Predicting Mortality for Non-alcoholic Fatty Liver Disease

Kavyalata Kothari

MS Computer Science

CSU Northridge

***Abstract*— By examining data sets from population research, this longitudinal study seeks to predict outcomes in people with non-alcoholic fatty liver disease (NAFLD). NAFLD patients and matched controls in the research were monitored for metabolic abnormalities, cardiac endpoints, and mortality. Using this substantial dataset, the project intends to create predictive models for metabolic abnormalities, cardiac events, and mortality in NAFLD. The results of this study will aid in improving the understanding of the long-term effects of NAFLD and in the creation of focused therapies to enhance patient outcomes.**

Keywords: NAFLD, non-alcoholic fatty liver disease, longitudinal study, predictive modeling, metabolic conditions, cardiac endpoints, mortality, population study, outcomes, matched control group.

I. INTRODUCTION

A significant global health issue is non-alcoholic fatty liver disease (NAFLD), defined by an accumulation of fat in the liver without excessive alcohol consumption. The NAFLD spectrum of liver issues includes anything from simple steatosis to non-alcoholic steatohepatitis (NASH), which can progress to severe liver diseases and increase the risk of cardiovascular events and mortality. Understanding the long-term impacts and the trajectory of NAFLD is necessary for providing effective patient care, risk assessment, and the development of targeted therapeutics.

In this paper, the relationship between NAFLD and important clinical outcomes like metabolic diseases, cardiac endpoints, and mortality is examined. It does this by leveraging a comprehensive dataset collected from community research on NAFLD. Body mass index (BMI), case identifications, follow-up duration, subject identifiers, age, gender, weight, height, and vital status at the last follow-up are just a few of the crucial factors that are captured in the dataset, dubbed "nafld1," which is made up of data from 17,549 observations.

By longitudinally examining this dataset, we want to create prediction models that can anticipate the probability of metabolic problems, cardiac events, and mortality in people with NAFLD. A matched control group is a useful point of comparison for comparative analysis, allowing for a better understanding of the particular risks related to NAFLD. Using age, gender, weight, height, and BMI as potential predictors, we aim to pinpoint the critical elements that influence unfavorable outcomes in NAFLD patients.

This study's conclusions have important clinical and public health ramifications. Medical personnel can identify patients who are more likely to experience metabolic problems, cardiovascular events, or mortality with the help of accurate prediction models for NAFLD outcomes. These models can assist with early intervention, tailor management tactics, and enhance patient care in general. The knowledge acquired from this study can also be used to help create public health initiatives and focused preventive efforts that aim to lessen the burden of complications caused by NAFLD.

II. BACKGROUND

Non-alcoholic fatty liver disease (NAFLD) is a common liver ailment that mostly affects people who drink very little or no alcohol and is characterized by a buildup of fat in the liver cells. Insulin resistance, metabolic syndrome, obesity, and other metabolic illnesses are all directly related to it. Simple fatty liver and non-alcoholic steatohepatitis (NASH), which includes inflammation, hepatocyte damage, and possible fibrosis, are also examples of NAFLD. If NASH is not treated, it may lead to more severe liver conditions. NAFLD pathophysiology is multifaceted and includes genetic predisposition, oxidative stress, inflammation, insulin resistance, and lifestyle variables. Blood tests, imaging studies, and clinical evaluations are frequently used to make diagnoses. While NASH may need additional therapies, management of the condition requires dietary and lifestyle changes, including weight loss, exercise, and a nutritious diet. Obesity, diabetes, and cardiovascular disease are all intimately related to NAFLD, which has a huge negative impact on public health. It's critical to address risk factors and put preventive measures in place to lessen the overall impact of NAFLD and its complications.

III. METHODOLOGY

The dataset employed in this study, "nafld1,csv," [11] is made up of a thorough compilation of data acquired from a population study on non-alcoholic fatty liver disease (NAFLD). It has 17,549 observations across 10 variables, providing crucial knowledge regarding the characteristics and prognoses of NAFLD patients. The dataset provides information on demographic, anthropometric, and clinical factors that are associated with metabolic disorders, cardiac endpoints, and mortality.

The dataset nafld1 is a CSV file with 17549 observations on the following 10 variables. They are as follows: - id: Each participant in the study is given a unique

identification number called a subject identifier. Using this variable, participants can be tracked and distinguished throughout the investigation. age: The "age" variable in the dataset tells us how old each person was when they joined the study. height: The height variable records the height of each subject, measured in centimeters. It enables the calculation of body mass index (BMI) and helps evaluate the impact of height on NAFLD-related complications. bmi: The body mass index represents the calculated BMI of each subject, which is derived from their weight and height measurements. BMI serves as a valuable indicator of obesity and its association with NAFLD progression. case.id: The case identifier variable links each subject to the corresponding NAFLD case to which they are matched. This matching allows for comparative analyses between NAFLD cases and control subjects. 'futime': The 'futime' variable denotes the duration from the initial entry into the study until either the occurrence of death or the most recent follow-up. It provides crucial information about the length of the follow-up period for each subject.

We can determine if each person was alive or had passed away at the time of the most recent follow-up by looking at the "status" field in the dataset. A number of 0 indicates that the person was still alive, whereas a value of 1 indicates that they were no longer alive.

As the first step, the file was imported into the pa ndas data frame. Then after doing the exploratory data analysis, the following observations were extracted.

The dataset had several missing values in the following features:

Features	**Missing Values**
weight	4786
height	3168
bmi	4961

As weight and height are very crucial features for predicting status, it might be appropriate to remove those variables to ensure the integrity of the analysis.
Body mass index (BMI) is a measure of body fat based on height and weight that applies to adult men and women. We already have the 'weight' and 'height' attributes which can be used to calculate the BMI and fill in the missing values.
Formula to calculate BMI: weight (kg)/ height(mt)squared.
As we had height in centimeters, we divided it by 100 to convert it into meters.
Feature engineering is required to transform raw data into a set of meaningful and informative features that can be used to train machine learning models. It involves creating new features or modifying existing ones to enhance the performance of the models and improve their ability to understand patterns and make accurate predictions.
We dropped two features: 'Unnamed: 0' as it was a duplicate of the index. We also dropped 'case.id' as it is the id of the NAFLD case to whom this subject is matched which has no relevance to our predictive model.

Machine learning model performance is assessed using performance metrics, which are numerical measurements. These measurements assist in evaluating the model's performance and determine how well it predicts or categorizes data. The type of problem (e.g., classification, regression) and the specific objectives of the investigation will determine the performance indicators to be used. The ratio of true positives to the total number of recalls is known as recall (sensitivity or true positive rate).

The following performance matrices in the models were used:

Accuracy: The number of correctly classified instances with respect to all instances.

Precision: It measures a model's accuracy in identifying positive instances as the ratio of true positives to the total of true positives and false positives.

Recall (Sensitivity or True Positive Rate): The ratio of true positives to the total of true positives and false negatives, indicating the model's sensitivity to detecting all positive occurrences.

F1 Score: The harmonic mean of recall and precision, which offers a balanced measurement considering both measurements.

Area Under the ROC Curve (AUC-ROC): This statistic measures the trade-off between the true and false positive rates at various classification points.

Mean Squared Error (MSE): A measure of the average squared departure from the true values, calculated as the average of the squared discrepancies between the predicted and true values.

Root Mean Squared Error (RMSE): This statistic provides a measurement in the same units as the target variable and is the square root of the MSE.

Mean Absolute Error (MAE): An indicator of average absolute deviation, calculated as the average of the absolute discrepancies between true and predicted values.

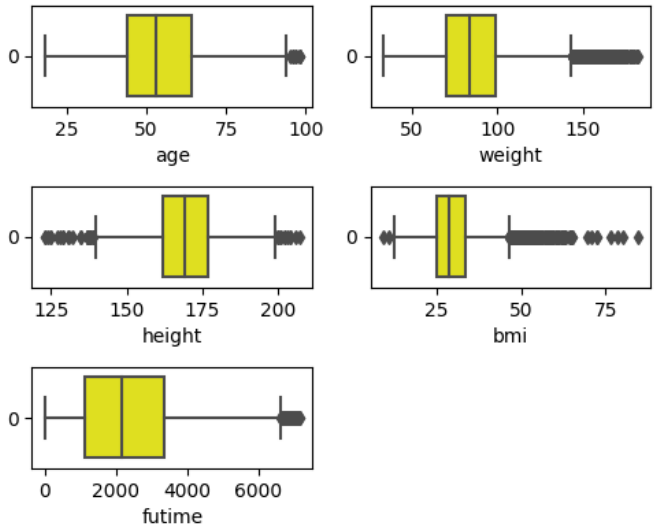
For the detection of the outliers, boxplots was used to visualize. It can be observed from the below image that the age and 'futime' features have the least number of outliers whereas features such as height, weight, and BMI. These outliers can potentially distort the analysis or modeling process. It was handled through the following steps:

For each column, the first quartile (q1) and the third quartile (q3) were calculated using the quantile () function from pandas. The interquartile range (IQR) was then computed as the difference between q3 and q1.

The lower bound and upper bound for outliers were defined by subtracting and adding 1.5 times the IQR from q1 and q3, respectively. This followed the commonly used Tukey's fences method, where values below the lower bound and above the upper bound were considered outliers.

The data frame was filtered using Boolean indexing to keep only the rows where the values in the respective columns

were within the calculated bounds. This effectively removed the outliers from the dataset.



Many machine learning estimators require the standardization of a dataset because they may behave poorly if the individual features do not more or less resemble standard normally distributed data (for example, Gaussian with 0 mean and unit variance). For example, several components of a learning algorithm's objective function (such as the SVM's RBF kernel or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have a variance that is distributed in the same order. In our dataset, the attributes have the following scales: -

	id	age	male	weight	height	bmi
min	1.000000	18.000000	0.000000	35.100000	141.000000	13.040536
max	17566.000000	94.000000	1.000000	143.200000	197.000000	45.843248
mean	8773.340557	53.713119	0.446074	84.370287	169.368311	29.301466

	futime	status
min	7.000000	0.000000
max	6656.000000	1.000000
mean	2382.172945	0.079513

Scaling is required for attributes like weight, height, and BMI as they have a large range of values. Scaling helps to normalize the attributes and ensures that they are on a similar scale as other attributes. In order to achieve the above, we used StandardScaler() of the Scikit Learn library on the dataset.

Various machine learning algorithms, such as logistic regression, random forest, and XGBoost will be employed to develop predictive models.

A total of 4 models were implemented which included:
Logistic Regression, Random Forest, XGBoosting, and Logistic Regression with Hyperparameter Tuning.
For this implementation first, the dataset was split into training and testing data. The split was 80-20% of
the entire data.
All the above models were implemented using Python's ScikitLearn Library.
Logistic regression, which is a linear model, was trained using the logistic regression function and its predictions were made using the prediction method. Accuracy, precision, recall, and AUC-ROC scores were calculated to evaluate the model. For Random Forest, a random forest classifier was trained using the Random Forest Classifier function, and predictions were made on the test data. Accuracy, precision, recall, and AUC-ROC scores were calculated to evaluate the model's performance. Similarly, for XGBoosting, a custom class called XGBoosting was defined. The XGBoosting class had methods to fit the model, make predictions, and calculate mean absolute error (MAE), mean squared error (MSE), and R-squared score. The XGBoosting model was trained using the XGBRegressor function from the xgboost library. The predicted values were obtained using the predict method. Metrics such as precision, recall, AUC-ROC score, MAE, MSE, and R-squared score were printed to assess the performance of the XGBoosting model. Overall, logistic regression, random forest, and XGBoosting were implemented with their respective evaluation metrics to address the classification task in the project.
For Logistic Regression using Hyperparameter tuning, a hyperparameter grid was defined using the param_grid dictionary. The grid included hyperparameters for the logistic regression model, such as the inverse of regularization strength (C), regularization penalty type (penalty), and the algorithm for optimization (solver). The GridSearchCV function was used to perform a grid search with cross-validation, where the logistic regression model is trained and evaluated using different combinations of hyperparameters.
The grid search identifies the best hyperparameter values that yield the highest performance based on the provided evaluation metric (default is mean accuracy) across the specified number of cross-validation folds (cv=5).
The best hyperparameter values found using grid_search.best_params_ were - 'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'. The best logistic regression model, best_logreg, was obtained using grid_search.best_estimator_.Predictions were made using the best logistic regression model on the test data (X_test). The accuracy, precision, recall, and AUC-ROC score were calculated to evaluate the performance of the best logistic regression model. Precision and recall were computed using precision_score and recall_score functions, respectively. The predicted probabilities for the positive class were obtained using predict_proba and stored in y_scores. The AUC-ROC score was calculated using roc_auc_score.
The grid search approach helps in finding the optimal hyperparameter values, leading to improved model performance.

IV. Related Work

Zobair Younossi and colleagues [1] (2016) reported on the global epidemiology of non-alcoholic fatty liver disease. Non-alcoholic fatty liver disease (NAFLD) is a major cause of liver disease worldwide. The authors estimated the global prevalence, incidence, progression, and outcomes of NAFLD and non-alcoholic steatohepatitis. All studies were reviewed by three independent investigators. The global prevalence of Non-alcoholic fatty liver disease is 25.24% with the highest prevalence in the Middle East and South America and lowest in Africa. Metabolic comorbidities associated with NAFLD included obesity, type 2 diabetes, hyperlipidemia, hypertension, and metabolic syndrome.

In ‘NAFLD’, Christopher Targher[2] (2015) reported that Non-alcoholic fatty liver disease is the most common cause of chronic liver disease in Western countries. Non-alcoholic fatty liver disease (NAFLD) is predicted to become the most frequent indication for liver transplantation by 2030. The primary liver pathology in NAFLD affects hepatic structure and function to cause morbidity and mortality. NAFLD increases the risk of type 2 diabetes mellitus (T2DM), cardiovascular (CVD) and cardiac diseases, and chronic kidney disease (CKD).

In ‘Non-alcoholic fatty liver disease and risk of incident cardiovascular disease’, Giovanni Targher and colleagues [3] (2016) reported that a total of 16 unique, observational prospective and retrospective studies with 34,043 adult individuals (36.3% with NAFLD) and approximately 2,600 CVD outcomes over a median period of 6.9years were included in the final analysis. Patients with more ‘severe' NAFLD were more likely to develop fatal and non-fatal CVD events. Non-alcoholic fatty liver disease (NAFLD) is associated with an increased risk of fatal and non-fatal CVD events. The observational design of the studies included does not allow us to prove that NAFLD causes cardiovascular disease. The research involved 34043 adult individuals.

Non-alcoholic fatty liver [4] disease is defined as the presence of hepatic fat accumulation. The ongoing persistence of obesity with an increasing rate of diabetes will increase the prevalence of NAFLD. As this population ages, many will develop cirrhosis and end-stage liver disease. Non-alcoholic fatty liver disease (NAFLD) is the most common liver disease in the world. NAFLD is commonly associated with metabolic comorbidities, including obesity, type II diabetes, dyslipidemia, and metabolic syndrome.

A team led by Qing Ye of the Cedars-Sinai Medical Center [5] (2020) described the global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease. Non-alcoholic fatty liver disease is commonly associated with obesity. The prevalence of non-obese NAFLD in the general population varied from 25% or lower in some countries to higher than 50% in others. They aimed to characterize the prevalence, incidence, and long-term outcomes of non-obese or lean NAFLD at a global level. The authors reviewed 10576383 studies.

NAFLD is a multisystem disease, affecting many extra-hepatic organs and cardiometabolic complications, including Type 2 diabetes and cardiovascular disease. [6]

Nonalcoholic fatty liver disease is predicted to become the most frequent indication for liver transplantation by 2030. There is compelling evidence that NAFLD is a multisystem disease, affecting many extra-hepatic organs. Nonalcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease in Western countries.

Non-alcoholic fatty liver disease (NAFLD) encompasses a spectrum of liver diseases from simple steatosis with hepatic lipid accumulation to end-stage liver disease with decompensated cirrhosis, liver failure, and hepatocellular carcinoma.[7] NAFLD increases the risk of extrahepatic diseases such as type 2 diabetes mellitus, cardiovascular disease, cardiac diseases, and chronic kidney disease. There is a triangular relationship between dysbiosis and T2DM and NAFLD. In 2013, NAFLD was the second most frequent indication for liver transplantation behind hepatitis C.

In ‘Non-alcoholic fatty liver disease and risk of fatal and non-fatal cardiovascular events’, Alessandro Mantovani and colleagues [8] (2021) reported that NAFLD was associated with a moderately increased risk of fatal or non-fatal CVD events. This risk markedly increased across the severity of NAFLD, especially the stage of fibrosis. Studies have reported a significant association between non-alcoholic fatty liver disease (NAFLD) and increased incidence of cardiovascular disease (CVD). The authors systematically searched PubMed, Scopus, and Web of Science from database inception to July 1, 2021, to identify eligible observational studies. The authors identified 36 longitudinal studies with aggregate data on 5 802 226 middle-aged individuals and 99 668 incident cases of fatal and non-fatal CVD events over a median follow-up of 6·5 years (IQR 5·0–10·2).

Non-alcoholic fatty liver disease (NAFLD)[9] is the most common chronic liver disease. Obesity is a common metabolic risk factor associated with NAFLD9,10,11. Men accumulate visceral and subcutaneous fat mainly in the trunk and abdomen with continuous changes before and after puberty. This study applied machine learning (ML) methods to identify significant classifiers of NAFLD. NAFLD Fibrosis Score and FIB-4 have the potential to detect advanced fibrosis and the progression of fibrosis.

The analysis involved 593 individuals. The authors’ results appear to back up prior work in this field: “Neck circumference reflects the amount of subcutaneous fat in the upper body. Neck circumference is a reliable factor in determining central obesity, study finds,” Razmpour posited. However, “ML is expected to minimize some of the limitations associated with ultrasound use, including limited capability in detecting fatty infiltration, operator dependency, and subjective assessment. Application of ML techniques on body composition and anthropometric measures can help physicians in their clinical decision making,” admit the authors. They suggest that using liver biopsy outcomes would generate more valid results. Future studies should include body composition and anthropometric measures such as sagittal abdominal diameter and peri-renal fat.

Non-alcoholic liver steatosis is the leading cause of chronic liver disease in Western countries.[10] The model they propose can be exploited to target only those subjects who have a real need for further investigation, leading to a reduction in waiting lists, costs, and time required for instrumental examinations. This condition increases the risk of cardiovascular disease, type 2 diabetes mellitus, and chronic kidney disease. The authors' goal was to develop and validate a simple Neural Network (NN)-based web app that could be used to predict NAFLD absence. 2970 subjects were involved in the research. Some of their findings appear to substantiate prior work in this field: “This limits the use of these models in large-scale epidemiologic studies and health database research. The performance in terms of AUC is superior to LAP65, Hepatic steatosis index, SteatoTest,

APRI68, NAFLD fibrosis score," Soriano posited. Discussing possible improvements, "There are several limitations to this work. The study was conducted in a single center and so has a rather limited sample size. The NN is strongly linked to the identification of the NAFLD condition only in a Mediterranean population," they observe.

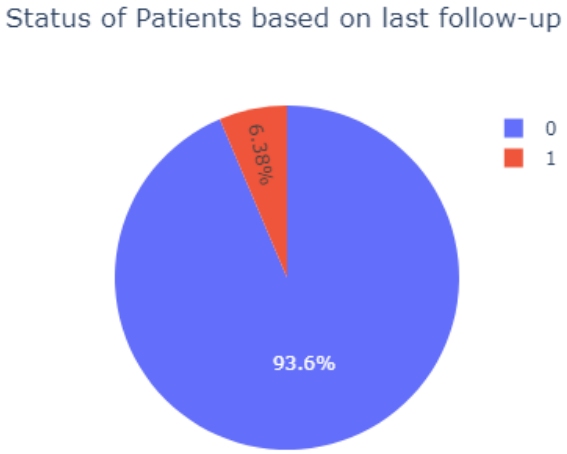
V. Environmental Setup

The project is set up on Jupyter Notebook installed over the Anaconda Platform. Following is the hardware configuration of the system: -

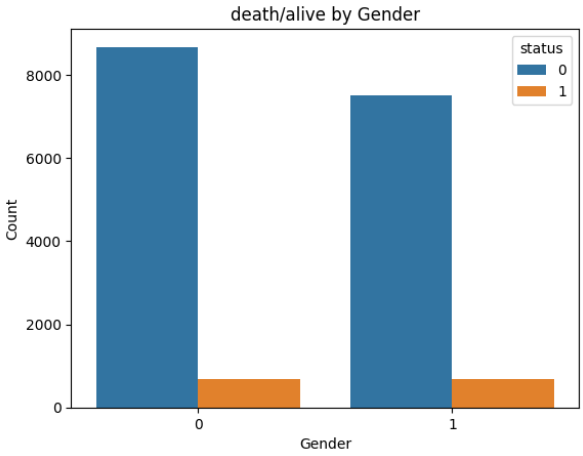
Processor	11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 2.80 GHz
Installed RAM	16.0 GB (15.8 GB usable)
Product ID	00342-21979-82910-AAOEM
System type	64-bit operating system, x64-based processor

VI. Result

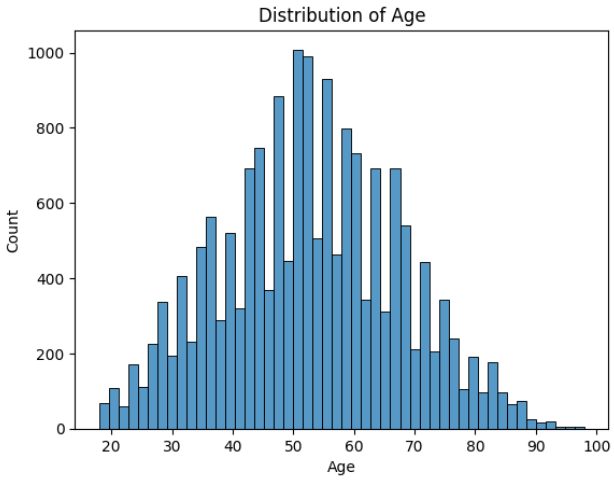
To predict the mortality of the patient suffering from the disease, the target attribute 'Status' was predicted. 'Status' has two classes 0 and 1. Class '0' represents patient is alive while 1 represents 'dead'.
Here is the exploratory data analysis: -
We can see that 6.38% of patients died even after the follow-up but the rest 93.6% survived.



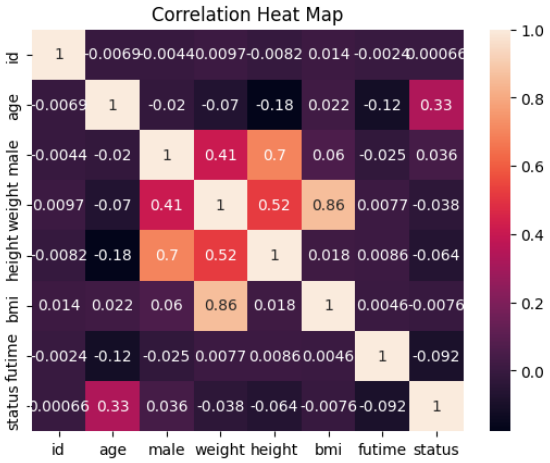
The below graphs help us to understand whether the gender of the patient has any impact on the mortality of the patient. We can see from the graph that the data is balanced between both genders. The Female ('male=1') and male('male=0') have significantly the same outcome in the status.



From the below plot, we can observe that most patients are aged between 40-70 years which signifies NAFLD is most common among people of 50-60 age group.



The correlation matrix helps us to find out the dependency of attributes on each other. The highly correlated values between weight and height represent that they are highly dependable while 'age' and 'height' have much less correlated to each other. However, all we are doing is calculating the linear correlation of the several variables. Again, this sheds some light on potential new traits that might come from the fusion of these, but these findings should be interpreted with care.



We used the following types of models to train and test our data: Logistic Regression, Random Forest, XGBoost, and Logistic Regression with Hyper Parameter tuning.

Logistic Regression

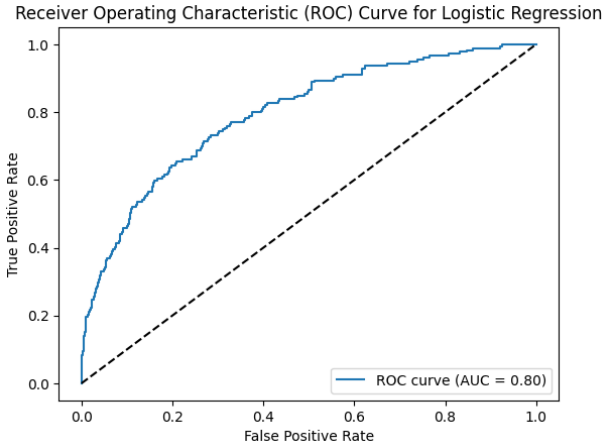
The logistic regression model properly identified 93.08% of the cases, according to its accuracy score of 0. 9308. The model's recall, however, was only 0.13, which indicated that it had trouble detecting true positives (dead people). The logistic regression model had a precision of 0.68, which indicates a moderate level of accuracy in predicting favorable outcomes. This model's AUC-ROC score was 0.79, indicating it had a strong ability to tell dead patients apart from living ones.

Logistic Regression Accuracy: 0.9308333333333333

Precision: 0.6857142857142857

Recall: 0.1340782122905028

AUC-ROC Score: 0.7951549329785014



Random Forest

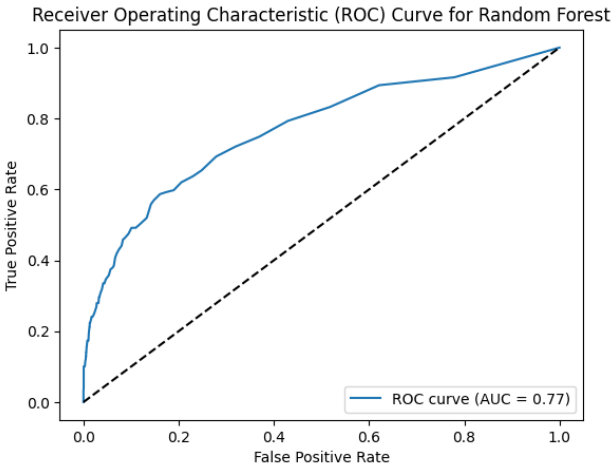
Compared to logistic regression, the random forest model produced results with a slightly lower accuracy of 0.9295. Improved recall (0.21) and precision (0.57) were seen, indicating greater performance in identifying the deceased. The random forest model's AUC-ROC score was 0.76, showing a fair capacity for differentiating between dead and surviving patients.

Random Forest Accuracy: 0.9295833333333333

Precision: 0.5757575757575758

Recall: 0.2122905027932961

AUC-ROC Score: 0.7679149509884067



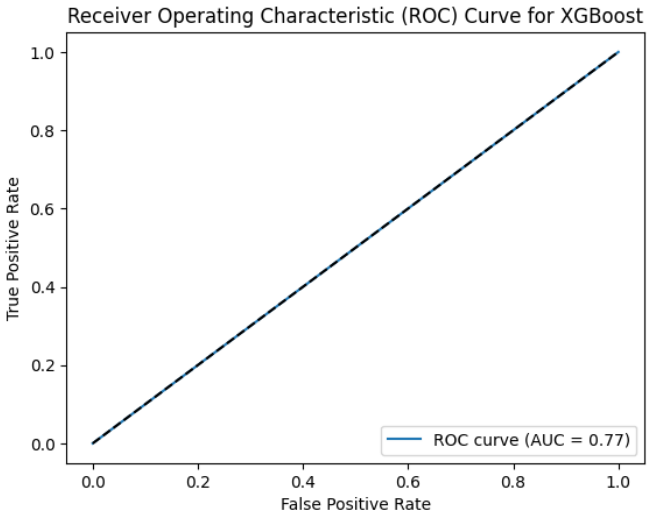
XGBoost

The accuracy of the gradient boosting algorithm XGBoost, which is comparable to logistic regression, was 0.9250. Both the recall and precision were 0.9254, demonstrating a balanced performance in making predictions of favorable events. The R2 rating for XGBoost, however, was -0.28, suggesting that the model did not adequately account for the variance in the data. With an AUC-ROC score of 0.75, XGBoost has a respectable level of discriminatory power.

Mean Absolute Error: 0.26 degrees.

Mean Square Error: 0.09 degrees.

R2 Score: -0.28 degrees.
xgBoost Accuracy: 0.9254166666666667
Precision: 0.9254
Recall: 0.9254
AUC-ROC Score: 0.7503



Logistic Regression with Hyperparameter Tuning

The logistic regression model with hyperparameter tuning, using the best hyperparameters (C = 0.1, penalty = l1, solver = 'liblinear'), achieved an accuracy of 0.93, similar to the other models. The precision was 0.68, indicating a moderate level of accuracy in predicting positive outcomes, while the recall was low (0.12). The AUC-ROC score for this model was 0.79, similar to the logistic regression model without hyperparameter tuning.

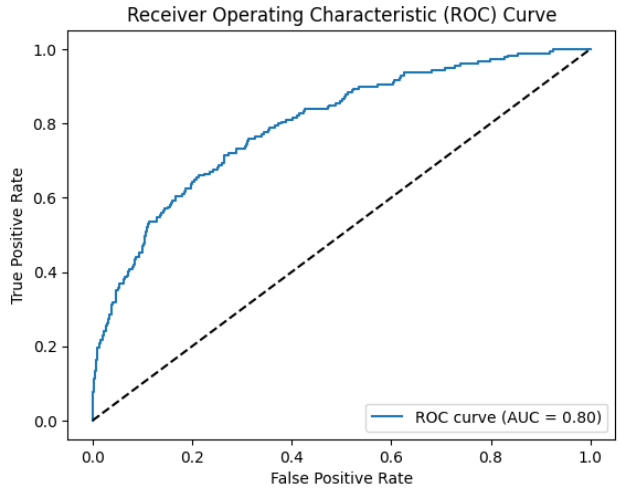
Best Hyperparameters: {'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}

Best Logistic Regression Accuracy: 0.9304166666666667

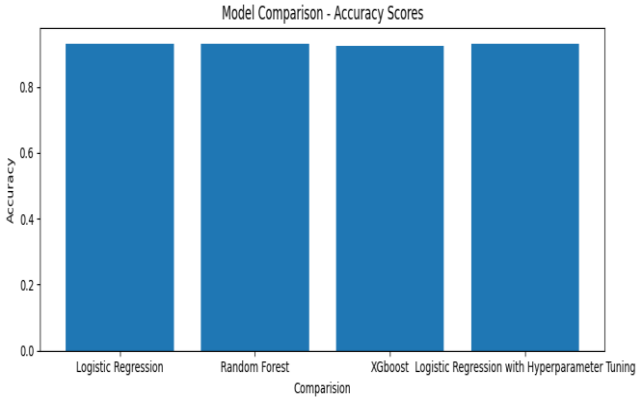
Precision: 0.6875

Recall: 0.12290502793296089

AUC-ROC Score: 0.7960881277998988



We were able to find that Logistic Regression with hyperparameter tuning was the best model to predict the correct label for the status feature.

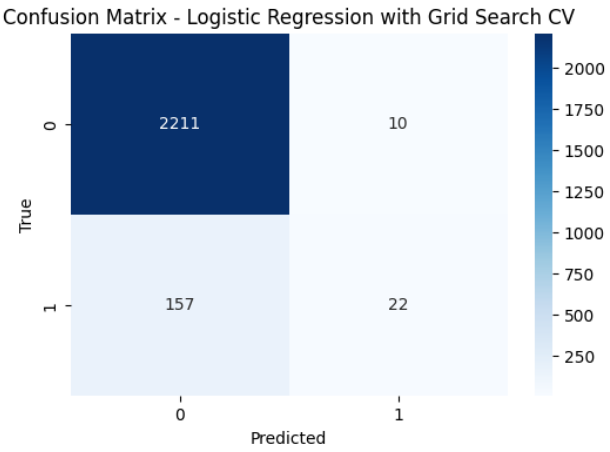


V. DISCUSSION

The confusion matrix is a table that summarizes the performance of a classification model by displaying the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

From the matrix, we can see 2211 values for class '0' were correctly predicted i.e status=0 while 157 values were predicted as status=1 but the actual status was '0'.

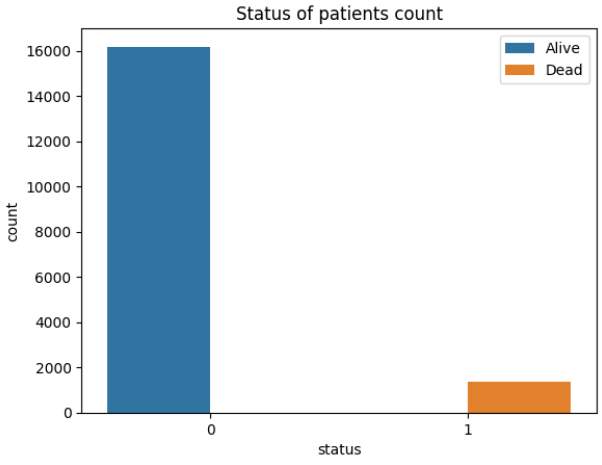
For status=1, 22 values were labeled correctly while 10 values were incorrectly labelled.



VI. LIMITATIONS

This study has made significant advancements in using machine learning models to predict mortality in individuals with non-alcoholic fatty liver disease (NAFLD). However, it is important to acknowledge several limitations and consider additional dataset details and previous findings:

Class Imbalance: The dataset used for training and evaluation exhibits an imbalance between the number of records indicating individuals who are alive and deceased. The 'status' has more than 16000 classes with status='0' (alive) while the negative class status='1' (dead) has approximately 1000 records. This class imbalance can impact the models' performance, especially in terms of recall and precision for the minority class. The below image illustrates the dominance of the alive class may lead to a higher tendency to predict individuals as alive, potentially affecting the overall accuracy of mortality prediction.



Improved Minority Class Prediction: Due to the smaller number of deceased individuals in the dataset, accurately predicting mortality cases can be challenging. The models may have lower recall and precision for the deceased class compared to the alive class. To overcome this limitation, techniques such as cost-sensitive learning, ensemble methods, or anomaly detection can be employed to improve the models' ability to identify and predict mortality, particularly for the minority class.

Generalizability to Diverse Populations: The developed models were trained and evaluated on a specific dataset,

raising questions about their applicability to diverse populations and healthcare settings. External validation using independent datasets representing different demographics and clinical contexts is necessary to assess the models' performance across diverse populations and ensure their reliability and applicability beyond the original dataset.

Longitudinal Analysis and Temporal Dynamics: The current models focus on predicting mortality based on the available data at the study's entry point, without explicitly considering the temporal dynamics and changes in risk factors over time. Incorporating longitudinal data and time-varying variables could provide a more comprehensive understanding of disease progression and improve the accuracy of mortality predictions in NAFLD.

Additional Data and Variables: While the current models utilize relevant variables such as age, gender, weight, and BMI, there may be other important factors not included in the dataset that significantly influence mortality prediction in NAFLD. Integrating additional data sources such as genetic information, biomarkers, socioeconomic factors, and lifestyle variables could enhance the predictive performance of the models and capture a broader range of risk factors.

Data Quality and Completeness: The accuracy and reliability of the models heavily depend on the quality and completeness of the input data. Inaccurate or incomplete measurements, missing data, or data entry errors may introduce noise and bias into the models, potentially impacting their predictive performance. Implementing thorough data quality assurance protocols and addressing missing data issues are crucial for ensuring the reliability and robustness of the models.

Despite these limitations, the results of this study offer valuable insights into the potential of machine learning models for predicting mortality in NAFLD. Further research and refinement are necessary to address these limitations and improve the clinical applicability and reliability of the models. In future work, addressing the class imbalance, incorporating additional data, conducting external validation, considering longitudinal analysis, and ensuring ethical considerations, future studies can advance the field of mortality prediction in NAFLD, leading to improved patient care and outcomes.

VII. Conclusion

The findings of this study show that mortality for people with non-alcoholic fatty liver disease (NAFLD) may be accurately predicted using machine learning algorithms. The accuracy ranged from 0.9282 to 0.9382 for the logistic regression, random forest, XGBoost, and logistic regression with hyperparameter tuning models.

The Logistic Regression with tuning model performed best at detecting deceased people among the models tested, displaying the highest accuracy, precision, and recall. However, compared to the other models, the Random Forest without hyperparameter adjustment had a slightly higher AUC-ROC score, indicating stronger discriminatory power.

While these models produced relatively accurate predictions, it's crucial to keep in mind that they have some drawbacks. The relatively low recall scores show that the models had trouble distinguishing between true positives. The XGBoost model also had a negative R2 score, indicating that it did not adequately account for the data's volatility.

The machine learning models created in this study's conclusion offer important insights into predicting mortality for people with NAFLD. Enhanced study and development of these models

VIII. References

1. Younossi, Z. M., Koenig, A. B., Abdelatif, D., Fazel, Y., Henry, L., & Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology (Baltimore, Md.), 64(1), 73–84. https://doi.org/10.1002/hep.28431

2. Byrne, C. D., & Targher, G. (2015). NAFLD: a multisystem disease. Journal of hepatology, 62(1 Suppl), S47–S64. https://doi.org/10.1016/j.jhep.2014.12.012

3. Targher, G., Byrne, C. D., Lonardo, A., Zoppini, G., & Barbui, C. (2016). Non-alcoholic fatty liver disease and risk of incident cardiovascular disease: A meta-analysis. Journal of hepatology, 65(3), 589–600. https://doi.org/10.1016/j.jhep.2016.05.013

4. Perumpail, B. J., Khan, M. A., Yoo, E. R., Cholankeril, G., Kim, D., & Ahmed, A. (2017). Clinical epidemiology and disease burden of nonalcoholic fatty liver disease. World journal of Gastroenterology, 23(47), 8263–8276. https://doi.org/10.3748/wjg.v23.i47.8263

5. Ye, Q., Zou, B., Yeo, Y. H., Li, J., Huang, D. Q., Wu, Y., Yang, H., Liu, C., Kam, L. Y., Tan, X. X. E., Chien, N., Trinh, S., Henry, L., Stave, C. D., Hosaka, T., Cheung, R. C., & Nguyen, M. H. (2020). Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. The lancet. Gastroenterology & hepatology, 5(8), 739–752. https://doi.org/10.1016/S2468-1253(20)30077-7

6. Miele L, Targher G. Understanding the association between developing a fatty liver and subsequent cardio-metabolic complications. Expert Rev Gastroenterol Hepatol. 2015;9(10):1243-5. doi: 10.1586/17474124.2015.1074860. Epub 2015 Jul 29. PMID: 26224399.

7. Scorletti E, Byrne CD. Extrahepatic Diseases and NAFLD: The Triangular Relationship between NAFLD, Type 2-Diabetes and Dysbiosis. Dig Dis. 2016;34 Suppl 1:11-8. doi: 10.1159/000447276. Epub 2016 Aug 22. PMID: 27548822.

8. Mantovani A, Csermely A, Petracca G, Beatrice G, Corey KE, Simon TG, Byrne CD, Targher G. Non-alcoholic fatty liver disease and risk of fatal and non-fatal cardiovascular events: an updated systematic review and meta-analysis. Lancet Gastroenterol Hepatol. 2021 Nov;6(11):903-913. doi: 10.1016/S2468-1253(21)00308-3. Epub 2021 Sep 21. PMID: 34555346.

9. Razmpour, F., Daryabeygi-Khotbehsara, R., Soleimani, D., Asgharnezhad, H., Shamsi, A., Bajestani, G. S., Nematy, M., Pour, M. R., Maddison, R., & Islam, S. M. S. (2023). Application of machine learning in predicting non-alcoholic fatty liver disease using anthropometric and body composition indices. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-32129-y

10. Sorino, P., Campanella, A., Bonfiglio, C., Mirizzi, A., Franco, I., Bianco, A., Caruso, M. G., Misciagna, G., Aballay, L. R., Buongiorno, C., Liuzzi, R., Cisternino, A. M., Notarnicola, M., Chiloiro, M., Fallucchi, F., Pascoschi, G., & Osella, A. R. (2021). Development and validation of a neural network for NAFLD diagnosis. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-99400-y

11. Singh, U. (2023, May 9). Non-alcohol fatty liver disease (NAFLD). Kaggle. https://www.kaggle.com/datasets/utkarshx27/non-alcohol-fatty-liver-disease