Liver Disease Prediction

Andrew Bessire

Hanisha Pasupulate

Nick Killian

Robert Jenkins

Sheikh-Sedat Touray

Abstract

Our study focused on analyzing a subset of the Health Checkup Results dataset available on Kaggle (https://www.kaggle.com/datasets/hongseoi/health-checkup-result). This dataset contains the health checkup results for South Korean citizens taken by the South Korean National Health Insurance Service spanning from 2002-2020. Our analysis specifically delved into data recorded in 2009 and 2019, encompassing information from one million individuals in each respective year.

Background and Introduction

Liver disease is a major global public health problem, accounting for more than two million deaths per year. Most of these deaths can ultimately be attributed to complications from cirrhosis or liver cancer, but the origins of these conditions are numerous and diverse. These include viral hepatitis, principally hepatitis B and C virus, but also including other pathogens; alcohol consumption; and non-alcoholic fatty liver disease (NAFLD). NAFLD can be associated with insulin resistance or metabolic syndrome, but is not necessarily dependent upon the presence of obesity or clinical indications of diabetes. In addition to the previous etiologies, acute liver damage may also be caused by ingestion of drugs or other substances, such as

acetaminophen or aflatoxin, or may originate from autoimmune conditions. Many of the aforementioned conditions are becoming more prevalent as society becomes more affluent, more sedentary, and is living longer.[1-3]

Liver failure is usually the result of years of chronic liver disease, but many of the factors involved can be avoided or controlled. Therefore, early intervention may allow for improved long-term outcomes when liver dysfunction is discovered. Although there are outward signs of liver dysfunction including jaundice and abdominal pain, these may not be observed until problems are more advanced. A diagnosis of liver dysfunction usually includes a panel of blood tests which measure the enzymes serum aspartate aminotransferase (AST), alanine aminotransferase (ALT), alkaline phosphatase (AP), and γ-glutamyl transpeptidase (GGT).[4,5] These enzymes are associated with the liver to a large degree, and are released into the bloodstream when liver cells are damaged. However, there may be opportunities to gain insight into the possible presence of liver dysfunction when a more limited analysis of blood is conducted, such as at health fairs, in which only blood glucose and standard lipids (low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides) are measured. Our project idea is to try to predict elevated levels of AST, ALT, or GGT from the more commonly measured health indicators, including blood pressure, weight, glucose levels, LDL, HDL, and triglycerides. GGT in particular has been associated with dyslipidemia in other studies.[6, 7]

The dataset we used for conducting the project was a portion of the Health Checkup Results dataset from Kaggle (https://www.kaggle.com/datasets/hongseoi/health-checkup-result). This dataset contains the health checkup results for South Korean citizens taken by the South Korean National Health Insurance Service spanning from 2002-2020. We limited our analysis to

the observations for the years of 2009 and 2019, both of which contain data from 1 million individuals. The pertinent variables in the data are the following:

- Age: 20 years to >80 years, expressed as binned categorical data
- Sex: Male/Female
- Blood pressure: BP_LWST (diastolic) and BP_High (systolic), both in mmHg
- Blood glucose (BLDS): pre-meal levels expressed as mg/dL
- Total cholesterol (TOT_CHOLE): normal ranges are 150-250 mg/dL
- Low-density lipoprotein (LDL_CHOLE): upper limit of normal is 170 mg/dL
- High-density lipoprotein (HDL_CHOLE): normal ranges are 35-65 mg/dL
- Triglycerides: normal ranges are 35-135 mg/dL
- Creatinine: normal ranges in serum are 0.8-1.7 mg/dL
- AST (SGOT_AST): normal ranges are 0-40 IU/L
- ALT (SGPT_ALT): normal ranges are 0-40 IU/L
- GGT (GAMMA_GTP): normal ranges: 11-64 (men), 8-35 (women) IU/L
- Smoking status (SMK_STAT): 3 categories - 1, never smoked; 2, previous smoker; 3 current smoker
- Drinking status (DRK_YN): 0, doesn't drink; 1 drinks (amount not specified)

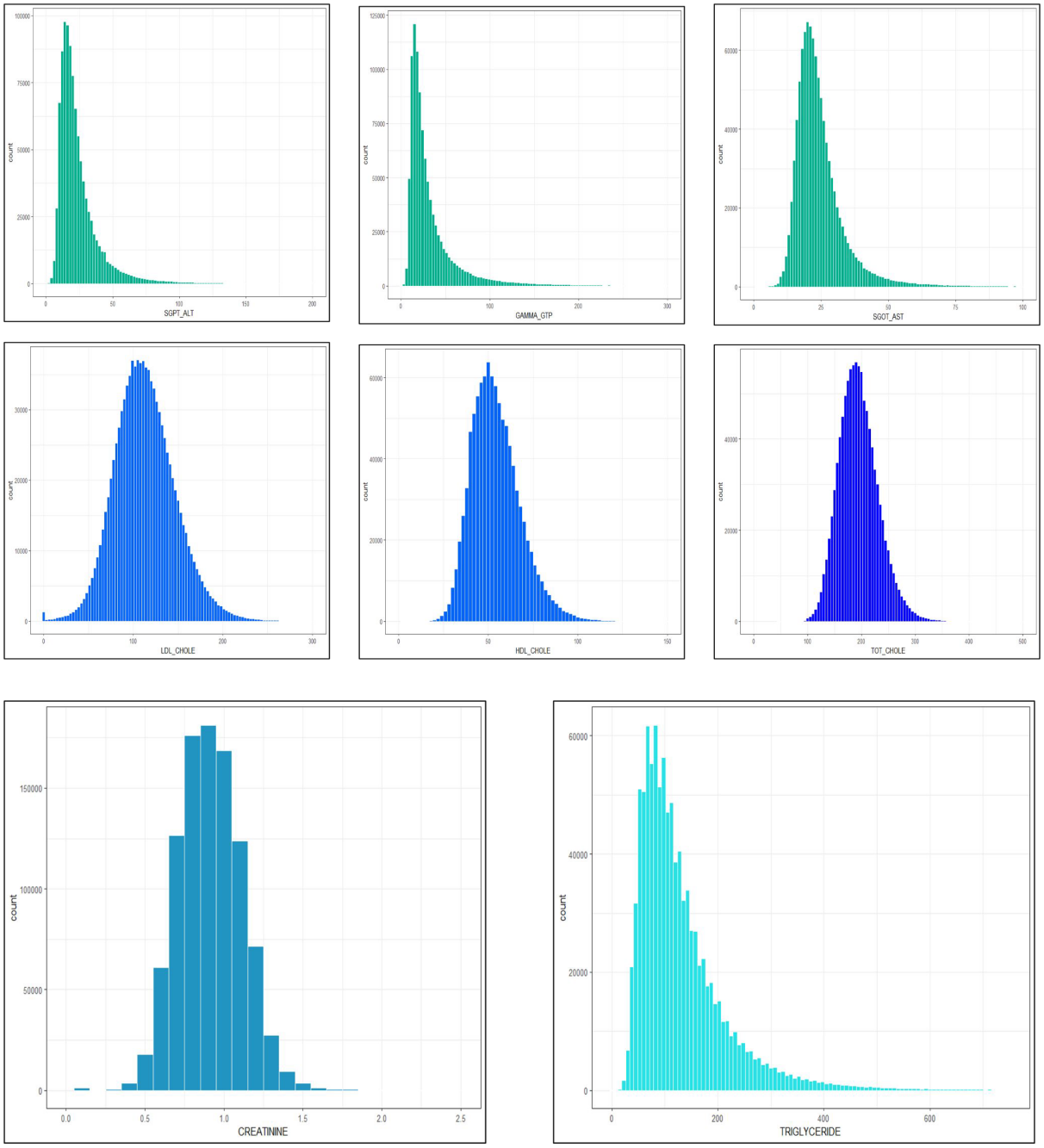
This data also included other variables, such as missing teeth, which are not pertinent to the project, hence we excluded them.

With this dataset we will focus on the variables that may relate to liver disease and we hope to generate a model that can accurately predict a likelihood of liver dysfunction. Some of the variables we will use for this prediction will be Sex, Age, Smoking Status, Drinking Status,

and blood test measurements for which there is some evidence of a relationship to liver function (triglycerides, total cholesterol, LDL, and HDL)[3], and SGOT-AST, SGPT-ALT, Gamma-GTP.[4,7] Since liver disease is not explicitly labeled in our dataset, we will apply supervised machine learning methods to predict liver enzyme levels (AST, ALT, and Gamma-GTP, all generally accepted as biomarkers of liver injury[4]) for use as an output variable in training and test sets. We then can train models using this output variable, and check their predictive ability. In addition to creating the predictive models, we hope that we can gain some insights into how these variables have changed over the ten-year period from 2009 to 2019.

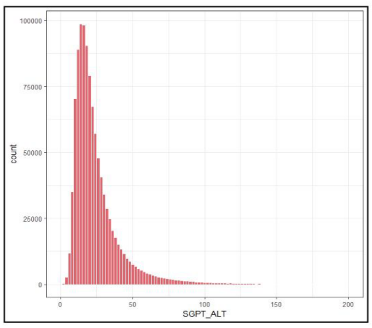
Key Variables from 2009

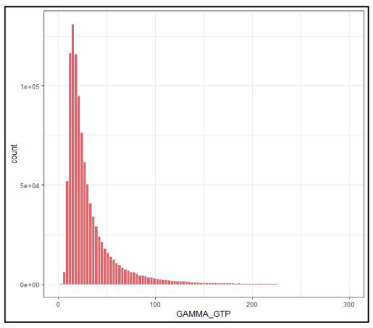
Figure 1. The histograms in this figure depict key variables used in our model. From top left to bottom right: SGPT_ALT, GAMMA_GPT, SGOT_AST, LDL_CHOLE, HDL_CHOLE, and TOT_CHOLE. Most observations fall within normal values while some fall outside this range. We see very few outliers, although we see a small spike at 0 in our LDL_CHOLE indicating a potential error in recording. Most observations fall in the normal ranges for creatinine and triglyceride and we have a small outlier spike at 0 in our creatinine.

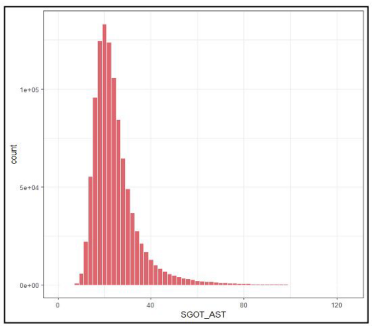


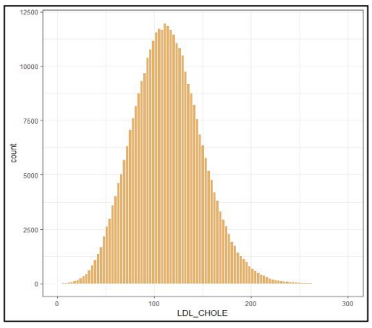
Key Variables from 2019

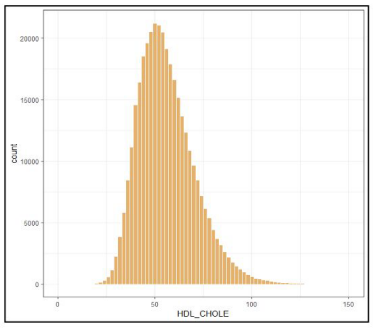
Figure 2. Here we see the same variables graphed in a very similar unimodal distribution to Figure 1. Most observations fall within normal ranges and a few outside their normal range. The outlier spikes at 0 for creatinine and LDL_CHOLE observed for 2009 are not present in the data from 2019.

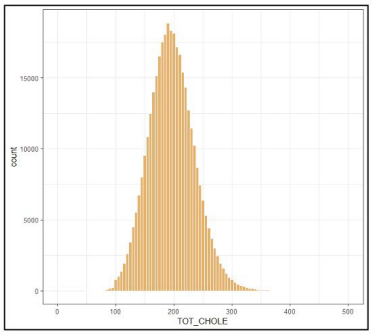

count
SGPT_ALT

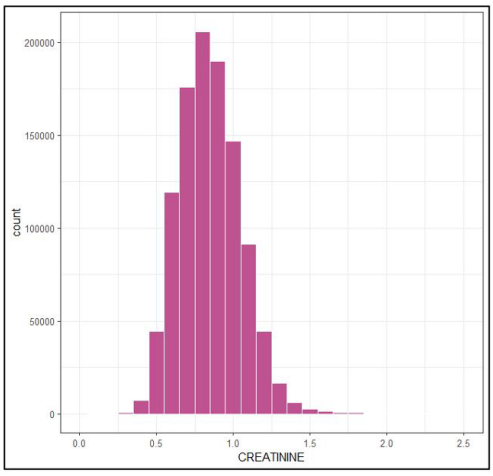

count
GAMMA_GTP

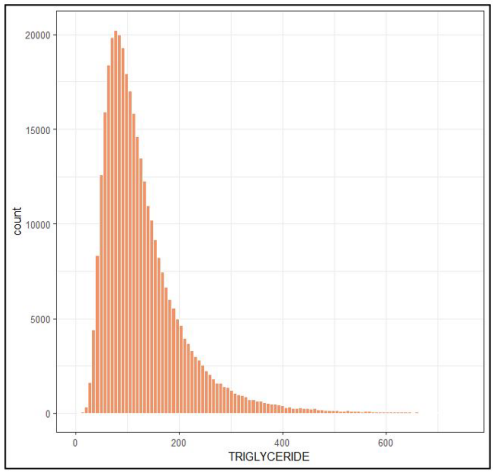

count
SGOT_AST


count
LDL_CHOLE


count
HDL_CHOLE


count
TOT_CHOLE


count
CREATININE


count
TRIGLYCERIDE

Methods

The methods we are going to follow in this project include some of the methods learned in this class. After exploring the data, attempting to clean it up, make visualizations, and make summary statistics. We also checked for "Not a number" (NaN) values in the datasets. For the 2009 data set when we dropped the columns that had these values we show that the remainder of the columns still includes the predictors we were interested in to predict these conditions whether someone has a likelihood to have a liver disease or not. However, for the 2019 dataset when we attempted to drop the columns that had NaN values we saw that it dropped more columns that we wanted and because of that and since we had over a million observations, we decided to only drop the rows with these 'Not a Number' values. Other than that the dataset is very clean.

In order to set up this classification problem we mapped all the values between 0 and 40 of the response variable (which is the AST enzyme) to 'yes' . It shows normal values and everything that is not within that range of values in the response variable is mapped to 'no'. However, since most machine learning algorithms can only take numerical values we went ahead and created dummy variables to map the 'yes' and 'no' to '1' and '0' respectively. This allowed us to set the problem into a classic binary classification problem.

	SGOT_AST	dummy_yes_no	dummy_no	dummy_yes
0	31	yes	0	1
1	11	yes	0	1
2	138	no	1	0
3	19	yes	0	1
4	32	yes	0	1
...	...	...	...	...
999995	18	yes	0	1
999996	21	yes	0	1
999997	19	yes	0	1
999998	27	yes	0	1
999999	26	yes	0	1

1000000 rows × 4 columns

```
[ ] y['dummy_yes_no'].value_counts()
```

```
yes    934316
no      65684
Name: dummy_yes_no, dtype: int64
```

Also, when we counted the values of the response variable we observed that the number of observations that show normal AST enzyme levels were significantly higher than those that have abnormal values. And for this reason when we initially trained our models we were getting between the higher nineties to a hundred percent accuracy on training and when we went to test our models on unseen data we were only able to see 30%-50% testing accuracy. So, we went back to the drawing board to perform gridsearch and cross-validation where the data is going to be shuffled each time before a different split of the sample is being used and sure enough

together with some hyper-parameter fine-tuning ceased our models from overfitting and began performing well on unseen data and as well as they did on training.

We split our dataset into 75% training and 25% testing and since it is a huge dataset which would require a more powerful computer to handle this problem within a reasonable amount of time, we decided to make a random sampling of a sizeable number that is not too large to give us issues but also not too small to be able paint us a brighter picture of the solution. For the Ensembles, we used Bagging, Boosting and Stacking. And then Individually, we used a Decision Tree Classifier, Random Forest Classifier and Logistic Regression.

Results

After determining our methods and preprocessing the data we began running both our weak learners and our ensemble models. We were pleased to see that the models performed quite well on the data subsets. For the 2009 data, the logistic regression training accuracy was 89.15% while the decision tree training accuracy was 94.66% and the random forest classifier training accuracy was 93.73%. The 2019 training accuracy was quite similar with logistic regression being almost the same while decision tree accuracy was slightly lower at 93.66 and random forest accuracy was also slightly lower at 93%.

2009 Training Set Accuracy:

```
[0] Logistic Regression Accuracy:  0.8914666666666666
[1] Decision Tree Accuracy:  0.9466666666666667
[2] Random Forest Classifier Accuracy:  0.9373333333333334
```

2019 Training Set Accuracy:

```
[0] Logistic Regression Accuracy:  0.8916
[1] Decision Tree Accuracy:  0.9366666666666666
[2] Random Forest Classifier Accuracy:  0.93
```

We then moved to create confusion matrices and calculate the test accuracy for the respective models for both 2009 and 2019. The results, displayed below, showed that the Decision Tree model performed with the highest accuracy for both years at around 85.2% and 87.4% respectively, while the logistic regression model performed the worst, however still at a respectable rate of ~81%. To evaluate the models and results, we decided to calculate the F1 scores of each model. For 2009, the Logistic Regression F1 score was .89, the Decision Tree F1 score was .92 and the Random Forest F1 score was .92. For 2014, the Logistic Regression F1 score was .88, the Decision Tree F1 score was .93 and the Random Forest F1 score was .92. All of these scores show models with a very high level of performance that were able to successfully capture and classify each observation.

2009 Test Set Accuracy:

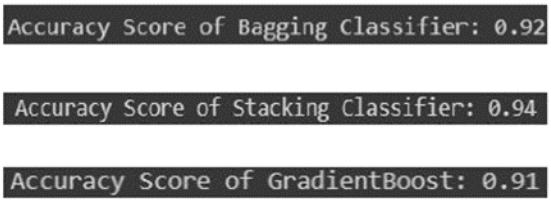
```
model 0
[[2311  189]
 [ 389 2111]]
Testing Accuracy =  0.8122157244964262

model 1
[[2392  108]
 [ 326 2174]]
Testing Accuracy =  0.8520790729379687

model 2
[[2407   93]
 [ 357 2143]]
Testing Accuracy =  0.847457627118644
```

2019 Test Set Accuracy:

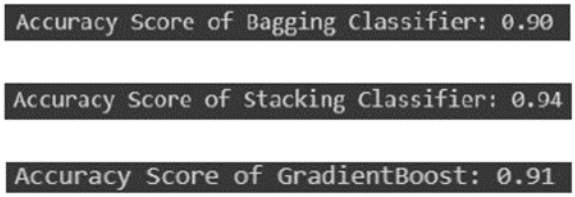
```
model 0
[[1125  132]
 [ 163 1080]]
Testing Accuracy =  0.8099226804123711

model 1
[[1151  106]
 [  75 1168]]
Testing Accuracy =  0.8741307371349096

model 2
[[1137  120]
 [  75 1168]]
Testing Accuracy =  0.8657024793388429
```

With the weak learners analyzed and evaluated, we moved on to our three ensemble models to see how they would perform. Initially with a small data subset, the ensemble models performed quite poorly and much worse than the individual models. However, after increasing the data subset size, we found the ensemble models produced very accurate predictions. As you can see below, all of the test accuracy scores were above 90% and the Stacking Classifier performed the best at 94% for both years. These high test accuracy scores clearly showed that our models were successfully able to predict the correct liver enzyme levels of SGOT_AST in a patient. In addition, because our models can accurately predict SGOT_AST levels, this indicates that we will also be able to predict with a high degree of accuracy whether or not a person has or will develop liver problems.

2009:



2019:



While these results are very promising, we did not evaluate them against other approaches, as other research incorporated techniques such as MRI, utilized other output variables that were not available in our dataset, or conducted studies in diseased populations.[4,7] Despite the lack of comparison studies, the confusion matrices and training/test accuracy scores demonstrate that the models are working as expected and predicting at a very accurate level. In addition, we have found that our given predictor variables act as accurate indicators of SGOT_AST levels. This is a very exciting development and could easily be implemented in a clinical setting to assist doctors in assessing possible liver problems.

Discussion and Future Work

Our goal was to use machine learning models that consider a variety of health indicators to accurately predict a patient's liver enzyme levels and therefore their probability of liver disease. Using this model in a clinical setting may allow doctors to predict the development of liver disease in patients. Treating liver disease early, or preventing it all together, will increase the health and longevity of patients. This model could ideally serve as another diagnostic tool for medical professionals. The model should be used in a clinical setting to verify its accuracy. No variable in the dataset confirmed the presence of liver disease in patients. We used the values of variables that have a known link to liver disease and used those as definitive indicators of liver disease. Being able to confirm liver disease would be helpful in determining the clinical accuracy

of our models. When working with a dataset this size, we had to spend some time cleaning the data and removing NA values. When it came time to train our models, we chose to select a sizable number of patients via random sampling to limit the required time to run the models. When working with this information we found that for this problem SGOT_AST was a strong indicator of liver disease and that the models we chose to work with were all fairly accurate.

The model is based entirely from data collected from patients across South Korea. South Korea's population is 50 million over an area of 100,000 km^2 for an average of 500 people per km^2. Seoul, South Korea has a population of just under 10 million - one fifth of the country's total population - over an area of 605 km^2 for an average of 16,528 people per km^2.[9,10] The increase in population density is significant. If we looked at the data for Seoul alone, would the additional health risks of living in such a densely-packed city environment increase the likelihood of liver disease? Our model does not currently consider the area code or region associated with the patient. If we were to link patients with their area codes, we might expect to see an increased number of liver disease cases in higher density area codes.

There are two well-known factors that increase the likelihood of liver disease for which our model did not account: alcohol consumption and cigarette smoking.[10,11] Alcohol consumption has a well-known negative effect on the liver. Adding a variable to the model that accounts for patients' alcohol consumption habits may better predict the likelihood of liver disease.
Smoking cigarettes or other tobacco products is another risk factor that is linked to liver health. The duration and frequency of a patient's smoking habitat should be incorporated into the model to improve accuracy.

Our model accounted for cholesterol levels of patients, and high cholesterol was associated with development of liver disease. How significant, however, is cholesterol level in

predicting the likelihood of liver disease? For example, some cultures (e.g., the United States) have relatively high-cholesterol diets. When looking at areas with high-cholesterol diets, would the increase in patients' HDL, LDL and total cholesterol lead to a higher frequency of liver disease, or would the increase result in our model overestimating the likelihood of liver disease in patients? We used AST, a liver enzyme, as our response variable. The use of alternative response variables that are powerful predictors of liver disease, such as GGT, may affect the accuracy of the model. Our model uses height and weight to determine obesity. Other indicators of obesity, such as BMI, and waist-to-hip ratio, may be more accurate variables.

References

1. Younassi ZM, Wong G, Anstee QM, Henry L. The Global Burden of Liver Disease. *Clinical Gastroenterology and Hepatology* **2023**; 21 :1978-1991.
2. Asrani S, Devarbhavi H, Eaton J, Kamath PS. Burden of Liver Diseases in the World. *J. of Hepatology* **2019**; 70:151-171.
3. Ray K. NAFLD - The Next Global Epidemic. *Nat Rev of Gastroenterol Hepatol.* **2013**; 10:621.
4. Ghadir MR, et al. The Relationship between Lipid Profile and Severity of Liver Damage in Cirrhotic Patients. *Hepat Mon*. **2010**; 10(4): 285-288.
5. McGill MR. The Past and Present of Serum Aminotransferases and the Future of Liver Injury Biomarkers. *EXCLI Journal* **2016**;15:817-828 (http://dx.doi.org/10.17179/excli2016-800).
6. Kathak RR, Sumon AH, Molla NH, Hasan M, Miah R, Tuba HR, Habib A, Ali N. The Association between Elevated Lipid Profile and Liver Enzymes: A Study on Bangladeshi Adults. *Scientific Reports* **2022**; 12:1711.
7. Bedogni G, Bellantani S, Miglioli L, Masutti F, Passalacqua M Castiglione A, Tirabelli C. The Fatty Liver Index: A Simple and Accurate Predictor of Hepatic Steatosis in the General Population. *BMC Gastroenterology* **2006;** 6:33.
8. South Korea country profile. BBC News. October 17, 2023. Retrieved October 17, 2023. (https://www.bbc.com/news/world-asia-pacific-15289563)
9. City Overview (Population). Seoul Metropolitan Government. Archived from the original on 26 November 2021. Retrieved 26 November 2021.

10. Maher JJ. Exploring alcohol's effects on liver function. Alcohol Health Res World. 1997;21(1):5-12
11. Abdul-Razaq, Sangar Najat, and Bakhtiar M. Ahmed. "Effect of cigarette smoking on liver function test and some other related parameters." Zanco Journal of Medical Sciences (Zanco J Med Sci) 17.3 (2013): 556-562.