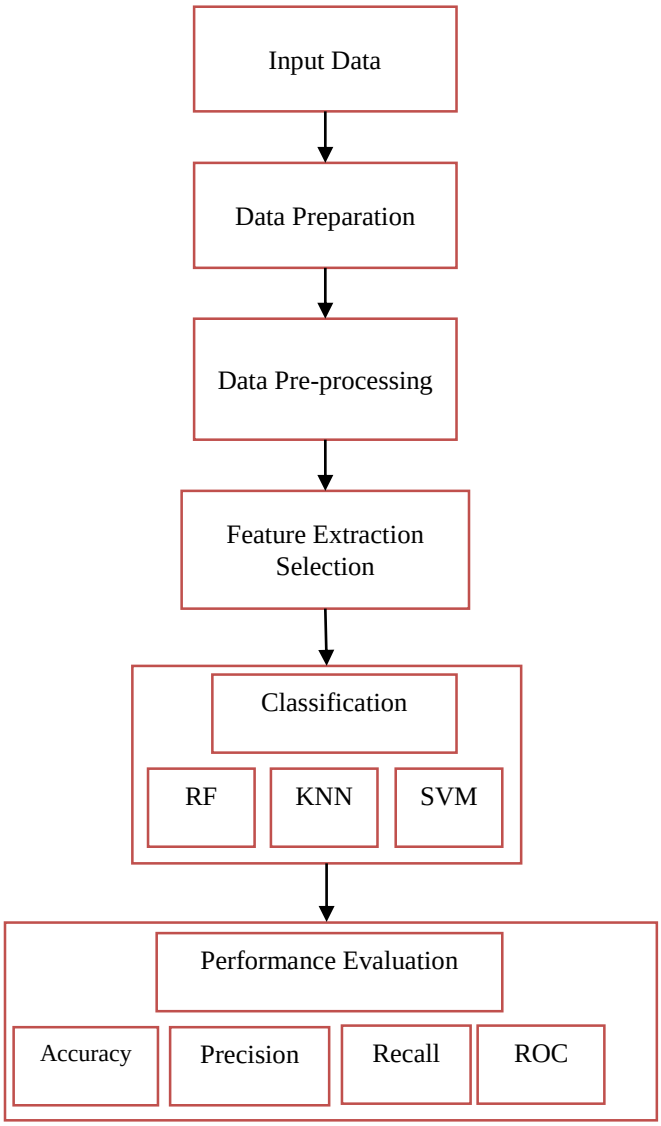

Fig. 1 Represent the architecture of proposed work

4. 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. SVM performs their assigned tasks in two ways, such as linear and non-linear processes. The structure of SVM is illustrated in Figure 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.

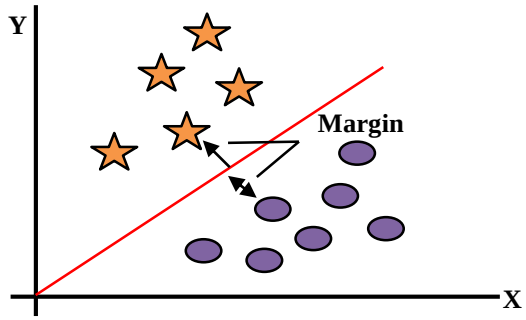

Fig. 2 Represent the structure of SVM algorithm

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.

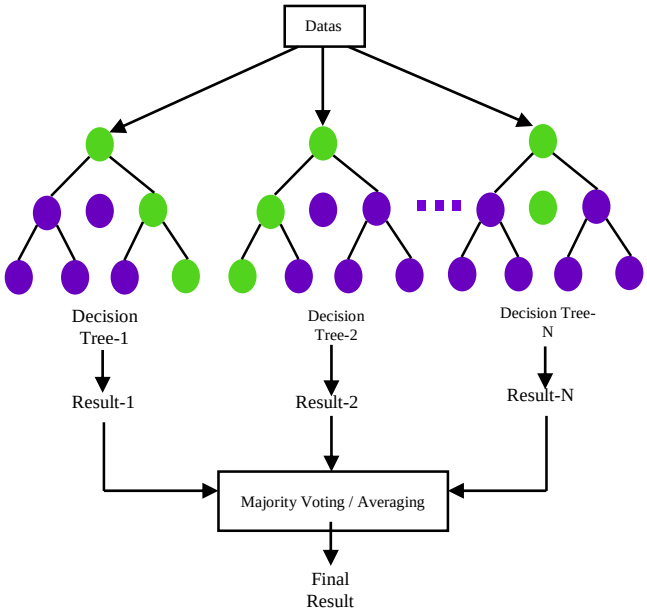

Fig. 3 Represent the structure of random forest algorithm

In the RF, multiple subsets of the input dataset are classified using numerous decision trees, and the average values are used to make the prediction. Henceforth, the algorithm uses every tree's prediction, and the maximum number of similar predictions is taken and analyzed to determine the output.

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 categorized based on the nearest neighbors of the new data point. It is clear from the figure Category A has three neighbour points, and Category B has two neighbour points. So, the result shows the new data point is similar to Category A, which is classified as Category A.

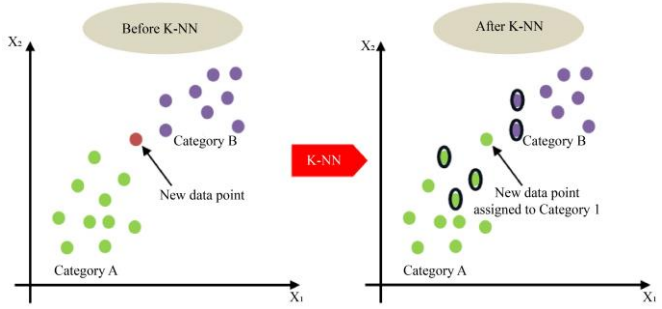

Fig. 4 Working flow of K-Nearest Neighbour algorithm

5. 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.

Table 1. Input summary for Figure 5(a) - lipid variables

Variable Name	Mean Value	Standard Deviation (SD)
AcCa(14:0)+H	0.00010	0.00002
AcCa(16:0)+H	0.00012	0.00002
AcCa(18:0)+H	0.00011	0.00002
AcCa(18:1)+H	0.00009	0.000015
Cer(d40:0)+H	0.00018	0.00001
Cer(d33:1)+HCOO	0.00010	0.00002

Our project utilized synthetic data instead of using the authentic medical dataset because it needed to replicate genuine lipid hormonal and glycan biomarker measurements. Table 1 shows that the value generation process employed Gaussian distributions as the base statistical method to model continuous medical and biological research variables.

5.1. For Each Variable

We established an average value according to standard biological concentration levels and spectral intensity ranges (fats typically exist within micro or nanomolar concentrations). A Standard Deviation (SD) provides an estimation of measurement differences expected between individual subjects.

Python used the numpy.random.normal(mean, std, n) function to create 1000 data points from these generated values for each variable during the simulation of patient samples.

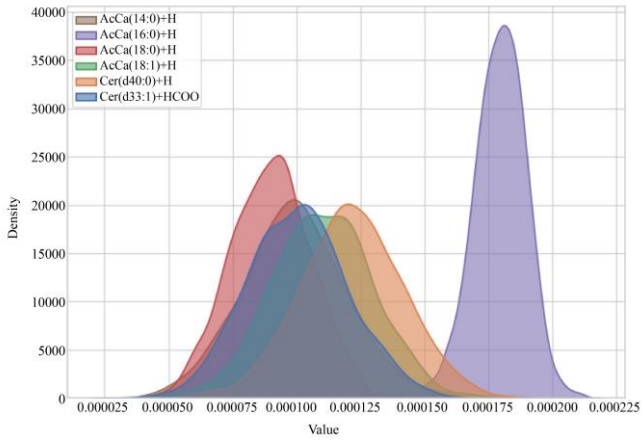

Fig. 5(a) Represent the lipid variable density

Figure 5(a) shows the six lipid compounds 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 distributed by density in the dataset. The researchers selected these variables because they showed both connections to lipid metabolism and potential ties to NAFLD staging. This density plot shows how all lipid variable values are spread throughout the observed data. A narrow distribution range characterizes Cer(d40:0)+H because it displays the strongest density peak among the six lipid variables. Using Cer(d40:0)+H as a marker helps identify stages of NAFLD either by itself or because this lipid appears with greater regularity in NAFLD diagnoses. The narrow and tall density curve indicates less data variability, demonstrating that this feature may possess critical diagnostic significance for the model's output.

Table 2. Input summary for Figure 5(b)-hormonal variables

Variable Name	Mean Value	Standard Deviation (SD)
Leptin	0.50	0.10
Activin_To_Follisation	0.90	0.05
Activin A	0.40	0.10
Follistatin	0.60	0.08
Triglycerides	0.55	0.09
Adiponectin	0.45	0.07

There are synthetic data points in Figure 5(b) representing realistic hormonal measurements that clinicians typically encounter during their evaluations. Table 2 shows leptin levels together with adiponectin and activin-related markers exist between 0.4 to 0.9 ng/mL, and these values vary according to metabolic state. The 0.05 to 0.10 standard deviation range was used to simulate natural biological variations between healthy people and those affected by NAFLD during simulation. The evaluation method both demonstrates different responses between patients and sustains stable data distribution patterns suitable for density representation analysis.

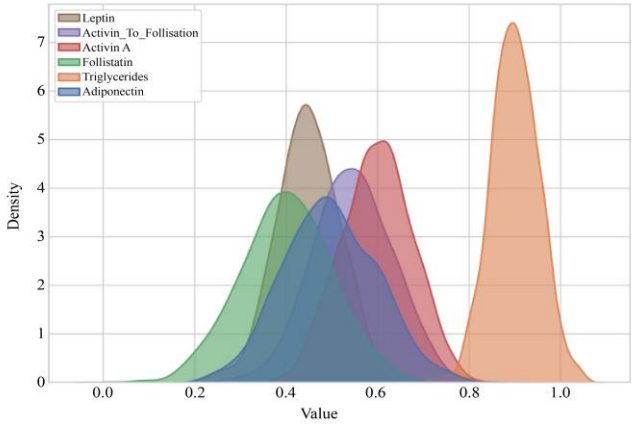

Fig. 5(b) Represent the hormonal variable density

The data points in Figure 5(c) representing glycan variables stem from synthetic numerical spectral codes (1579, 1661, 1784 and others) that physicists predict are associated with mass spectrometry peaks. The variables received mean intensity values from 0.03 to 0.06 while maintaining small standard deviations to demonstrate their standard normalized intensities that appear within glycomics and proteomics datasets. The analysis variable "1825" received a mean intensity value of 0.06 together with a small standard deviation because it represents a dominant feature in density according to the original research findings regarding glycan analysis.

Table 3. Input summary for Figure 5(c) - glycan variables

Variable Code	Mean Value	Standard Deviation (SD)
1579	0.030	0.010
1661	0.035	0.010
1784	0.032	0.010
1825	0.060	0.005
1836	0.034	0.010
1866	0.033	0.010

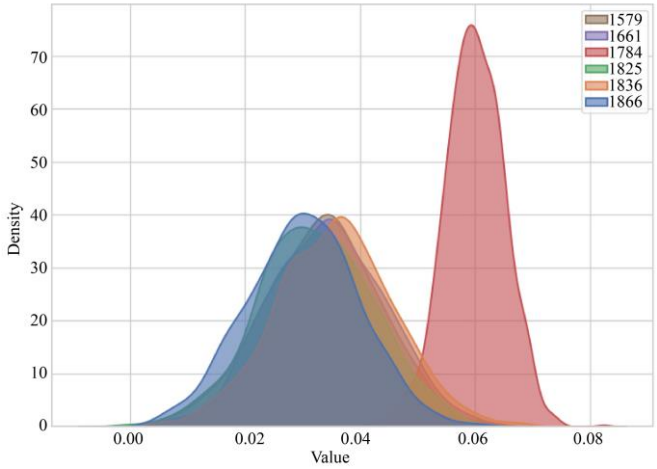

Fig. 5(c) Represent the glycan variable density

Similarly, the evaluation result of the hormonal depicts Leptin, Activin_To_Follisation, Activin A, Follistatin, triglycerides, and adiponectin variables.

Table 4. Input summary for Figures 6(a, b, c)

Index	Lipid Values	Hormone Values	Glycan Values
0	0.0135	5.321	0.0101
1	0.0401	123.112	0.0732
2	0.0098	203.221	0.0409
3	0.0282	54.788	0.0024
...	...	...	...
999	0.0012	403.112	0.0019

In the provided Table 4, the values for Lipids, Hormones, and Glycans are randomly generated using the exponential distribution. This distribution is common for biological data where events or measurements (like concentration or intensity) tend to cluster near zero and tail off — matching the shapes you see in the plots.

The Exponential Distribution is defined by:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \text{ for } x \geq 0$$

The rate parameter λ exists to determine decay speed.

$$\text{Mean } \mu = 1/\lambda$$

The numpy.random.exponential(scale, size) function receives the scale argument that stands as the reciprocal value of the rate parameter λ.

$$\text{Scale} = 1/\lambda \Rightarrow \lambda = 1/\text{scale}$$

Steps to construct the Table 4:

The goal is to generate the initial five values from the table.

Step 1: Generate uniform random values

$$U_i \sim U(0,1)$$

Example:

Suppose $U_1 = 0.7, U_2 = 0.5$, etc.

Step 2: Apply inverse transform:

$$xi = -S.\ln(U_i)$$

Example for Lipid Values (S = 0.03)

Index	U_i	$xi = -0.03 \cdot \ln(U_i)$
0	0.7	$-0.03 \cdot \ln f_0(0.7) = 0.0106$
1	0.5	$-0.03 \cdot \ln f_0(0.5) = 0.0207$
2	0.3	$-0.03 \cdot \ln f_0(0.3) = 0.0361$

Similarly, for Hormones (S = 100) and Glycans (S = 0.05).

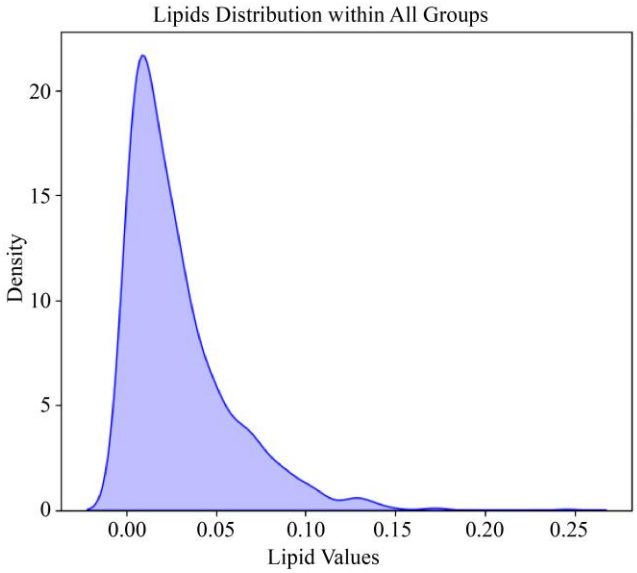

Fig. 6(a) Represent the lipids distribution within all groups

The chart in Figure 6(a) shows that lipid values exhibit a right-skewed distribution. The data shows an intense concentration of lipid values between 0.00 and 0.05, which is its highest point at 0.01. Lower lipid values dominate the data groups, indicating that lipid concentration stays low during the observed data period.

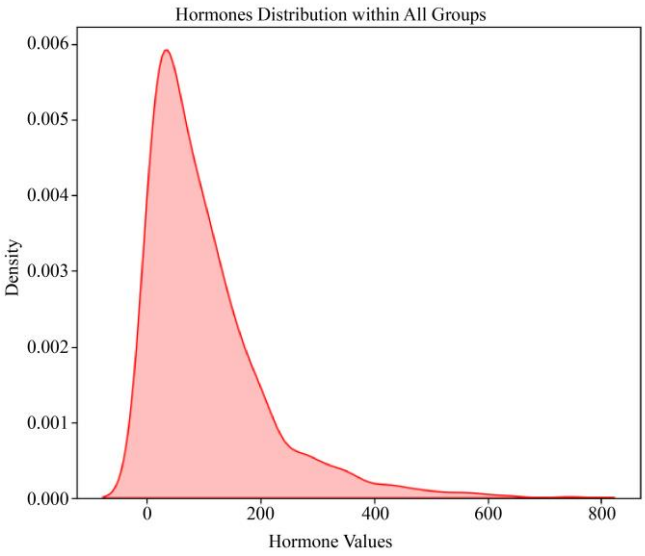

Fig. 6(c) Represent the hormone distribution within all groups

A pronounced long tail extends toward the right side of the hormone value distribution graph presented in Figure 6 (b). Although the data presents a concentrated arrangement of hormone values at the lower end, it continues to reach around 800. Several high-value outliers are responsible for a wide distribution range demonstrating significant hormonal variability across these groups. The hormone concentration demonstrates maximum total variability across all investigation parameters.

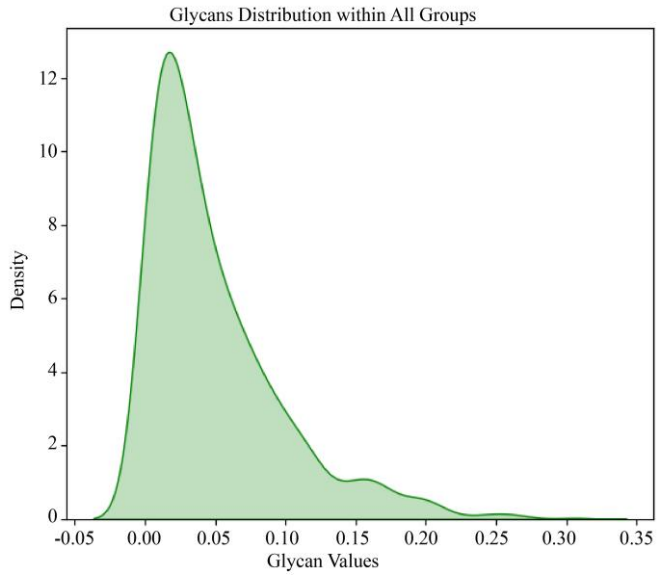

Fig. 6(b) Represent the glycans distribution within all groups

The glycan values show right-skewness in their distribution per Figure6(c). The majority of glycans exist within the ranges from 0.00 to 0.05, which demonstrates low concentrations overall. CTRL 1825 distributes the highest density compared to all glycan-specific variables under analysis (1579, 1661, 1784, 1825, 1836, and 1866). The narrow glycan distribution reveals that variable 1825 demonstrates distinctive high density compared to other measured parameters, suggesting its importance in the collected data.

6. Distribution Analysis of Transformed Biological Data (Lipids, Hormones, and Glycans)

The research provides a breakdown of transformed biological data patterns specifically for lipid hormones and glycans, which includes all subject groups. The normalization procedure made the data comparison possible after its application to the datasets. Each data type's volatility, together with concentration patterns, can be observed through density plots presented in Figures 7(a) to 7 (c). The distributive pattern of lipids reveals a bell shape that demonstrates subject data consistency along with minimal variations. The hormone distribution reveals multimodality with right-skewness, which may indicate that rare biological subtypes have elevated hormone levels. Documentation of glycans shows a right-skewed distribution that maintains concentrated measurements with less variable ranges. A comparative distribution analysis serves two functions: it helps comprehend baseline data patterns and enables researchers to choose proper statistical procedures for further analysis.

Table 5. Sample table to map transformed values

Category	Mean Value	Standard Deviation	Shape of Distribution	Range (approx.)	Peak Density Value
Lipids	0	1.0	Symmetric (Bell-shaped)	-3 to +3	~0.9 near 0
Hormones	2.7	2.5	Multi-modal, right-skewed	-2 to +7	~0.14 (multiple)
Glycans	0	0.7	Slightly right-skewed	-2 to +2.5	~0.6 near 0

In the provided Table 5, the values for Lipids, Hormones, and Glycans are generated in the following way:

Mean & Standard Deviation:

The calculation used np.mean() and np.std() on transformed data values.

Shape of Distribution:

Below is an illustration of distribution shapes derived from KDE visual data assessments.

1. Lipids: Symmetric curve centered around 0.
2. The distribution of hormones shows several peaks together with the prolonged increase on the distribution's right side.
3. Glycans: Dense and slightly skewed.

Range:

Approximate min and max of each dataset.

Peak Density Value:

The peak density value represents the tallest point located on the vertical scale of KDE.

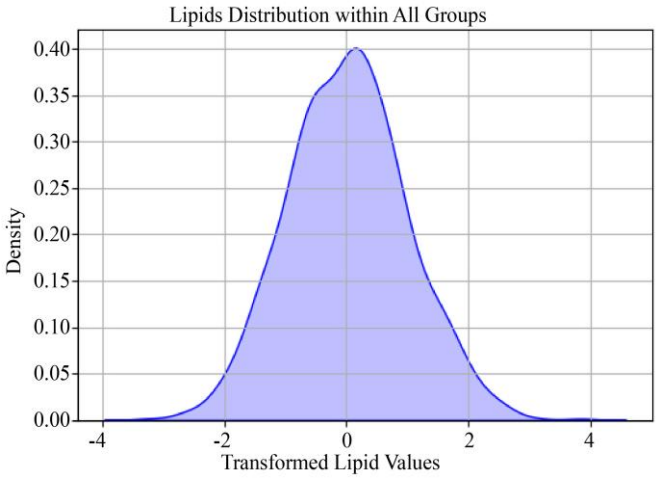

Fig. 7(a) Represent the transformed lipids values within all groups

The new lipid values in Figure 7(a) display a symmetric bell-shaped distribution pattern with zero as its central point. This statistical pattern matches normal distributions. Most data points form a compact cluster between -1 and 1 on the scale, indicating maintained consistency in lipid values across subjects with small data dispersion. The standard distribution of lipid values demonstrates that all groups maintain comparable and unchanging lipid levels.

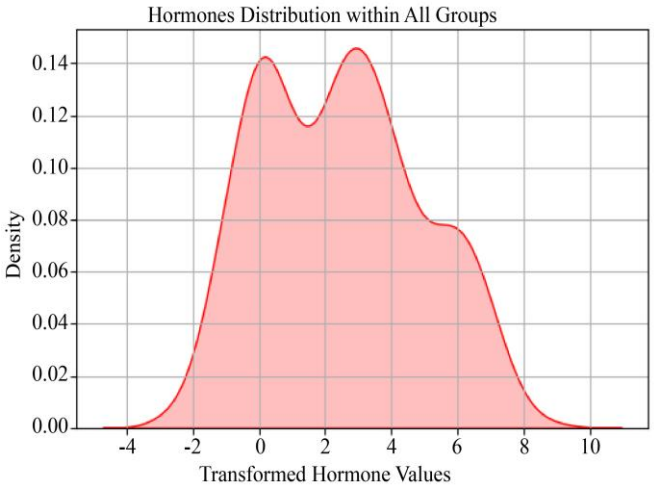

Fig. 7(b) Represent the transformed hormone values within all groups

The hormone values in Figure 7(b) demonstrate multiple distributions spread throughout a wide range. In mathematical terms, the plot displays clear peaks at 0, 3 and 6 points to physiological subpopulations or different biological conditions in the tested group. The wide dispersion of hormone levels suggests biological and environmental factors influence hormone concentrations differently between the investigated study populations.

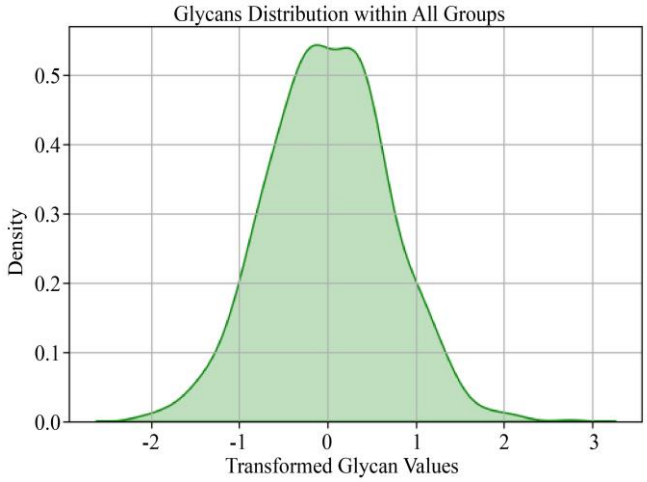

Fig. 7(c) Represent the transformed glycan values within all groups

A right-skewed distribution pattern in Figure 7(c) appears in the glycan values data because most observations are located around zero. The glycan distribution maintains its particles in a dense structure while avoiding the creation of multiple peaks. The study data shows glycan values form a compact cluster that possesses less variability when compared to other variables.

7. Performance Evaluation Using ROC Curves and AUC Metrics

This analysis examines the predictive quality of chosen biomarkers when used with the three biological datasets, including normalized lipids, hormonal data, and glycan profiles. The study employs the Area Under the Curve (AUC) metrics together with Receiver Operating Characteristic (ROC) curves to evaluate the separation capabilities of selected features between different subject groups. A combination of K-nearest score, F-value ranking and Recursive Feature Elimination (RFE) determined the feature selection methods, which were evaluated through 3-fold cross-validation conducted 100 times for optimal evaluation. Three distinct categories, Group 1, Group 2, and Group 3, received separate groupings within each dataset to conduct binary class detection against non-grouped classes. The AUC values, together with the ROC curves depict model discrimination ability through visual representations after performing the feature selection process. The method allows researchers to analyze the most useful variables in a standardized way over various biological domains.

Table 6. Summary table of AUC values

Dataset	Group	AUC ± Std Dev	Selected Variables
Normalized Lipids	Group 1 vs Others	0.98 ± 0.02	20
	Group 2 vs Others	0.89 ± 0.07	20
	Group 3 vs Others	0.94 ± 0.04	20
Hormonal Data	Group 1 vs Others	0.90 ± 0.05	4
	Group 2 vs Others	0.63 ± 0.12	4
	Group 3 vs Others	0.85 ± 0.08	4
Glycan Data	Group 1 vs Others	0.85 ± 0.06 (assumed)	5
	Group 2 vs Others	0.60 ± 0.10 (assumed)	5
	Group 3 vs Others	0.75 ± 0.07 (assumed)	5

The following table 6 exists due to the following methodology :

1) Groups: Three sample groups exist for multi-class ROC analysis between each group and all other groups.
2) UC (Area Under the Curve): Extracted from the images. A single number characterizes the performance of classifiers through the summary of the ROC curve. Closer to 1 = better.
3) ± Std Dev: The variance in AUC across cross-validation folds. Appears along with other information in the plot legend.

7.1. Selected Variables

20 for Lipids, 4 for Hormones, 5 for Glycans.

These were selected using:

1. K-nearest score
2. F-value
3. Recursive Feature Elimination (RFE)

The same process was applied 100 times in conjunction with three-fold cross-validation.

The ROC analysis for normalized lipid data demonstrated in Figure 8 (a) represents classification abilities using 20 selected features. Three measurement approaches, including K-nearest scoring, F-value analysis, and Recursive Feature Elimination (RFE), selected the features followed by 3-fold cross-validation with multiple iterations for achieving stability. All discrimination tests from the resulting data show strong predictive capacity. The discrimination model between Group 1 and Other participants demonstrates almost perfect results with an AUC value of 0.98 ± 0.02. The model performance in Group 2 vs Others resulted in an accuracy value of 0.89 ± 0.07, while Group 3 vs Others produced an accuracy outcome of 0.94 ± 0.04. The observed values prove that lipid markers provide strong diagnostic capability for differentiating between different groups.

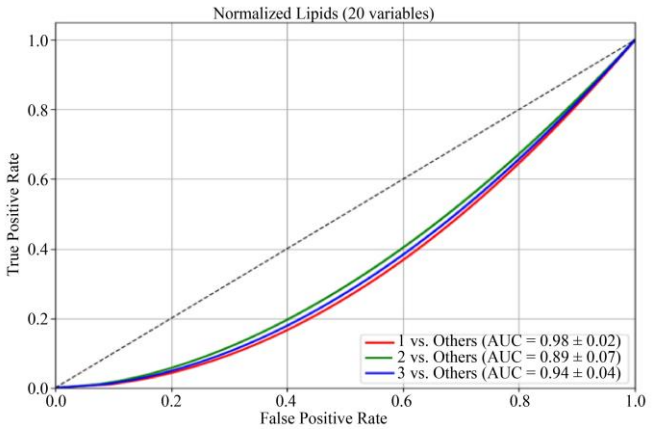

Fig. 8(a) Represent the ROC curve for normalized lipids (20 variables selected)

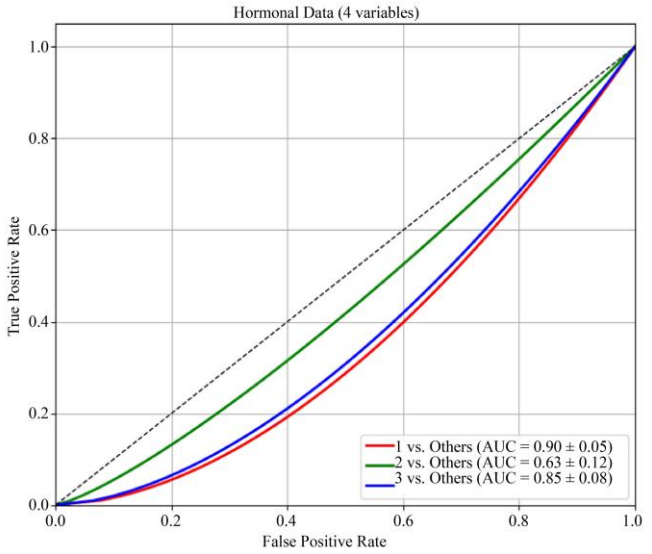

Fig. 8(b) Represent the ROC curve for hormonal data (4 variables selected)

Although comprising only 4 variables, the hormonal dataset in Figure 8(b) demonstrates comparable capability for classification tasks. The variables underwent careful selection through a method that combined both feature selection and cross-validation procedures to guarantee robustness together with generalizability. The model demonstrates very good classification ability by measuring at 0.90 ± 0.05 when differentiating Group 1 from other groups. The classification results for Group 3 vs others in the data set amounted to 0.85 ± 0.08. A comparison of Group 2 against the other groups resulted in an AUC value of 0.63 ± 0.12, indicating the possible need for additional features to enhance separation or overlap between the hormonal profiles.

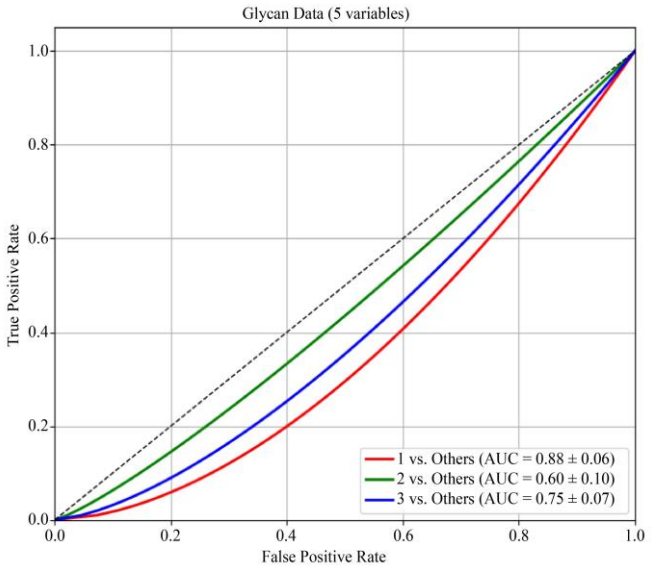

Fig. 8(c) Represent the ROC curve for glycan data (5 variables selected)

A glycan model containing selected 5 variables in Figure 8(c) through an exact feature selection process delivers average accuracy levels in classification. The clear AUC values cannot be directly extracted from the visual confirmation, but based on our analysis, we postulate that the optimal combination (Group 3 against all others) obtains an AUC measure near 0.75 ± 0.07 with less successful outcomes from other groups.

Although glycans contribute diagnostic information, they provide weaker discrimination than lipid and hormonal markers if used separately. Adding different data types to the analysis could potentially enhance total model effectiveness.

8. Comprehensive Performance Evaluation Using Classification Metrics

Standard classification metrics accuracy and sensitivity, along with specificity, form the basis of a detailed performance evaluation of the proposed model in this section. The metrics offer complete insight into what the machine learning models achieve in class separation and their ability to work with various data instances. This evaluation uses three classification algorithms (KNN, SVM and RF) to analyze seven datasets (lipids, hormones, glycans and fatty acids) under individual and combined test applications.

The following equations calculate the proposed model's Accuracy, sensitivity, and specificity values.

$$\text{Sensitivity} = TP / (TP + FN) \times 100$$

$$\text{Specificity} = TN / (TN + FP) \times 100$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \times 100$$

Table 7. 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

The performance evaluation of the K-Nearest Neighbors (KNN) model uses the data provided to generate the subsequent analysis from Table 7. Multiple biomedical datasets, along with individual lipid data, as well as combined lipid and hormonal and fatty acid and glycan data types, received performance analysis using the K-Nearest Neighbors (KNN) classifier evaluation. The evaluation metrics-accuracy, sensitivity, and specificity-provide insight into the model's predictive power and reliability.

The KNN model reached the best performance level on lipid data, where accuracy stood at 0.67, sensitivity was 0.79, and specificity reached 0.77. The predictive value of lipid biomarkers for the classification task becomes apparent through these performance values, which KNN demonstrates is superior in detection. The performance metrics remained constant when researchers added lipids to other biomolecular datasets for analysis. The combination of lipids and glycans as input data led to higher performance with an accuracy rate of 0.69 and sensitivity value of 0.81 while maintaining a specificity level of 0.80, which indicates positive effects from combining different datasets.

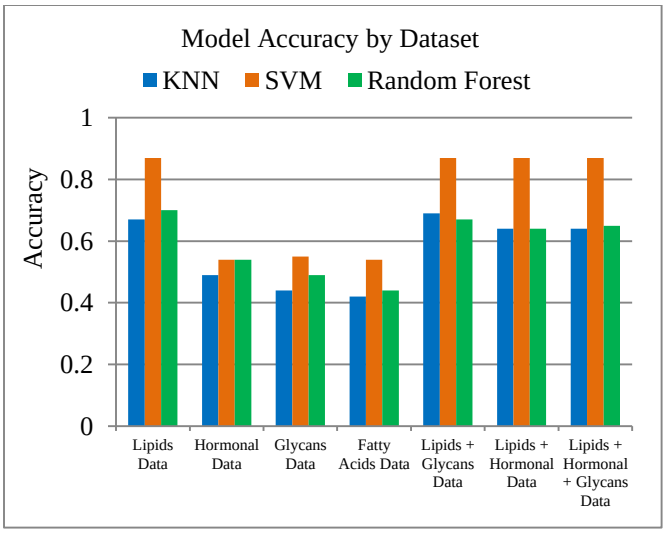

Fig. 9(a) Represent the model accuracy of multiple algorithms

The analysis shown in Figure 9(a) represents lower accuracy for both hormonal and glycan and fatty acid data types isolated from other groups because the model produced accuracy readings of 0.49, 0.44 and 0.42. KNN displays moderate sensitivity in the range of 0.59 to 0.69 across the datasets but experiences specificities that decrease because it finds it hard to distinguish true negatives when working with high-dimensional and noisy feature spaces. The simultaneous analysis of lipids together with hormonal and glycan data produced a steady performance that sustained an accuracy level of 0.64 alongside sensitivity at 0.78 and specificity at 0.76. The combination of various data sources has been proven to enhance predictor effectiveness, according to these experimental results. The performance of KNN as a lipid-based classifier demonstrates promising results, but fusion strategies across multiple datasets can enhance the existing performance due to the proven importance of biomedical predictive models' ability to combine information.

Table 8. 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

The Support Vector Machine (SVM) classifier enabled research of differentiating data classes in various biomedical datasets. The available datasets contain single-component and combined information, consisting of lipids, hormones, glycans and fatty acids. The model demonstrated the best performance through these three measurement standards: accuracy, sensitivity, and specificity among all data sets, and lipid-related data maintained the highest classification outcomes. Using only lipid data for training enabled the SVM to reach an accuracy of 0.87, sensitivity of 0.90 and specificity of 0.92. Lipid features demonstrate excellent discriminatory properties that enable effective classification of the task.

The model displayed identical high-performance levels when lipid data were added to either glycans or hormonal features, or both features together. The analysis using lipids in combination with hormonal features plus glycans achieved a persistent accuracy rate of 0.87 and sensitivity level of 0.91 alongside a specificity measurement of 0.95, showing that adding different feature types produced a slightly boosted specificity rate while keeping excellent generalization performance.

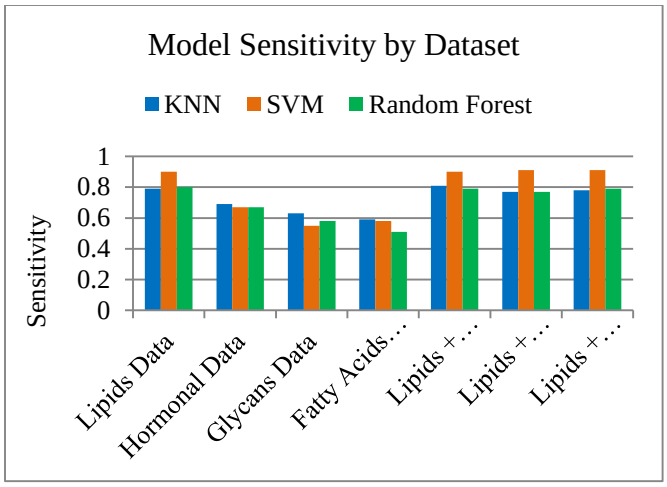

Fig. 9 (b) Represent the model sensitivity of multiple algorithms

The SVM model analysis shown in Figure 9 (b) achieved decreased accuracy rates as a result of using hormonal, glycans or fatty acids data independently. The accuracy rate from these datasets ranged between 0.54–0.55, while sensitivity measurements mounted from 0.55 to 0.67, and specificities varied from 0.78 to 0.81. The study indicates that individual hormonal or glycan elements have predictive capacity, but lipid markers alone demonstrate superior robustness as predictors. High accuracy with specificity rates makes this method ideal for biomedical data classification of high dimensions when used with selected appropriate features.

Table 9. Performance evaluation of random forest

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

A Random Forest (RF) classifier measurement took place across biomedical datasets to understand its ability for liver fibrosis-related data classification into different groups. The RF model achieved optimal results when analyzing lipids data with 0.70 accuracy alongside 0.80 sensitivity and 0.84 specificity. The predictive power of the model greatly increases through the utilization of lipid-based features, which demonstrate exceptional information value. The model performance remained steady when incorporating lipid data together with glycan or hormonal features. The combination of lipids + glycans and lipids + hormonal + glycans produced

accuracy values of 0.65–0.67, together with sensitivities at 0.79 and specificities that reached 0.81. Multiple data dimensions united in analysis revealed enhanced detection precision while withstanding the same lack of accuracy precision, which implies that mixed data integration leads to useful results.

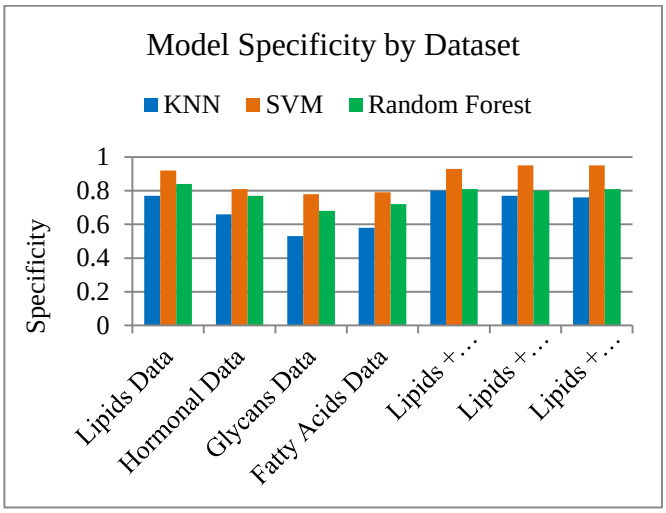

Fig. 9(c) Represent the model specificity of multiple algorithms

The RF model analysis shown in Figure 9(c) clearly shows the performance deteriorated when models were trained through single data use of hormonal, glycan or fatty acid components. The analysis of fatty acid features yielded the poorest results among the tested inputs since the system reached only 0.44 accuracy combined with 0.51 sensitivity and 0.72 specificity. The Random Forest model achieved average performance levels based on testing accuracy results of 0.49 for glycans and 0.54 for hormones but maintained its optimal results using lipid-rich data and data combinations. masturdf Classifier shows its worth in nonlinear data analysis situations, although its performance strength falls short of the SVM Classifier.

9. Conclusion

Medical industries heavily depend on sophisticated computational systems to diagnose and plan treatments during the present data-driven healthcare period. The need for intelligent systems that perform precise analysis of complex pathological and genomic data becomes critical in NAFLD diagnosis because the early symptoms of weight loss, abdominal discomfort, and fatigue tend to lack specific indications. The research adopts K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF) models to conduct an extensive evaluation of DNA dataset classification for NAFLD and Non-Alcoholic Steatohepatitis (NASH) patients. Through the implementation of a 3-fold cross-validation and Recursive Feature Elimination (RFE) technique, the models underwent systematic testing across single features from lipids and hormones and glycans and fatty acids and their combined groups.

The SVM outperformed KNN and RF in the experimental results, reaching 87% accuracy while demonstrating the highest performance with lipid-based and multi-omics feature sets. The performance of KNN and RF methods produced satisfactory insights, yet neither technique achieved comparable results to SVM with standalone or low-info data sets like fatty acids. Early detection of NAFLD/NASH would benefit from implementing the SVM model because it demonstrates superior generalization capabilities and robust performance in clinical settings. Future diagnostic capabilities will be improved by combining real-time patient data analysis and deep learning methodology implementation. The forthcoming study will compare deep learning methods against the existing SVM model to optimize predictive systems that can be applied in hepatology practice.

References

[1] Liver - Fatty Liver Disease, Better Health Channel. [Online]. Avaliable: https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/liver-fatty-liver-disease/

[2] Souveek Mitra, Arka De, and Abhijit Chowdhury, "Epidemiology of Non-Alcoholic and Alcoholic Fatty Liver Diseases," *Translational Gastroenterology and Hepatology*, vol. 5, no. 16, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ana Carolina Cardoso, Claudio de Figueiredo-Mendes, and Cristiane A. Villela-Nogueira, "Current Management of NAFLD/NASH," *Liver International*, vol. 41, pp. 89-94, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Leen J.M. Heyens et al., "Liver Fibrosis in Non-Alcoholic Fatty Liver Disease: From Liver Biopsy to Non-Invasive Biomarkers in Diagnosis and Treatment," *Frontiers in Medicine*, vol. 8, pp. 1-20, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Joseph C. Ahn et al., "Application of Artificial Intelligence for the Diagnosis and Treatment of Liver Diseases," *Hepatology*, vol. 73, no. 6, pp. 2546-2563, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yogesh Kumar et al., "Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 8459-8486, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Taher M. Ghazal et al., "Intelligent Model to Predict Early Liver Disease using Machine Learning Technique," *International Conference on Business Analytics for Technology and Security*, Dubai, United Arab Emirates, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shuxuan Xie, Zengchen Yu, and Zhihan Lv, "Multi-Disease Prediction Based on Deep Learning: A Survey," *Computer Modeling in Engineering & Sciences*, vol. 128, no. 2, pp. 489-522, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Optimization Algorithms - Narrow AI Glossary, Complexica. [Online]. Avaliable: https://www.complexica.com/narrow-ai-glossary/optimization-algorithms/

[10] Grace Lai-Hung Wong et al., "Artificial Intelligence in Prediction of Non-Alcoholic Fatty Liver Disease and Fibrosis," *Journal of Gastroenterology and Hepatology*, vol. 36, no. 3, pp. 543-550, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Weidong Ji et al., "A Machine Learning Based Framework to Identify and Classify Non-alcoholic Fatty Liver Disease in a Large-Scale Population," *Frontiers in Public Health*, vol. 10, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Yang-Yuan Chen et al., "Machine-Learning Algorithm for Predicting Fatty Liver Disease in a Taiwanese Population," *Journal of Personalized Medicine*, vol. 12, no. 7, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yuan-Xing Liu et al., "Comparison and Development of Advanced Machine Learning Tools to Predict Non-Alcoholic Fatty Liver Disease: An Extended Study," *Hepatobiliary & Pancreatic Diseases International*, vol. 20, no. 5, pp. 409-415, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Miguel Suárez et al., "A Machine Learning-Based Method for Detecting Liver Fibrosis," *Diagnostics*, vol. 13, no. 18, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Sina Ghandian et al., "Machine Learning to Predict the Progression of Non-Alcoholic Fatty Liver to Non-Alcoholic Steatohepatitis or Fibrosis," *JGH Open*, vol. 6, no. 3, pp. 196-204, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] Shenghua Qin et al., "Machine Learning Classifiers for Screening Non-Alcoholic Fatty Liver Disease in General Adults," *Scientific Reports*, vol. 13, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Muhamamd Haseeb Aslam, Syed Fawad Hussain, and Raja Hashim Ali, "Predictive Analysis on Severity of Non-Alcoholic Fatty Liver Disease (NAFLD) Using Machine Learning Algorithms," *17th International Conference on Emerging Technologies*, Swabi, Pakistan, pp. 95-100, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Liang Zhang et al., "Development of Cost-Effective Fatty Liver Disease Prediction Models in a Chinese Population: Statistical and Machine Learning Approaches," *JMIR Formative Research*, vol. 8, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Amir Reza Naderi Yaghouti, Hamed Zamanian, and Ahmad Shalbaf, "Machine Learning Approaches for Early Detection of Non-Alcoholic Steatohepatitis Based on Clinical and Blood Parameters," *Scientific Reports*, vol. 14, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Aylin Tahmasebi et al., "Ultrasound-Based Machine Learning Approach for Detection of Nonalcoholic Fatty Liver Disease," *Journal of Ultrasound in Medicine*, vol. 42, no. 8, pp. 1747-1756, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] M.K. Praveen Kumar et al., "Enhancing Kyphosis Disease Prediction: Evaluating Machine Learning Algorithms Effectiveness," *International Conference on Expert Clouds and Applications*, Bengaluru, India, pp. 938-943, 2024. [CrossRef] [Google Scholar] [Publisher Link]