PREDICTION OF NON-ALCOHOLIC FATTY LIVER DISEASE (NAFLD) USING DNA PATHOLOGICAL DATA AND SUPPORT VECTOR MACHINES

[1]T. V. K. P. PRASAD, [2]SRI LEKHA BANDLA, [3]N. SRIKANTH, [4]K KAVYA RAMYA SREE, [5]INAKOLLU ASWANI, [6]PAMULA UDAYARAJU, [7]BODDU L V SIVA RAMA KRISHNA

[1]Department of CSE, SRKR Engineering College, Bhimavaram, AP, India.
[2]Master's in biomedical engineering, University of Bridgeport, USA.
[3]Department of CSE, Lakireddy Balireddy college of engineering, Mylavaram, AP, India.
[4]Department of AIML, Aditya University, Surampalem, AP, India.
[5]Department of CSE, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, AP, India.
[6,7]Assistant Professor, Department of CSE, School of SEAS, SRM University – AP, India.

tvkpp@srkrec.ac.in[1], sbandla@my.bridgeport.edu[2] , srinekkalapu51@gmail.com[3], kavyaramyasreek@adityauniversity.in[4] , inakolluaswani2104@gmail.com[5], udayaraju.p@srmap.edu.in[6], sivaramakrishna.b@srmap.edu.in[7]

ABSTRACT

Non-Alcoholic Fatty Liver Disease (NAFLD) has emerged as one of the most prevalent liver disorders globally, affecting nearly one-third of the population, with particularly high incidence rates in countries like the UK. Despite its widespread occurrence, accurate estimation of its prevalence remains a challenge. Early-stage NAFLD, typically characterized by simple steatosis, can silently progress to more severe conditions such as non-alcoholic steatohepatitis (NASH), fibrosis, and cirrhosis if left untreated. This progression significantly compromises liver function and increases the risk of cardiovascular complications. However, current diagnostic methods, including magnetic resonance spectroscopy and ultrasound imaging, are often limited by cost, accessibility, and diagnostic specificity. Given the clinical urgency and the limitations of conventional diagnostics, this study addresses the critical need for an accessible and accurate method to detect early-stage liver disease—specifically, to predict NASH within the NAFLD spectrum. We propose a machine learning-based approach that leverages clinical and pathological data, including blood parameters and ultrasound-derived tissue characteristics, to support early detection. Using a dataset of 181 patients, we applied preprocessing techniques such as normalization and categorical encoding to prepare the data for modelling. Features such as integrated backscatter (IB), Q-factor, and homogeneity factor (HF) were extracted to quantify liver tissue characteristics. Support Vector Machine (SVM), chosen for its balance of simplicity and efficiency in handling high-dimensional datasets, was employed for classification and regression tasks. Experimental validation using Python-based implementations demonstrated the model's effectiveness, achieving an average accuracy of 89.95% across both clinical and imaging-derived datasets. This study underscores the potential of machine learning in improving early diagnosis of liver diseases and reducing their long-term clinical burden.

Keywords: *Fatty Liver Diseases, Non-Alcoholic Fatty Liver Diseases, NASH, SVM, Pathological Information, Machine Learning.*

1. INTRODUCTION

Alcoholic Fatty Liver disease also affects people who do not take alcohol. It became a chronic disease affecting common people due to food habits and obesity. Fatty liver has emerged as a global health concern, characterized by the excess of fat that accumulates in the liver that is 10% greater than the weight of a normal liver. The failure of the liver to break down lipids is the cause of the excess deposit of fat and causes fatty liver disease (FLD). People with obesity, diabetes, or high triglycerides tend to have fatty liver. Even though this fatty liver will not harm anything at the earlier stage, it creates inflammation at the severe stage. They are also known as hepatic steatosis and steatosis liver disease (SLD) [1]. FLD is caused by factors such as fatty diets, ready-made foods, beverages, sedentary lifestyles, excessive alcohol consumption, and fructose metabolism. FLD caused by the consumption of alcohol develops alcoholic steatohepatitis. In the alternative scenario, Non-

Alcoholic Fatty Liver Disease (NAFLD) is caused by inflammation in the liver, which affects mainly non-alcoholic persons. The earlier symptoms of NAFLD cannot be identified easily. In contrast, the symptoms of its severity level can be identified through weariness, weight loss, disorientation, and stomach discomfort if the condition worsens [2]. Fatty Liver Diseases (FLD) affect people who drink alcohol regularly, whereas Non-Alcoholic Fatty Liver (NAFLD) affects people who are not consuming alcohol. Whenever an alcoholic patient is undergoing a medical diagnosis, medical experts immediately take liver tests, not for non-alcoholic patients. This confusion increases the severity level of FLD for non-alcoholic patients and causes heart attacks and sudden deaths. The death ratio of NAFLD is increasing nowadays because of wrong diagnosis and not diagnosing livers and its dangerous conditions. Patients with advanced stages of NAFLD develop symptoms that include Esophageal varices, cirrhosis, and liver carcinoma [3]. As shown in Figure 1, NAFLD will transform into cirrhosis, which is one of the severe stages of fatty liver, leading to tissue damage, liver failure, and cancer if untreated [4]. Patients frequently find out about their fatty liver after testing for unrelated conditions. Fatty liver patients are generally middle-aged and overweight. The most frequent risk factors connected to fatty liver disease are overweight, with a 25–30 body mass index, obesity, higher triglyceride levels, and heart attacks.

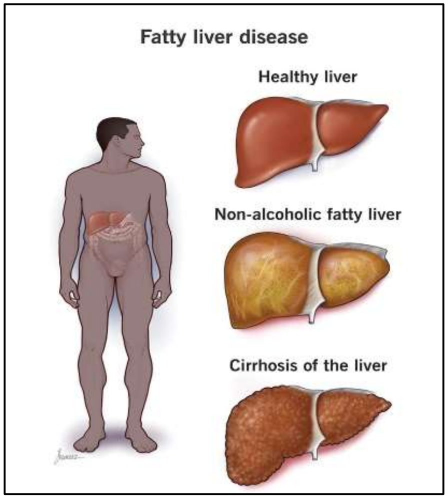


Figure 1: Morphological Difference Between Normal and Fatty Liver [4]

NAFLD raises mortality and morbidity rates [5], which leads to economic loss, mainly in Western countries. It affects 25% of adults in the US and Europe and causes severe problems like NASH, fibrosis, cirrhosis, and even death. Worldwide, the prevalence of NAFLD is higher in South America (31%) and lower in Africa (14%), and there was a 47.15% increase in cirrhosis death cases globally [6, 7, 8]. Compared to women with NAFLD, men had a higher risk of this liver-related disease above 20 of their age [5]. Primary screening is needed to overcome this fatal disease, which can be identified using various diagnosis methods like medical images, scans, and blood tests. There are three main ways to diagnose non-alcoholic fatty liver disease (NAFLD) they are blood tests, which are used to detect inflammation in the liver, and several imaging tests (MRI, CT scan, and ultrasound) to visualize the liver, and classy testing (transient elastography) to measure stiffness in the fatty liver. However, ultrasound-based testing might not be as dependable in recognizing advanced liver diseases such as NASH and fibrosis [9]. The need for NAFLD screening is essential since there is a lack of effective treatment; if NAFLD remains undiagnosed, cirrhosis will double by 2030 [10,11]. These days, genetic algorithms are frequently used in robotics, image processing, machine learning, automatic control, and other fields. However, very little work has been done using these algorithms to process DNA datasets for diagnosing FLD. Liver disorders were diagnosed by combining an Artificial Immune System (AIS) with a Genetic Algorithm [12]. The AIS was a vital component of the system architecture, and a genetic algorithm was used in the learning process to deduce how the antigen and antibody population evolved. Thus, it is found that AIS and GA algorithms [12] are highly suitable for processing and diagnosing NAFLD data. One of the studies [14] proved that 4,312 patients were affected by NASH from their overall data sample of 26,404 collected from the medical industry. They have used GWAS and SAIGE software tools to predict NASH.

This remains a real and urgent problem because Non-Alcoholic Fatty Liver Disease (NAFLD) affects a significant portion of the global population—nearly one in three people—yet its early detection and accurate diagnosis are still major clinical challenges. The disease often progresses silently from benign fat accumulation to serious conditions like NASH, fibrosis, and cirrhosis, which can severely impair liver function and increase cardiovascular risk. Current diagnostic tools, such as magnetic resonance spectroscopy and ultrasound, are either too expensive, not widely accessible, or lack the specificity to distinguish disease stages effectively. As a result, many cases go undetected until the disease has significantly advanced, highlighting the critical need for affordable, non-invasive, and accurate diagnostic approaches—such as the machine learning method proposed in this

study—to support early intervention and reduce long-term health burdens.
In the earlier days of medical industries, various sequencing technologies were applied in metagenomics to provide sequencing fragments obtained from thousands of species of human gene patterns. Only by analyzing the human genomic data can it be very easy to identify the presence of FLDs. Predicting abnormal genes in metagenomics fragments is a crucial task and one of the most fundamental problems in metagenomics, and it provides less accuracy. The FLD and its severity levels can also be predicted using pathological data and are very easy to analyze without using complex algorithms.

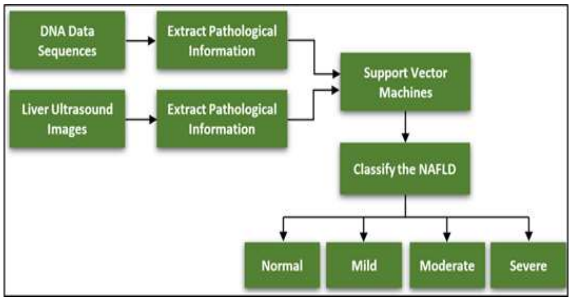


Figure-2. Proposed Model

This paper aims to understand humans' pathological information and the various blood parameters used in the clinical traits to analyze and predict liver diseases. Still, there are very few tools available to diagnose the DNA dataset for predicting various tumor and cancer-based diseases, not for NASH. Thus, this paper aims to implement an SVM algorithm for NASH prediction from the pathological information obtained from the DNA and Liver Ultrasound dataset. The overall structure of this paper's proposed method is illustrated in Figure-2 and provides four different classes.

a machine learning model using clinical and ultrasound-derived features can accurately predict NASH, offering a cost-effective alternative to current diagnostic methods.

The prediction accuracy of NAFL disease detection is increased by contributing the following processes:

- The proposed methodology of the paper is explained to understand the problem of fatty liver disease detection and its importance. It helps to design the proposed model for the FLD data.
- It explained various factors of pathological data well, through which FLD can be identified.
- The theoretical model of the SVM is explained figuratively with mathematical expressions.
- SVM is implemented and experimented with the benchmark pathological dataset, and the prediction accuracy is verified.
- The model's performance is evaluated by verifying the output in different aspects.

2. LITERATURE SURVEY

In this section, various earlier research work on fatty liver diseases is reviewed and discussed in detail. This research mainly defines the optimal model for diagnosing fatty liver diseases. Non-alcoholic is one of the common types of liver diseases. Most researchers have suggested many methods to detect NALD from the earlier stage. Some of the recent research works are discussed. AI-based techniques are widely used in the medical field to solve more problems [11]. In that sense, diagnosing fatty liver diseases is also achieved through AI-based models. It brought many changes in the healthcare sector in detecting diseases at an early stage. The AI-based technique is applied with invasive and non-invasive treatment techniques to diagnose fatty liver diseases accurately. Some of the most common AI-based models are logistic regression, random forest, XGBoost, and decision trees used for single data; the RNN model is used for sequential data, and the DNN model is used to learn the images and histology data.
A machine learning-based approach was developed to detect fatty liver diseases [12-13]. Around 577 patient records were taken as input samples to evaluate the model performance. Of these, 377 patients are with FLD; the remaining records are healthy reports. These data are classified using proposed RB, NB, ANN, and LR algorithms. Then, the model's efficiency is evaluated using ROC and performance metrics. The model's accuracy is assessed by applying the 10-fold cross-validation with proposed approaches. The analysis results indicate that the proposed model [12] has achieved 87.48, 82.65, 81.85, and 76.96 accuracy in the RF, NB, ANN, and LR models. The ML-based approach proposed by the authors in [13] has achieved 76.30, 74.10, and 64.90 accuracy, sensitivity, and specificity, respectively.
A deep learning approach is developed and implemented to improve further the model's accuracy and processing speed [14]. Ultrasound-based detection is one of the most common imaging

techniques for detecting FLD. To improve the model accuracy, the DL model is evaluated with TL models (CNN and VGG16). The analysis shows that the proposed model has achieved 90.6% accuracy in classifying fatty liver diseases. The efficiency of the deep learning model in detecting NAFLD is defined by comparing the uses of the DL model with three imaging techniques [15]. For this, 240 patients' reports are taken, which include four categories of patient reports: normal, mild, moderate, and severe. This model has classified the input data with 0.958 accuracy. An extensive study is conducted to describe the bacterial metataxonomic signature in the NAFLD [16]. The study comprised cohort patients with NAFLD, as proven by a biopsy. The differences in the phenotypic features of the patients are studied. The patients suffered from moderate obesity. The changes in the microbial characteristics of the diseases are identified using DNA profiling—the sample tissues of 116 patients with 47 overweight and 50 morbidly obese patients. The study showed that NAFLD patients had a diverse repertoire of DNA sequences.

Usually, ultrasound images are used to predict fatty liver diseases. However, the low-quality images make it difficult for radiologists to predict the disease. It is overcome with computer-aided diagnosis models that use ML and classification algorithms for categorizing the tissues. The authors in [17] have used genetic algorithms to classify fatty liver disease automatically. It is a supervised learning algorithm trained using ultrasound liver images for better prediction. The proposed voting-based classification model provided 95.71% accuracy, well above the J48 algorithm used for comparison. Big data effectively manages the large amount of fatty liver disease data available. A genetic algorithm is used to classify fatty liver disease data [18]. The physical examination records are analyzed using the proposed model and classified based on the tissue type. A genetic algorithm is usually used in classification, which involves more data and difficulty defining classes. Among the various methods used for NAFLD prediction, liver biopsy is the most considered, with sample errors that are difficult to interpret. So, a new non-invasive model is developed to predict NAFLD, using an intelligent scheme utilizing forward, Viterbi, and Baum-welch algorithms [19]. It provided better results in the prediction process compared to the biopsy. Classification algorithms like KNN, K-Means, RF, SVM, and other models are also used to predict fatty liver disease. However, these algorithms provided poor results compared to the genetic algorithm [20]. Machine learning algorithms also classify the gut microbiome signature for predicting fatty liver diseases. This is done based on the insulin resistance studied by Kang et al. 2022, providing a better prediction result of 0.83% accuracy.

When compared to literature:

- Recent deep learning models using **CNNs and VGG16** achieved higher accuracies (up to **95.71%**) on larger datasets and with more complex architectures, particularly in imaging tasks ([17], [14]).
- Other approaches, such as the **genetic algorithm-based models**, have shown robust classification results for fatty liver using ultrasound data, achieving even higher accuracies in some cases ([17], [20]).
- Studies combining **clinical data, imaging, and microbiome profiling** (e.g., Kang et al., 2022) demonstrated holistic prediction strategies, achieving accuracies around **83%**, while capturing biological complexity not addressed in this work.
- The deep learning models evaluated across imaging modalities ([15]) also outperformed traditional ML techniques, with reported accuracies as high as **0.958**.

Thus, while this study's SVM-based approach offers a **simpler and computationally efficient** solution, it may be **less effective than cutting-edge deep learning and hybrid approaches** in terms of scalability and real-world application—particularly in image-rich or multi-modal data environments.

3. LIMITATION AND MOTIVATION

The prediction models proposed earlier for liver disease are based on statistical analysis and machine learning. Still, it should not be considered a substitute for professional prediction models for medical data diagnosis. A screening tool is used to help healthcare professionals detect fatty liver disease, but they need more accuracy and future help in evaluation. The output depended on the data quality and number of features used for disease prediction. Most models cannot provide good accuracy due to the impurity of raw data. This means the raw data has missing elements, wrong data, mismatched data, and others that affect the prediction accuracy and reliability of the performance. Also, the training process is inaccurate because the novel factors are unavailable in the past historical dataset used for training the model. Most models do not find the correlation between individual features and liver diseases since each feature may cause the disease individually. Hence, this paper motivated to predict fatty liver disease by analyzing the performance parameters concerning individual and combined features, like IB, QF, HF, IB+QF, IB+HF, QF+HF, and IB+QF+HF using the SVM model.

The Following are the Realtime Considerations for the Given Work.

1. **Limited Dataset Size**: The study uses data from only **181 patients**, which may not be sufficient for building a generalized or robust predictive model, especially when applied to diverse or larger populations.
2. **Lack of External Validation**: The model was evaluated only on the internal dataset. Without external validation on independent or multi-center datasets, its real-world applicability remains uncertain.
3. **Feature Scope Constraints**: While the model uses features like IB, Q-factor, and HF, it may overlook other influential clinical, genetic, or lifestyle factors that contribute to NAFLD progression.
4. **Model Selection Simplicity**: Although SVM is efficient, it may not capture complex non-linear relationships as effectively as more advanced deep learning models (e.g., CNNs or ensemble techniques).
5. **Imaging Quality and Variability**: Ultrasound-derived features are subject to image quality, operator variability, and machine differences, potentially affecting model consistency across settings.
6. **No Real-Time Clinical Integration**: The abstract doesn't address how the model can be integrated into clinical workflows or decision-support systems, limiting its immediate clinical utility.
7. **Potential Overfitting Risk**: With high accuracy reported on a small dataset and no mention of robust cross-validation strategies (e.g., k-fold CV or bootstrapping), there is a risk of overfitting.

These limitations suggest that while the study is a promising step toward early NAFLD detection using ML, further work is needed to validate, expand, and clinically integrate the model.

4. PROBLEM STATEMENT

Non-Alcoholic Fatty Liver Disease (NAFLD) is a growing health issue worldwide, affecting more people. Detecting and predicting the severity of NAFLD, especially its advanced form called non-alcoholic steatohepatitis (NASH), is especially to prevent serious complications like liver failure and cirrhosis. However, traditional methods like imaging are not always accurate or easily available. Because of this, there is a need for a better way to predict these liver diseases using machine learning.

This study focuses on developing a predictive model using the Support Vector Machine (SVM) algorithm. The model will analyze data from ultrasound scans of liver tissue, looking at factors like fat distribution using integrated backscatter (IB), Q-factor, and homogeneity factor (HF), along with clinical information from patients. The challenge is managing large amounts of data, converting it into useful features, and making sure the model gives accurate results.

The SVM algorithm works by finding the best way to separate several types of data using a boundary called a hyperplane. In this case, it will use both clinical and tissue-related data to classify liver conditions like NAFLD and NASH. The goal is to create a reliable system that helps in early diagnosis and better treatment decisions.

The SVM constructs a hyperplane (in a higher-dimensional space if necessary) that maximizes the margin between the two classes. The objective is to find the hyperplane that minimizes classification error and maximizes the distance (margin) between the classes. This is given by the equation:

$$w \cdot x + b = 0$$

Where, w is the weight vector, x is the feature vector, and b is the bias term.

The SVM solves the optimization problem to maximize the margin while minimizing classification errors. The optimization problem can be formulated as:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

Subject to the constraint that for all training samples:

$$y_i(w \cdot x_i + b) \geq 1, \qquad i = 1,2,\ldots,N$$

where y_i is the class label of the i-th sample (1 or -1), and x_i is the feature vector for the i-th sample.

Since the data may not always be linearly separable, the kernel trick is used to map the data into a higher-dimensional space where a linear hyperplane can be found. The kernel function $K(x_i,x_j)$ computes the dot product of transformed feature vectors without explicitly transforming the data. Common kernel functions include:

$$K(x_i, x_j) = (x_i \cdot x_j + c)^d$$

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right)$$

The goal is to minimize the following objective function for SVM with a kernel:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

where ξ_i is the slack variable for each data point, and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error.

The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score. The accuracy A of the model is defined as:

$$A = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions}$$

The model achieves an average accuracy of 89.95% in predicting NAFLD and NASH, as confirmed through the experiment.

5. METHODS AND MATERIALS

The dataset is taken from pathological information of ultrasonic tissues of fatty liver diagnosis collected from patients in China. It is available in GitHub storage under the name "CGMH_gastro.xlsx." It comprises 1765 rows and 35 columns of data elements about the patients. This paper trains and tests the input data using the Python 3.6.2 version. The overall training and testing process is performed in four steps, as discussed below.

Step 1: Initially, the input data are classified as A and B and loaded for training and testing.

Step 2: The ultrasonic characteristics of the train data parameter are set with intervals [0 and 1]. Based on the scaled value of training data, the testing data scaled value is generated to avoid large values in the resultant.

Step 3: C and Y are the parameters used to define the relationship between the noise tolerance and maximum margin and adjust the proposed model's complexity, respectively. Based on the parameter space value {C = 2^−5, 2^−3, …, 2^15, γ = 2^−15, 2^−13, …, 2^3, degree = 1, 2, 3}, the SVM model is trained through LeaveOneOut (LOO) cross-validation technique with the help of optimal value generated by the grid search method.

Step 4: The ROC and confusion matrix results are evaluated and plotted using test data to predict the final analysis result. This will increase the model's diagnostic accuracy.

Excess fat accumulates and causes fatty liver disease, which should be controlled and removed through physical exercises, a good health diet, and other doctor-advised home remedies; it is prevalent in terminal liver diseases like cancer. Medical industries all over the world consider pathology to be a standard way of diagnosing fatty livers in the earlier days. However, non-invasive medical image processing methods are used instead of invasive methods because of their side effects and controversies. People prefer ultrasound for diagnosis due to safety, convenience, and price. While using ultrasound images, some limitations were found, such as some of the ultrasonic parameters not being opted for in many circumstances. Based on this, some parameters are extracted from the original ultrasound signal, representing the physical and tissue characteristics that help diagnose fatty liver diseases.

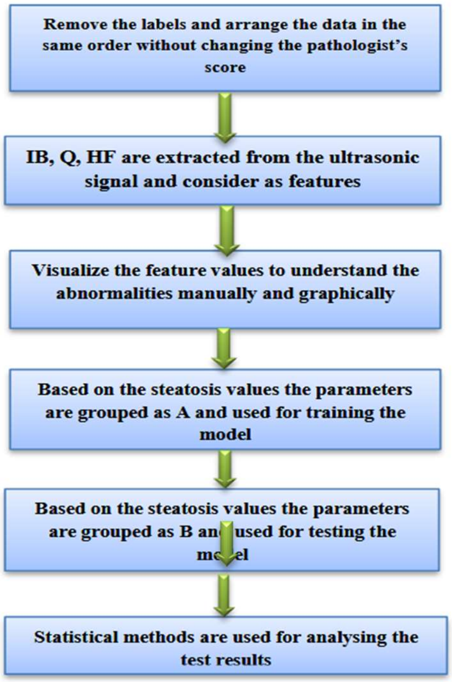

Remove the labels and arrange the data in the same order without changing the pathologist's score
IB, Q, HF are extracted from the ultrasonic signal and consider as features
Visualize the feature values to understand the abnormalities manually and graphically
Based on the steatosis values the parameters are grouped as A and used for training the model
Based on the steatosis values the parameters are grouped as B and used for testing the model
Statistical methods are used for analysing the test results

Figure-3. Proposed Pipeline Data Analytics Process

The Integrated backscatter (IB), the Q factor of the Hilbert-Huang transition (Q factor), and the Homogeneity factor (HF). The backscatter signal intensity is measured by IB, the frequency decay by Q, and the fat evenness is by HF. Analyzing single parameter and predicting NAFLD cannot provide accuracy and thus, the SVM algorithm is used for analyzing each and combination of all the parameters called features which overcomes the limitations and increase the true positive rate. From the dataset, two groups are created as A having 111 samples, and B having 75 samples and that are used for training and testing the SVM algorithms. Finally, a ground truth data with 10% steatosis is used to confirm the FLD.

From the experimental output, it is found that the parameters extracted can be able to determine the FLD with their respective performances. For improving the other performances than sensitivity, all the features are combined and used in the analysis. The accuracy of predicting fatty patients is 86.49% and ROC is 89.29%. Analysing the combination of parameters can increase the versatility and accuracy of the SVM model and decrease the computational complexity. And it is proved that SVM is highly suitable and potential in fatty liver diagnosis. The overall data analysis process is illustrated in Figure-3. This paper's dataset is the pathological information obtained from ultrasound images and DNA data sequences for fatty liver patients. Several machine learning algorithms are available for medical data analytics and can be applied in the real-time medical industry concerning the nature of the dataset and its complexities. The Support Vector Machine algorithms can perform well with reduced computational and time complexities compared to all machine learning algorithms. Only the hyperplane in the search space can be increased based on the constraints. Thus, this paper implements a Support Vector Machine (SVM) algorithm for analyzing the pathological information of fatty liver disease patients' data. SVM can do processes using different mathematical models built into it. For example, it used 2D point classification using vector representation. For example, $\overrightarrow{OA}$ represents a vector that connects points O to A. The distance between two points, called norms, and the distance of the vector can be obtained using the formula:

$x(x_1, x_2, x_3)$ is expressed as $\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$

At the same time, the data moving direction is called vector direction and is obtained by the formula:

$x(x_1, x_2, x_3)$ is given as $\{\frac{x_1}{||x||}, \frac{x_2}{||x||}, \frac{x_3}{||x||}\}$

Two vectors (data points) can be extended into the same directions as represented as

$$u \,.\, v = |u||v|\cos(\theta) = x_1 \times x_1 + y_1 \times y_2$$

u & v are vectors, their dot product is evaluated as: (.) denotes the inner product, and θ represents the angle between u and v. SVM uses hyperplanes to classify the data in the search space. It is called a hyper-line in 2D space, and it is called a 3D hyperplane. It divides the data into two classes. The data points x are represented using SVM is y, and it is expressed as:

$$y = a * x + b$$

$$a * x + b - y = 0$$

Given Vector $X = (x, y)$ and $W = (a, -1)$ Hence, the vector equation of the hyperplane is $W.X + b = 0$. If the data points are not linear, they can be separated by

$$Z = X^2 + Y^2$$

This ensures that a linear classifier can be applied to the data points.

SVM is one of the robust machine learning models that can perform regression, outlier detection, and linear and non-linear classification over any alpha-numeric dataset. It is used in various applications like the classification of text and images, pattern recognition, gene expression analysis, etc. It is adaptable and practical in multiple applications since it manages nonlinear association and high-dimensional data using a customized number of hyperplanes to differentiate the classes based on the features. It is a supervised machine-learning model used for classification and regression processes. The novelty of SVM is to obtain optimal hyperplanes for differentiating data points in high-dimensional data. Each hyperplane is drawn between data points, and they try to get various classes based on the closest point as much as possible. Figure-4 shows the three different hyperplanes used for classifying the data points based on the distance values d1, d2, and d3. The closest data points are selected based on the closest distance-based data points.

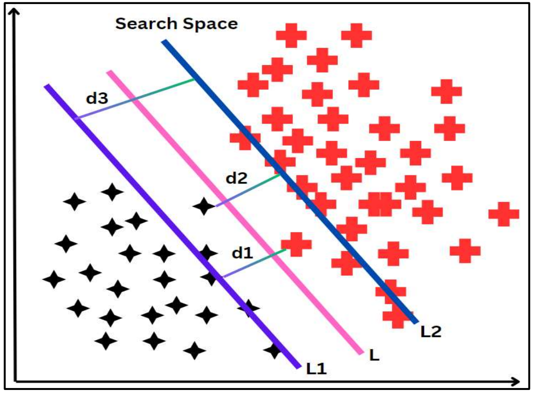


Figure-4. Support Vector Machine

In the SVM model, the input feature vector X and class label Y are used to train the dataset and to perform binary classification with two labeled classes, +1 and -1. The hyperplane of the SVM model is evaluated using the equation.

$$\boldsymbol{w^T x + b = 0} \qquad \textbf{(1)}$$

Here, W and b represented the average vector of the hyperplane and the distance between the hyperplane from the average vector to the origin. Then, using Equation (2), the distance between the decision boundary and the data point is evaluated.

$$\boldsymbol{d_i = \frac{w^T x_i + b}{\|w\|}} \qquad \textbf{(2)}$$

Where $||w||$ defines the Euclidean norm value of the normal weight vector W. The Euclidean norm value for the linear SVM model is expressed as:

$$\hat{y} = \begin{cases} \mathbf{1}: \boldsymbol{w^T x + b \geq 0} \\ \mathbf{2}: \boldsymbol{w^T x + b < 0} \end{cases}$$

The following expression is then applied to find the optimization result of the soft margin linear SVM model.

$$\underset{\omega,b}{\textbf{minimize}} \tfrac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\zeta\, i \textbf{ subject to } y_i(\omega^T x_i + \mathrm{b}) \geq 1 - \zeta_i \textbf{ and } \zeta_i \geq 0 \textbf{ for i} = 1,2,3, \ldots, \mathrm{m}$$

the optimized output obtained from the SVM is given in Y. The overall process of data analytics is shown in Figure-3.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed preprocessing method is experimented with Inter Pentium Core i7, 7th generation, 1TB HDD, 16GB RAM, and 2.36GHz processor speed. Python is installed with all essential libraries to enable the artificial intelligence algorithms with Kera's model. To improve the diagnostic accuracy of the model, the input data are pre-processed using the filter. The data before and after applying the pre-processing technique is visualized in Figure 1. The X and Y axes represent the Integrated backscatter and homogeneity factors, respectively. Figure-4 (a) shows the scatter plot graph depicting the raw input data before preprocessing. Figure-4 (b) shows the after-pre-processing scatter plot result. The input FLD data are classified into four categories: Normal, Mild, Moderate, and Severe. It is classified based on the fat level present in the liver. If analyzed, fat levels <5%, 5-33%, 33-66%, or >66% are termed Normal, Mild, Moderate, or Severe, respectively. These different stages of the disease data in the input dataset are classified using various factors.

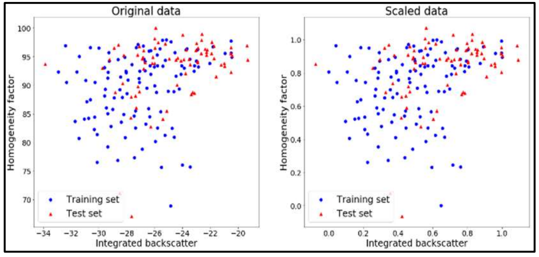


Figure-5. Before and After Preprocessing Data Data Visualization

The experiment is carried out by analyzing the best parameters as features, such as IB, HHT, and HF, to detect abnormalities in the fatty liver condition. The accuracy obtained from the experiment is evaluated by comparing the accuracy using individual features and combined features. Initially, IB, HHT, and HF factors are used individually to compute the sensitivity, specificity, and accuracy. Then, two eatures, IB and HHT, IB and HF, and vice versa, will be combined to verify the efficiency of sensitivity, specificity, and accuracy. For example, Figure-5 illustrates the AUROC estimated concerning the individual and combined features. It shows that the AUROC value is high for all three features combined to identify fatty liver diseases. The overall AUROC values obtained using all the features and their combination are received within the range of 79.27% to 89.29%. Figure-6 illustrates the accuracy estimated concerning the individual and combined features. It shows that the accuracy value is high for all three features combined to identify fatty liver diseases. The overall accuracy values obtained using all three features and their combination are 68.92% to 86.49%.

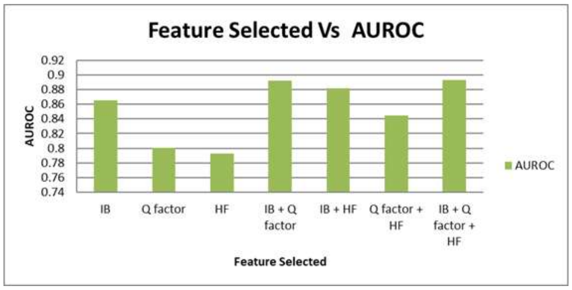


Figure-6. AUROC For Individual and Combination of Features

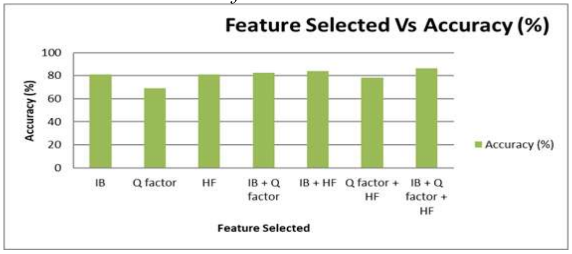


Figure-7. Accuracy For Individual and Combination of Features

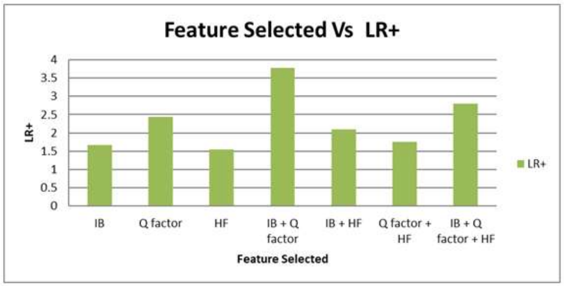


Figure-8. LR+ For Individual and Combination of Features

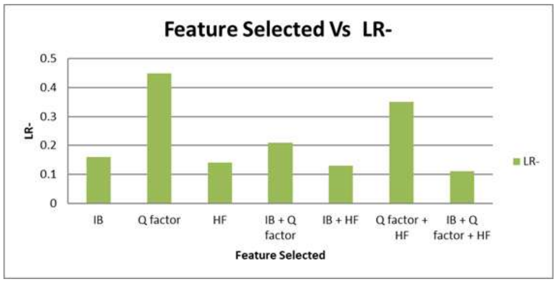


Figure-8a. LR For Individual and Combination of Features

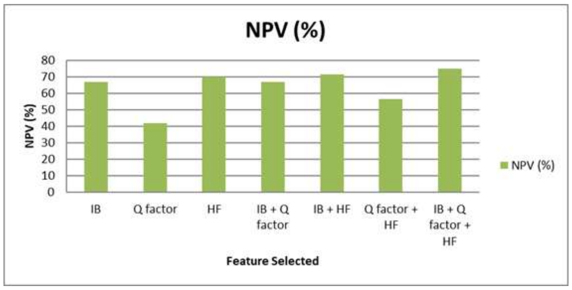


Figure-9. NPV For Individual and Combination of Features

Figure-7 illustrates the positive likelihood ratio (LR+) estimated concerning the individual and combined features. It shows that the positive likelihood ratio value (37.8%) for connecting features IB and Q-Factor is high in identifying fatty liver diseases. The overall positive likelihood ratio values obtained using all three features and their combination are 15.5% to 37.8%. Similarly, the negative likelihood ratio (45%) is high for only the Q-Factor feature in identifying fatty liver diseases. The overall negative likelihood ratio values obtained using all three features and their combination are 11% to 45%, as shown in Figure-8. The NPV (Negatively Predicted Value) estimated concerning the individual and combined features is shown in Figure-9. The highest NPV value (75%) is obtained by combining all three features: IB, QF, and HF. The overall negative likelihood ratio values obtained using all three features and their combination are 41.94% to 75%. Similarly, the PPV (Positively Predicted Value) is estimated concerning the individual and combined features, as shown in Figure-10. The highest PPV value (92.16%) is obtained by combining two features, IB and QF. The overall PPV obtained using all three features and their combination is 82.81% to 92.16%.

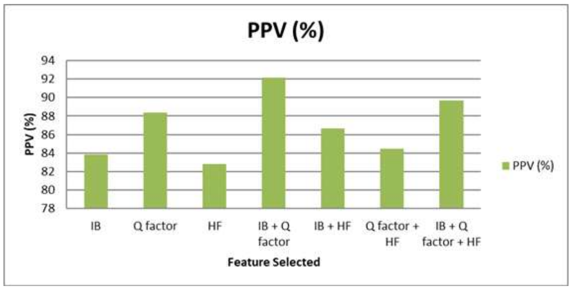
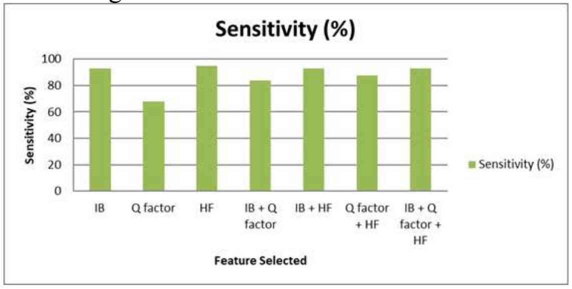


Figure-10. PPV For Individual and Combination of Features

The sensitivity is calculated based on individual and combined features to verify the efficiency of FLD detection. Figure-11 shows the highest sensitivity, 94.64%, obtained only by analyzing the dataset concerning HF features for FLD detection.



igure-11. Sensitivity For Individual and Combination of Features

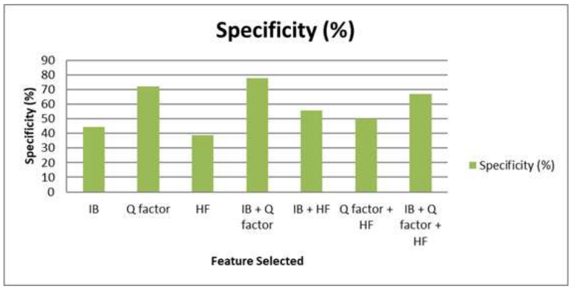


Figure-12. Specificity For Individual and Combination of Features

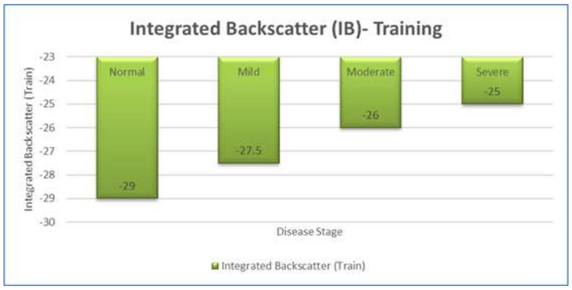


Figure-13(a). Disease Stage Classification w.r.t IB in Training

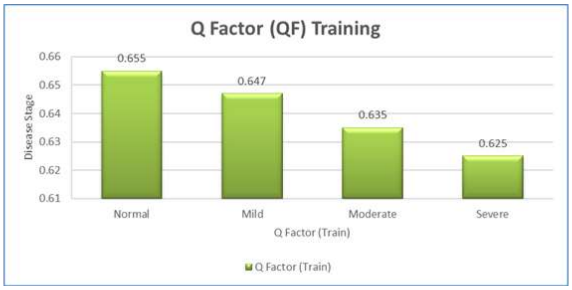


Figure-13(b). Disease Stage Classification w.r.t QF in Training

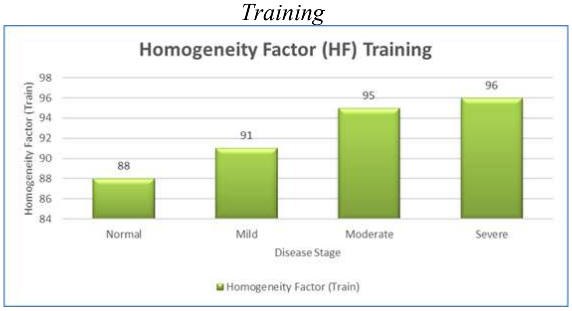


Figure-13. Disease Stage Classification w.r.t HF in Training

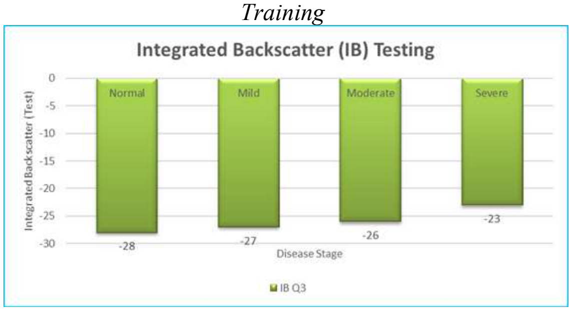


*Figure-1***4(a). Disease Stage Classification w.r.t** *IB in Testing*

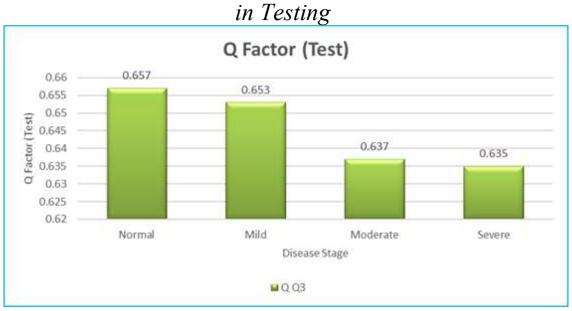


Figure-14(b). Disease Stage Classification w.r.t IB in Testing

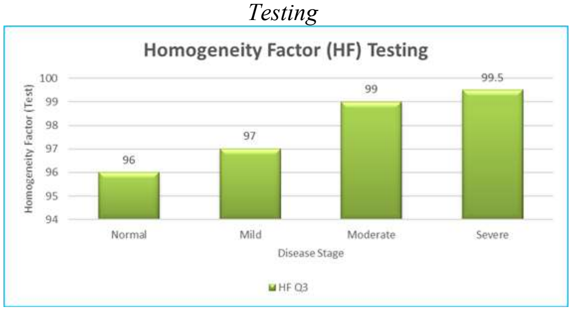


Figure-14(c). Disease Stage Classification w.r.t HF in Testing

The overall sensitivity range obtained in the experiment is from 67.86% to 94.64%. The specificity is calculated based on individual and combined features to verify the efficiency of FLD detection. Figure-12 shows the highest sensitivity, 77.78%, obtained by analyzing the dataset concerning IB and Q-Factor features for FLD detection. The overall specificity range obtained in the experiment is 38.89% to 77.78%. Figure-13 (a), (b), and (c) show the classification result of the input trained data through the IB factor, HTT Q factor, and Homogeneity factor, respectively. The classification result of the IB factor depicts that the input dataset has more severe data than the others. The HTT Q factor classified a maximum number of data as normal data. The homogeneity factor results show that most of the dataset's input are predicted and classified as moderate stages. Figure-14 (a), (b), and (c) depict the testing data classification result using three different factors, namely the IB factor, HTT Q factor, and Homogeneity factor, respectively. The classification result of the IB factor on testing data also classified most of the data as severe stage. The HTT Q factor classified a maximum number of data as normal data. The homogeneity factor results show that most of the dataset's input are predicted and classified as severe stages.

7. CONCLUSION

Predicting abnormal genes in metagenomic fragments remains a significant and challenging task in the field of metagenomics, often yielding limited accuracy. In this study, a Support Vector Machine (SVM) algorithm is applied to analyse pathological data derived from ultrasound imaging of fatty liver patients in China. After preprocessing the data, the SVM model is used to classify disease stages—normal, mild, moderate, and severe—based on quantitative tissue parameters such as integrated backscatter (IB), Q-factor, and homogeneity factor (HF), which reflect fat distribution and tissue behaviour. The SVM model demonstrates its versatility through regression, outlier detection, and both linear and non-linear classification on high-dimensional data. Implemented using Python, the model achieved an average accuracy of 86.49% across the clinical and imaging datasets. Future work will explore the application of SVM to additional pathological and genomic datasets and compare its performance with other machine learning algorithms to assess its robustness and generalizability.

REFERENCES

[1]. Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J. (2019). Prediction of fatty liver disease using machine learning algorithms. *Computer methods and programs in biomedicine*, *170*, 23-29.

[2]. Islam, M. M., Wu, C. C., Poly, T. N., Yang, H. C., & Li, Y. C. J. (2018). Applications of machine learning in fatty live disease prediction. In *Building continents of knowledge in oceans of data: the future of co-created eHealth* (pp. 166-170). IOS Press.

[3]. Reddy, D. S., Bharath, R., & Rajalakshmi, P. (2018, September). A novel computer-aided diagnosis framework using deep learning for classification of fatty liver disease in ultrasound imaging. In *2018 IEEE 20th international conference on e-health networking, applications and services (Healthcom)* (pp. 1-5). IEEE.

[4]. Cao, W., An, X., Cong, L., Lyu, C., Zhou, Q., & Guo, R. (2020). Application of deep learning in quantitative analysis of 2-dimensional ultrasound imaging of nonalcoholic fatty liver disease. *Journal of Ultrasound in Medicine*, *39*(1), 51-59.

[5]. Sorino, P., Caruso, M. G., Misciagna, G., Bonfiglio, C., Campanella, A., Mirizzi, A., ... & MICOL Group. (2020). Selecting the best machine learning algorithm to support the diagnosis of Non-Alcoholic Fatty Liver Disease: A meta learner study. PLoS One, 15(10), e0240867.

[6]. Gaber, A., Youness, H. A., Hamdy, A., Abdelaal, H. M., & Hassan, A. M. (2022). Automatic classification of fatty liver disease based on supervised learning and genetic algorithm. Applied Sciences, 12(1), 521.

[7]. Zhao, M., Song, C., Luo, T., Huang, T., & Lin, S. (2021). The fatty liver disease prediction model is based on big data from electronic physical examination records. Frontiers in Public Health, 9, 668351.

[8]. Singh, A., Nath, P., Singhal, V., Anand, D., Verma, S., & Hong, T. P. (2020). A new clinical spectrum for assessing nonalcoholic fatty liver disease using intelligent methods. IEEE Access, 8, 138470-138480.

[9]. Poonguzharselvi, B., Ashraf, M. M. A., Subhash, V. V., & Karunakaran, S. (2021). Prediction of liver disease using machine learning algorithm and genetic algorithm. Annals of the Romanian Society for Cell Biology, 2347-2357.

[10]. Kang, B. E., Park, A., Yang, H., Jo, Y., Oh, T. G., Jeong, S. M., ... & Ryu, D. (2022). Machine learning-derived gut microbiome signature predicts fatty liver disease in the presence of insulin resistance. Scientific Reports, 12(1), 21842.

[11]. https://www.pacehospital.com/fatty-liver-symptoms-grade-causes-complications-risk-factors

[12]. Singh, S., Osna, N. A., & Kharbanda, K. K. (2017). Treatment options for alcoholic and non-alcoholic fatty liver disease: A review. World journal of gastroenterology, 23(36), 6549.

[13]. https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/liver-fatty-liver-disease

[14]. Topal, E., Aydemir, K., Çağlar, Ö., Arda, B., Kayabaşı, O., Yıldız, M., ... & Erbaş, O. (2021). Fatty liver disease: Diagnosis and treatment. Journal of Experimental and Basic Medical Sciences, 2(3), 343-357.

[15]. https://my.clevelandclinic.org/health/diseases/15831-fatty-liver-disease

[16]. https://onlinelibrary.wiley.com/doi/full/10.1111/liv.15004

[17]. https://www.frontiersin.org/articles/10.3389/fpubh.2022.909455/full

[18]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8185553/

[19]. https://www.frontiersin.org/articles/10.3389/fimmu.2020.609900/full

[20]. https://jamanetwork.com/journals/jama/fullarticle/2754794

[21]. https://jamanetwork.com/journals/jama/fullarticle/2754794

[22]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10029948/

[23]. https://www.researchgate.net/publication/235777763_An_Automated_Diagnosis_System_of_Liver_Disease_using_Artificial_Immune_and_Genetic_Algorithms

[24]. https://www.mdpi.com/2073-8994/11/1/33

[25]. https://www.sciencedirect.com/science/article/pii/S2666379121002998