IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

Machine Learning Insights into Non-Alcoholic Fatty Liver Disease Prediction

Dr.V. Surya Narayana[1], S. Varun Sai Srinivas[2], M. Sai Ram[3], S. Mahesh[4]
[1]Department of Artificial Intelligence and Data Science, Professor of Engineering, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, AP, India
[2]Department of Artificial Intelligence and Data Science,Project Student, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, AP, India
[3]Department of Artificial Intelligence and Data Science, Project Student, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, AP, India
[4]Department of Artificial Intelligence and Data Science, Project Student, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, AP, India

ABSTRACT

This research addresses a prominent concern in contemporary society, non-alcoholic fatty liver disease (NAFLD) arises from an accumulation of fat in the body, posing a substantial health challenge. In essence, obesity-related liver illness is more harmful since it shortens people's lives. The liver releases fats, such as triglycerides and hyperlipidemia. These are a few of the lipids that the liver has evolved for the bloodstream. The liver may be impacted by the slowing down of blood flow that occurs when these fats are consumed in excess. Inflammation and damage to the liver are possible outcomes if fat storage is found in the liver cells. Therefore, identifying liver diseases is essential to minimizing their detrimental effects. To create a good prediction model, a few machine learning methods are being studied the proposed CROSS VALIDATION and boosting techniques system is intended to treat non-alcoholic fatty liver disease. We thoroughly suggested the Randomized search CV and Grid search CV algorithms in addition to other widely used models. According to the experimental results, the suggested architecture generally improves the accuracy of the disease predictions, ensuring a high level of model resilience and robustness.

Keywords:

Hepatic System, Gradient Boosting, XG Boosting, Randomized search cv, Grid Search cv, Random Forest, Liver Illness Detection.

1. INTRODUCTION

The prevalent condition known as non-alcoholic fatty liver disease (NAFLD) is a widespread ailment that may progress to hepatic complications, including non-alcoholic steatohepatitis (NASH) and cirrhosis. The liver is essential for digestion, waste removal, and energy storage. Using random search CV and grid search CV for training data augmentation, a unique strategy combines random forest with boosting algorithms such as gradient boosting and XGBoost to improve disease identification. Higher accuracy results demonstrate that this approach performs better than typical machine learning models, offering reliable and adaptable illness diagnosis in practical settings. This article effectively presents a combined approach that successfully combines boosting algorithms and cross-validation approaches for real-time detection of non-alcoholic fatty liver disease using text data, demonstrating the technique's promise in real-world applications. According to experimental findings, boosting and cross-validation combined work better.

The paper underscores the growing significance of machine learning in the diagnosis of complex liver ailments, the paper "Prediction of Liver Disorders Using Machine Learning Algorithms: A Comparative Study" investigates ML techniques including LR, DT, RF, and ET for early diagnosis. Despite advances in machine learning, it highlights the importance of medical experience in managing complex data and liver problems [1].

With an emphasis on examining LR, DT, RF, and ET techniques using the Indian Liver Patient Dataset, the literature review emphasizes the growing significance of machine learning in the diagnosis of liver illnesses. It emphasizes how crucial medical knowledge is to advancing machine learning, particularly when managing complicated data and treating the complexities of liver diseases [2].

Exploring the realm of liver disease prediction through machine learning, the paper titled “Accuracy Prediction using Machine Learning Techniques for Indian Patient Liver Disease”. During the investigation outlined in this research paper, we explore the applications of Decision Tree, Naive Bayes, Support Vector Machine, Random Forest, and Artificial Neural Network methodologies. The objective is to enhance prediction accuracy and expedite diagnostics using datasets specific to liver diseases in the Indian population [3]. One possible drawback, though, is the set number of features per data point, which could introduce bias and affect the effectiveness of supervised learning algorithms in different clinical scenarios [3].

The study, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," employs the Naïve Bayes and K Nearest Neighbour (KNN) algorithms for the early diagnosis of liver illness in patients. It uses comprehensive patient data mining to forecast the incidence of liver disease. Evaluation findings, however, show inconsistent performance; for example, Naïve Bayes scores 72.5% whereas KNN scores 63.19% for Area Under Curve (AUC). One disadvantage of the study is that, because of the different performance levels of the algorithms, more optimization based on training data and critical factors is required [4].

The study uses the Indian Liver disease dataset and a variety of boosting algorithms and feature reduction strategies to investigate the early diagnosis of liver illness through machine learning. It highlights the potential bias produced by these methodologies while analysing the aspects and global impact of liver illnesses, underscoring the necessity of giving careful thought to guarantee the model's flexibility across clinical circumstances [5].

The research aims to enhance the accuracy of diabetes prediction in the medical industry using machine learning techniques. This study delves into cutting-edge computational techniques, exploring algorithms like Decision Trees, Support Vector Machines, and Neural Networks to enhance predictive capabilities. Previous research has identified several challenges, including data heterogeneity, model interpretability, and potential biases in training datasets [6]. The study proposes new techniques to overcome these constraints and achieve more accurate diabetes predictions. It highlights the significance of data quality, model interpretability, and bias reduction for successful real-world healthcare applications.

2. PROPOSED METHODOLOGY

2.1. Structural Framework

The structural blueprint outlines the proposed approach for constructing an effective model catering to Non-Alcoholic Fatty Liver Disease (NAFLD) functionalities. To initiate the process, an 80% percentage split and data pre-processing are employed, yielding distinct training and test datasets. Leveraging Python programming and relevant libraries, the five algorithms undergo implementation on the training dataset with notable instances. Subsequently, each algorithm is trained, and the acquired models are evaluated on the test dataset, encompassing the remaining instances. The outcome facilitates a comparative analysis of accuracy, precision, recall, confusion matrix, and f1-score and MCC for each algorithm. The determination of the most proficient algorithm is based on this comprehensive evaluation. Emphasizing our focus on machine learning algorithms manipulating "internal parameters," this work aims to discern the optimal algorithm through a meticulous comparison of performance metrics.

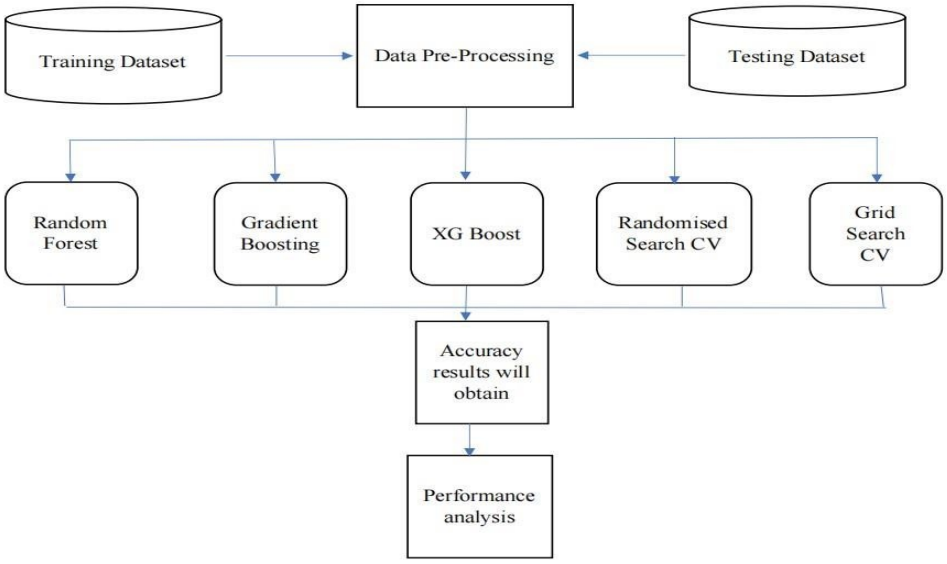

Fig 1: Structural Framework

2.2. Dataset Collection

This study aims to explore the predictive capabilities of machine learning algorithms in determining the presence of non-alcoholic fatty liver disease (NAFLD) based on clinical and demographic factors. The latent progression of undiagnosed NAFLD can lead to exacerbation. Enhancing patient outcomes necessitates timely identification and proactive intervention. The data collection for the research project comes from Kaggle, where there are initially 11 columns and 483 rows. Oversampling techniques are utilized to rectify the imbalanced data, culminating in an enlarged dataset consisting of 1056 rows. Training more robust models will be aided by this expanded dataset, which includes a balanced representation of NAFLD cases.

2.3. Dataset Description

The Dataset contained 10 distinct qualities of 1057 patients. We have 753 medical records with liver-related conditions and 303 records with non-liver health conditions in our dataset. Gathered from the northeastern area of Andhra Pradesh, India, the dataset uses the class label 'Dataset' to differentiate between individuals with liver disease and those without it. There are 263 patient records and 793 patient records for male patients in the dataset. Patients were assigned a score of 1 or 2 according to the condition of their livers. The attached table provides an extensive summary of the dataset and provides insights into a variety of traits and characteristics. This special dataset was used to evaluate prediction algorithms in an effort to reduce the workload for physicians. Liver datasets are a specific type of text data that requires specific features to be extracted for effective analysis. some of the main features are described in the below:
The dataset includes gender (male or female) and age (without restrictions), which are critical patient characteristics needed to predict liver disease. It includes a number of blood indicators, including Alkaline Phosphatase (ALP) levels between 30 and 120 IU/L, which may indicate liver damage at higher values, Total Bilirubin (TBIL) levels between 0.1 and 2.0 mg/dL, Direct Bilirubin (DBIL) levels between 0 and 1 mg/dL, and so on. Furthermore, liver cell injury may be indicated by the enzymes aspartate amino transferase (AST), which ranges from 4 to 40 IU/L, and aspartineaminotransferase (ALT), which ranges from 4 to 45 IU/L. Additionally taken into account are total proteins between 6.0 and 8.5 g/dL and albumin between 3.4 and 5.4 g/dL. Generally, the Albumin to Globulin Ratio (A/G) ranges from 1 to 2. The Class parameter of the dataset indicates the existence.

2.4. Algorithms Used

2.4.1. Random forest:

The Random Forest algorithm is a popular choice for classification and regression applications. The outputs of multiple decision trees are combined in this ensemble learning technique to produce predictions. Being a supervised learning algorithm, random forest obtains the highest accuracy results when learning from labelled training data.

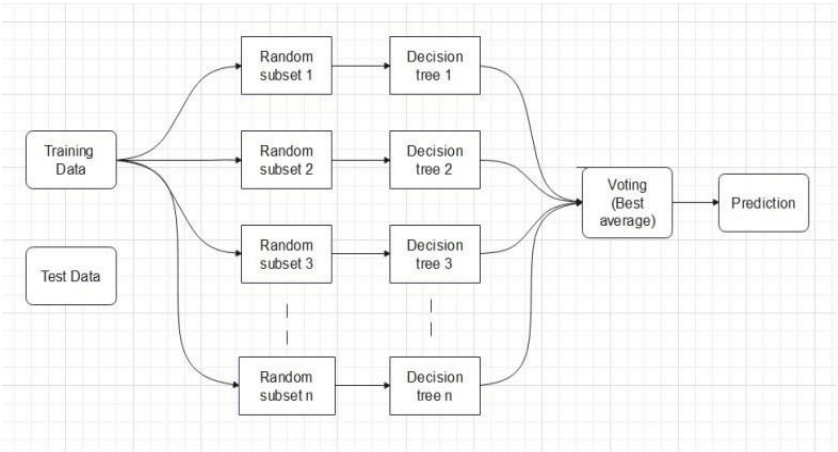

Fig 2: Random Forest Architecture

2.4.2. Gradient Boosting:

Gradient boosting is a widely used supervised learning algorithm that works especially well for regression and classification tasks. Unlike the Random Forest method, which builds multiple decision trees at once, this ensemble algorithm builds decision trees one after the other in order to gradually reduce residual errors from the previous tree. In gradient boosting, the resultant prediction model creates an ensemble of weak prediction models, which are usually represented as decision trees.

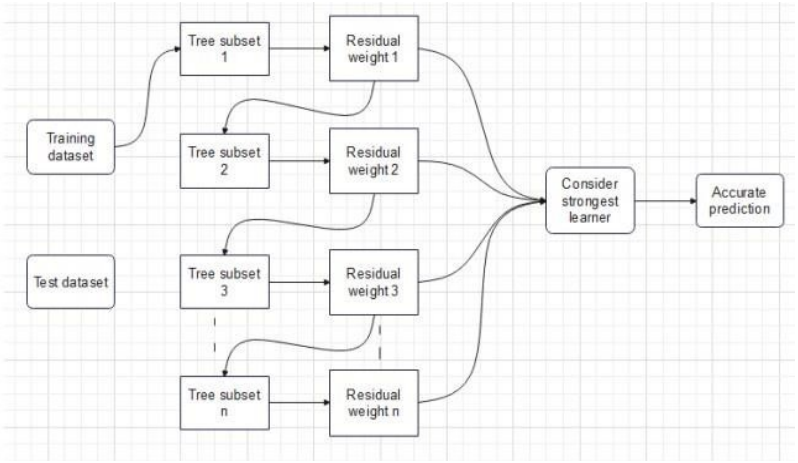

Fig 3: Gradient Boosting Architecture

2.4.3. XG Boosting:

Within the gradient boosting framework, the machine learning algorithm XGBoost uses decision trees. XGBoost is a distributed gradient boosting library that has gained recognition for its exceptional efficiency, versatility, and portability. Said to be a precise and effective data science solution, it performs exceptionally well when handling objective functions such as regression, ranking, and classification. Its superior efficiency over other frameworks is ascribed to sophisticated algorithms and a more refined model formalization.

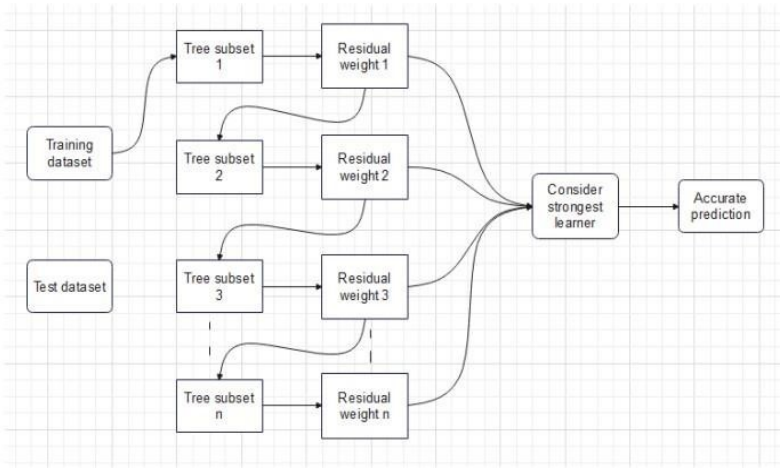

Fig 4: XG Boost Architecture

2.4.4. Randomized Search CV:

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

Randomized Search CV is an innovative approach to hyperparameter tuning, seamlessly integrating the merits of cross-validation and a randomized search strategy. This technique is employed to pinpoint the most effective combination of hyperparameters for a given machine-learning model.

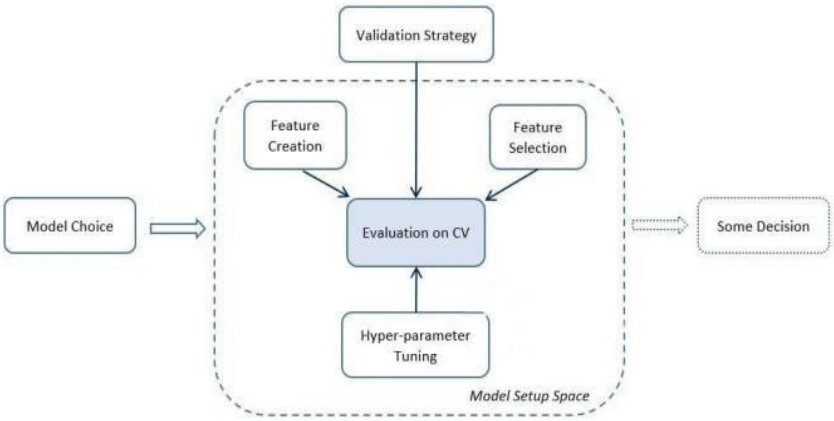

Fig 5. Randomized Seach CV Architecture

2.4.5. Grid Search CV:

Grid Search Cross Validation, or Grid Search CV, is a method for optimizing machine learning model hyperparameters. To find the optimal set of hyperparameters that maximizes the model's performance on a given dataset, Grid Search CV essentially employs a brute-force approach that thoroughly explores the hyperparameter space**.**

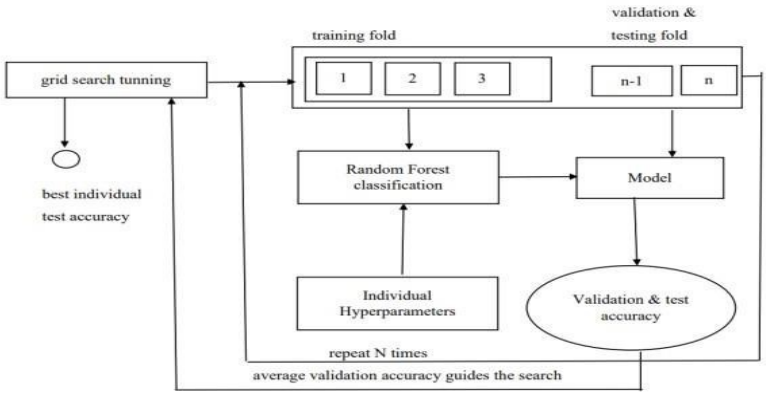

Fig 6. Grid Search CV Architecture

2.4. Implementation

- **Gathering Data:** This study used a dataset that was downloaded from Kaggle, a well-known source of machine learning datasets, and contained 483 cases and 11 pertinent columns. The dataset was expanded to 1057 rows while preserving the original features through the use of an oversampling technique, which strengthened the model's resilience. In addition to total and direct bilirubin, ALP, ALT, ASP, total proteins, albumin, albumin ratio, and globulin, age and gender are significant biochemical and demographic factors. The "Dataset" column is the target variable to be taken into account when predicting non-alcoholic fatty liver disease.
- **Data pre-processing**: During the pre-processing stage, we first split the dataset into independent and dependent features using an indexing method. Training and testing data were subsequently separated from the independent and dependent data. Remember that there are always more training data available than testing data (use 25% of the test data and 75% of the training data). Next, use feature scaling to

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

standardize the dataset's independent instances. Feature scaling is not necessary for pre-processing when using ensemble models.

- **Model Training:** This refers to the data that has been used to train the machine learning system. An important factor affecting the outcome of illness prediction is the alignment of the input data sets and the corresponding sample output data. The independent and dependent training data will be used to train the model. At the moment, the model yields the most accurate result when given full access to the train data. Nevertheless, this data is not appropriate for our analysis of the final results.
- **Model Testing:** The evaluation of a fully trained model's performance on a test dataset is referred to as model testing within the realm of machine learning. To find out how well the model works, we must test it. We now take into account the test data, both independent and dependent. We use this test set of data to apply the model, and we can evaluate the model's accuracy by comparing its output to the results of the training set.
- **Evaluating the model:** Building the model with training data and testing it later with test data to see how well it performed. Based on the predictions we get and the initial values of the test data, we will evaluate the model using evaluation metrics like confusion matrix, accuracy, precision, and recall. Furthermore, the amount of data used for preparation has a big influence on how accurate the information is. As the train size increases, the precision gets better. The accuracy of each model is evaluated, and the model that yields the best results is selected to indicate the presence or absence of liver disease.

2.5. Confusion Matrix

A discrepancy matrix is a tabular representation that contrasts the genuine and anticipated classifications, revealing the efficacy of a classification algorithm. It enables assessment metrics such as precision, recall, and accuracy and gives a summary of a model's performance on unknown data. It is essential for machine learning predictive analysis because it enables comparison, quantitative evaluation of hyperparameters, and the identification of subpar predictions that need more investigation.

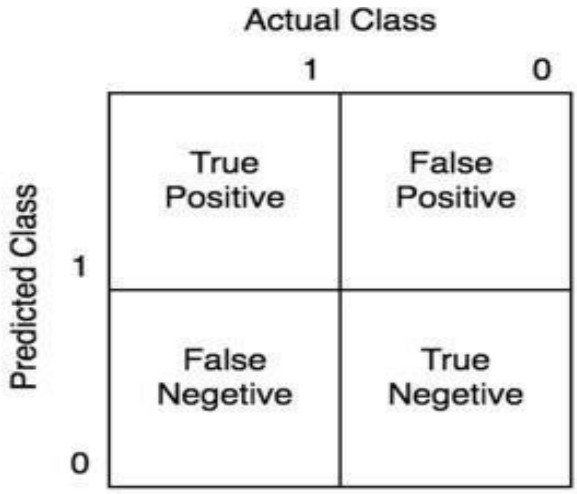

Fig 7: Confusion Matrix

Positive Correct (PC): The genuine value is positive, and the model accurately predicts it as positive.
Positive Incorrect (PI): The genuine value is negative, yet the model incorrectly predicts it as positive.
Negative Incorrect (NI): The genuine value is positive, but the model wrongly predicts it as negative.
Negative Correct (NC): The genuine value is negative, and the model accurately predicts it as negative.

Accuracy: Accuracy is one of the key metrics that can be ascertained from the confusion matrix. It shows the overall effectiveness of the classifier.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

Precision: An additional significant metric that can be obtained from the confusion matrix is precision. It gauges a classifier's capacity to correctly identify instances of one class without labeling them as belonging to another.

$$Precision = \frac{TP}{TP+FP}$$

Recall: A key performance indicator for classification models is recall, particularly when dealing with unbalanced datasets.

$$Recall = \frac{TP}{TP+FN}$$

F1-score: One metric used to assess test accuracy is the F1 score. It is derived from the test results by taking into account recall and precision.

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

MCC: The Matthews Correlation Coefficient, or MCC, is widely regarded as one of the most accurate indicators of a classification model's effectiveness. This is mainly because it considers every scenario for a prediction, in contrast to any of the metrics discussed earlier. The MCC ranges from +1 to -1.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

3. RESULTS

The patient records dataset was pre-processed to balance classes using SMOTE and extract nine pertinent features. Subsequently, XGBoost, Random Forest, and Gradient Boosting models were trained using these features. The model's performance was optimized by hyperparameter tuning through GridSearchCV and RandomizedSearchCV. The most robust algorithm was found by using a variety of metrics, such as accuracy, precision, recall, F1 score, and MCC, to evaluate the model's performance on test data. This machine-learning workflow is a comprehensive approach to categorizing liver disease based on medical data.

Results Obtained:

Random Forest Evaluation Matrix: The correlation between true and predicted labels is displayed in the matrices provided here. Whereas column entries correspond to predicted labels, row entries correspond to true labels. Elements with diagonals indicate cases where the true and predicted labels coincide. Non-diagonal elements indicate misclassified observations. The predictions made by the classifier are shown in each column, and each row precisely reflects the true label. See the approved random forest system's confusion matrix in the attached picture.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

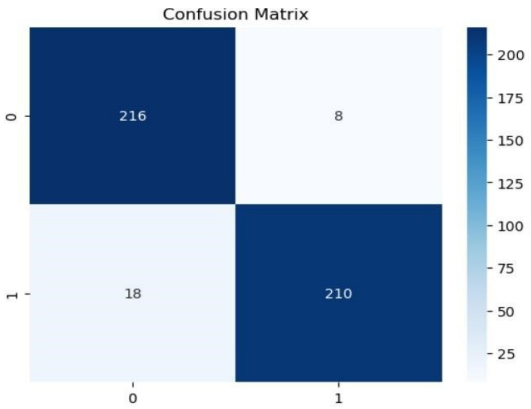

Fig 8: Random Forest Evaluation Matrix

Gradient Boosting Evaluation Matrix: The expected labels are shown in the columns and the true labels are shown in the rows of the following diagrams. The frequency of agreement between the expected and actual labels is indicated by diagonal elements. The classifier has mislabeled some observations, as indicated by the remaining cells. The rows indicate the correct label, while the columns show the classifier's prediction. You can find the confusion matrix for gradient boosting in the figure below.

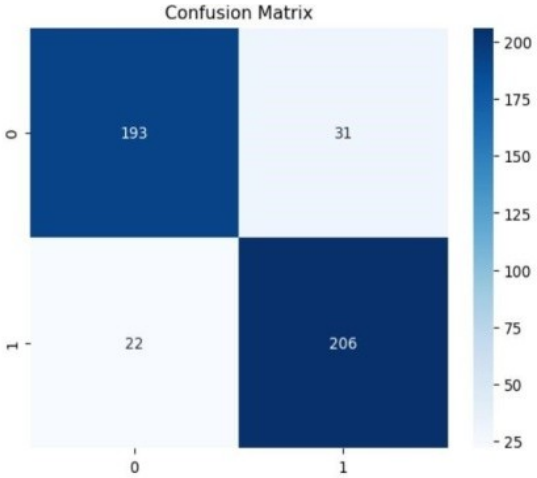

Fig 9: Gradient Boosting Evaluation Matrix

XG Boosting Evaluation Matrix: The number of correct predictions made by a classifier and the location of the classifier's confusion during an incorrect prediction can both be expressed using a confusion matrix. Actual labels are shown as rows and predicted labels as columns in the confusion matrices that follow. The values on the diagonal represent cases where the true label and the predicted label match. The classifier mislabeled observations in the other cells, as indicated by the values in those cells. The row indicates the correct label, while the column indicates the classifier's prediction. Below is a figure that displays the XG Boost Confusion Matrix.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

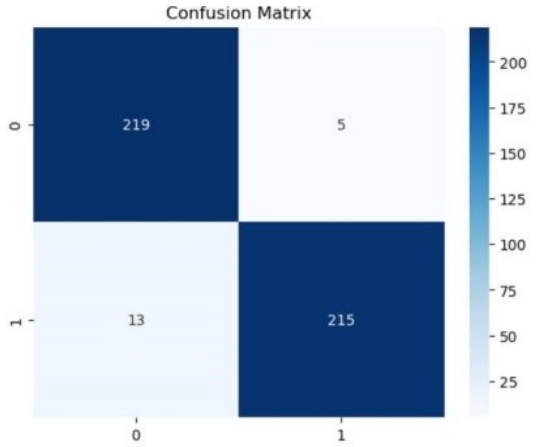

Fig 10: XG Boosting Evaluation Matrix

Randomized search CV Evaluation Matrix: The confusion matrices are presented below, where the true labels are displayed in the columns and the expected labels are depicted in the rows. Instances where the values align on the diagonal signify the number of instances where the actual and predicted labels match. The incorrect predictions made by the classifier are shown in the remaining cells; the correct label is shown in the row and the predicted label is shown in the column. The Confidence Matrix for the Randomized Search CV can be found in the figure below.

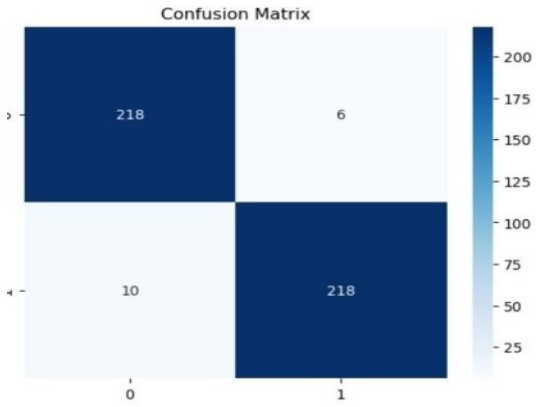

Fig 11: Randomized Search CV Evaluation Matrix

Grid Search CV Evaluation Matrix: Real labels are shown in rows and predicted labels are shown in columns in the matrices below. The diagonal values indicate how frequently the predicted and true labels match. The remaining cells show the classifier's incorrect labels for observations; the row represents the correct label, while the column shows the classifier's prediction. The figure below depicts the Grid Search CV Confusion Matrix.

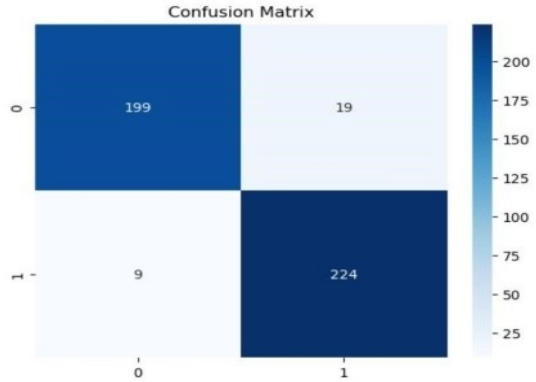

Fig 12: Grid Search CV Evaluation Matrix

S.NO	Model	Accuracy	Precision	Recall	F1-Score	MCC
1.	Random Forest	94.25	96	96	94	0.87
2.	Gradient Boosting	88.27	90	86	96	0.75
3.	XG Boosting	96.02	94	98	88	0.94
4.	Randomized Search CV	96.46	97	97	96	0.90
5.	Grid Search CV	93.79	96	91	93	0.86

Table 1: Performance Analysis of Models

4.CONCLUSION

To predict the development of non-alcoholic fatty liver disease (NAFLD), we recommend using machine learning algorithms such as Gradient Boosting, XG Boost, Randomized Search CV, Grid Search CV, and Random Forest. Gradient Boosting achieves an accuracy of 88%, precision of 90%, recall of 86%, F1-score of 88%, and MCC of 0.75. The system processes input instances smoothly, as supported by a detailed comparison. XG Boost, on the other hand, achieves impressive results with a recall of 98%, accuracy of 96%, precision of 94%, F1-score of 96%, and MCC of 0.94.

Additionally, there are some noteworthy performances from Random Forest, Grid Search CV, and Randomised Search CV. We give Randomized Search CV extra weight in our strategic selection process. The reason for this decision is that it is the best option for our predictive system because it can produce accurate results with less processing time. The efficiency and accuracy trade-off makes Randomized Search CV especially well-suited for cross-validation models, even though XG Boosting achieves the highest accuracy among the algorithms. This is a calculated move that maintains accuracy while satisfying the practical requirement for computing efficiency. The utilization of Randomized Search CV strategically improves overall system effectiveness and efficiency, and our system successfully predicts patient data with optimal accuracy. However it has the potential to receive a high score in the future given its present strengths and strategic choices. Its impact and relevance in the field of nonalcoholic fatty liver disease prediction can be further increased through ongoing development, validation on outside datasets, and practical application.

REFERENCES

[1] Rabbi, Md Fazle, et al. "Prediction of liver disorders using machine learning algorithms: a comparative study." 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT). IEEE, 2020.
[2] Geetha, C., and A. R. Arunachalam. "Evaluation based approaches for liver disease prediction using machine learning algorithms." 2021 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2021.
[3] Auxilia, L. Alice. "Accuracy prediction using machine learning techniques for indian patient liver disease." 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018.
[4] Hartatik, Hartatik, Mohammad Badri Tamam, and Arief Setyanto. "Prediction for diagnosing liver disease in patients using KNN and Naïve Bayes algorithms." 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS). IEEE, 2020.
[5] Shobana, G., and K. Umamaheswari. "Prediction of liver disease using gradient boost machine learning techniques with feature scaling." 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, 2021.
[6] Sarwar, Muhammad Azeem, et al. "Prediction of diabetes using machine learning algorithms in healthcare." 2018 24th international conference on automation and computing (ICAC). IEEE, 2018.
[7] Veeranki, Sreenivasa Rao, and Manish Varshney. "Intelligent Techniques and Comparative Performance Analysis of Liver Disease Prediction." International Journal of Mechanical Engineering 7.1 (2022): 489-503.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal Volume 13, Iss 02, 2024

[8] Kulkarni, Adeep, Suprit Shinde, and Dipali Kadam. "Automated Prediction of Non Alcoholic Fatty Liver Disease using Machine Learning Algorithms." International Research J. of Engineering and Technology (IRJET) 7.9 (2020): 488-491.
[9] Chen, Ming, and Xudong Zhao. "Fatty liver disease prediction based on multi-layer random forest model." Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence. 2018.
[10] Islam,Md Mohaimenul, et al. "Applications of machine learning in fatty live disease prediction."Building continents of knowledge in oceans of data: the future of co-created eHealth. IOS Press, 2018. 166-170.
[11] Sontakke, Sumedh, Jay Lohokare, and Reshul Dani. "Diagnosis of liver diseases using machine learning." 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI). IEEE, 2017.
[12] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of Data Mining & Knowledge Management Process 8.2 (2018): 01-09.
[13] Kefelegn, Shambel, and Pooja Kamat. "Prediction and analysis of liver disorder diseases by using data mining technique: survey." International Journal of pure and applied mathematics 118.9 (2018): 765-770.
[14] Kuzhippallil, Maria Alex, Carolyn Joseph, and A. Kannan. "Comparative analysis of machine learning techniques for indian liver disease patients." 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2020.
[15] Ramana, Bendi Venkata, and Raja Sarath Kumar Boddu. "Performance comparison of classification algorithms on medical datasets." 2019 IEEE 9th Annual computing and communication workshop and conference (CCWC). IEEE, 2019.