American University of Sharjah
Department of Mathematics & Statistics
STA401 – Introduction to Data Mining
Instructor: Dr. Ayman Alzaatreh

Fall 2021
Final Project Report

Using ANNs for Hepatitis C Diagnosis: A Novel Approach of Variable Selection in Neural Networks

Submitted by:
Ahmad A. Bilal - B00077899
Dara M. Varam - B00081313

Submission Date: 5th December 2021

Abstract

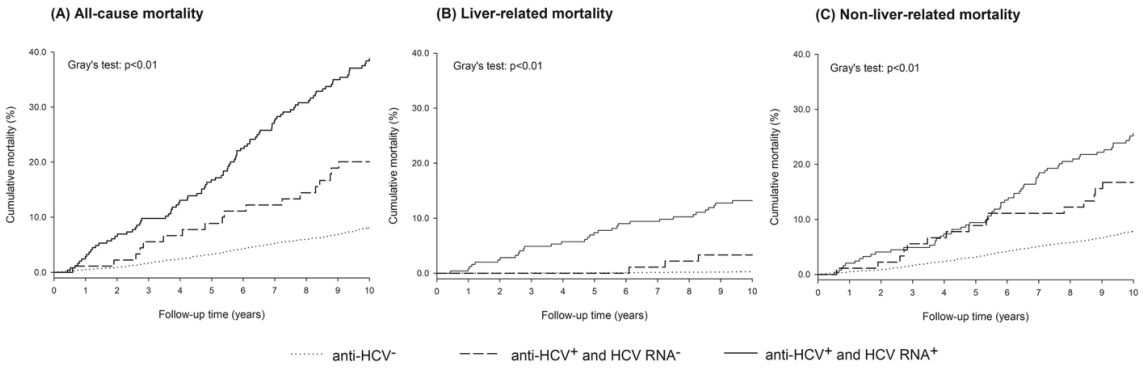
This paper covers the prediction and diagnosis of Hepatitis C in patients using a dataset containing said-patients' blood test results. It aims at comparing different neural network models based on variable selection methods that have, to the best of our knowledge, never been employed in this specific field. The variable selection process consists of pre-processing the data and putting it through three different machine learning techniques: K-nearest neighbor method, random forest method, and decision tree method. Upon processing the data through these three models, the variables with highest importance are selected through an average of their importance through each respective technique, and five different neural network models are built using these selected variables. The inspiration behind using this method for variable selection was because of the research done on the shortcomings of traditional machine learning techniques and the benefits of the newly emerging "hybrid" neural network architectures. The results indicate that this approach to selecting variables results in little to no difference in the accuracy of the neural network (when compared to the full model, with all the variables present) and we can easily obtain accuracies of over 90% in models with reduced variables.

Keywords: Neural Networks (ANNs), Variable Selection for Neural Networks, Hepatitis C diagnosis / prediction, Machine Learning, Neural Networks for healthcare diagnosis.

1. Introduction

It is estimated that around 354 million people currently live with chronic hepatitis B and C globally [1]. Hepatitis is a virus that infects the liver and causes it to inflate. The inflammation prevents the liver from carrying out its vital roles in the body, such as the production of certain proteins for blood plasma, production of cholesterol and special proteins to help carry fats through the body, and removing waste and breaking down fats in the small intestine. This highlights the importance of the liver and the severity of hepatitis. There are many factors that can lead to hepatitis, which could range from an overconsumption of alcohol to certain medical conditions or exposure to the virus. Hepatitis can be categorized into specific subcategories, the most common ones being hepatitis A, hepatitis B, and hepatitis C. They are all liver infections that are caused by three different viruses, and despite sharing similarities in the symptoms that they produce, they spread differently and affect the liver differently. For example, hepatitis A usually causes short-term infections, whereas hepatitis B and C usually begin as a short-term infection but can remain in the body to develop into a chronic infection. Moreover, there are currently vaccines that can help prevent hepatitis A and B but there is no vaccine for hepatitis C.

People who are diagnosed with hepatitis C might suffer from mild illness lasting a couple of weeks or they can develop long-term infection from the virus. In fact, more than 50% of people who get infected with hepatitis C virus develop a chronic infection [2]. It is difficult to deal with hepatitis C virus since some people might not develop symptoms and most people are not aware of the commonality of the virus. Indeed, symptoms of chronic viral hepatitis can take decades to develop [2]. Moreover, around 2.4 million people were living with hepatitis in 2018 [2]. Considering the severity of hepatitis C, this places the previous statistics in a new light. For example, if left untreated or in exacerbated cases, hepatitis C can lead to long-term health problems such as liver damage, failure, or cancer. Moreover, hepatitis C can even lead to death in some cases, as illustrated in the figure below, where there it can be for a number of reasons. In 2019 alone, there were 1.1 million reported death cases that were due to infections of hepatitis B and C and their effects [3]. To this end, despite there being no vaccine for hepatitis C, it can be treated or cured but it is economically costly. The World Health Organization estimates that to reduced hepatitis infection by 90% and deaths by 65%, 58.7 billion US dollars is needed eliminate viral hepatitis disease as a public threat from 67 low- and middle-income countries by 2030 [4]. Furthermore, treatment can even be expensive on an individual basis/scope. A study conducted in 2018 found that a single pill of one hepatitis C drug costs $1,000, and the total cost for a 12-week course treatment would cost the average patient around $84,000 [5]. Furthermore, another drug used to treat hepatitis C was found to cost around $24,000 per month for a treatment plan that could take 6 to 12 months, which brings up the total to around $288,000 for a one-year treatment plan [5]. However, if the disease was diagnosed and detected early on, then it can be greatly beneficial as there are highly effective treatments that can be used to treat approximately 95% of patients who are diagnosed early [6]. These findings highlight the importance of early detection in hepatitis C and the ever rising need to optimize current tools and measures to battle the virus.



2. Related Works

In recent years, the medical field has been integrating the use of machine-learning and data mining techniques into their research to help them optimize pre-existing tools to extract information from different specific datasets. Regarding hepatitis C, there have been many attempts at finding forms of detection with the use of artificial neural networks (ANN) in previous literature, some of which can be used to make health care organizational decision-making as demonstrated in [7]. Most of the previous literature discusses the use of different neural networks architectures with different properties and characteristics to assess their performance and use in real-life application. For example, some of the differing characteristics include the use of supervised and unsupervised ANN, where each one is can be further categorized into a sub-category. Some of the supervised category include the Feedforward Backpropagation Neural Network (FFNN), Generalized Regression Neural Network (GRNN) and Decision Trees, whereas the unsupervised category includes Self Organizing Map (SOM). There are certain predictors that can be used in machine learning algorithms to predict hepatitis C in patients, which include: sex, age, platelet, albumin, and aspartate aminotransferase, etc. There have been seven thoroughly reviewed literature works considered when working on this report.

In [8], the donors of the UCI dataset that we used, employed decision trees as a machine learning technique to predict diagnostic pathways for hepatitis C in patients. However, in their findings the authors note that the performance of the decision trees was well but minor changes led to "markedly deviant decision trees." This is another way of saying that it performed poorly in general, especially in comparison to an ANN architecture as discussed in previous literature such in [9] [10] [11] [12].

In [9], the authors attempted to use SOM for the diagnosis of hepatitis B in a UCI dataset, but it proved to be inefficient. [9] report that the SOM network was diagnose the hepatitis virus correctly as it required a greater number of examples for training to contain the target set. Moreover, the network was not able to handle the entire data inputs presented to it, whereas the FFNN and GRNN yielded almost a 91% accuracy.

In [10], the authors have designed and implemented ANN and rule based models for patients with normal and infected students. They concluded that the prediction of ANN-based models was specific and accurate than rule-based or unsupervised methods. This aligns with the previous literature, since even the past ANN models have been able to be more self-adaptive and learn with each additional input provided to it as described in [10]. One limitation in this literature, as the authors mention, is that the dataset used included a very small and limited number of patients, but strongly suggest that such models that can predict Hepatitis C can assist doctors in the diagnosis of the virus.

Moving on to [11], the authors used Support Vector Machine (SVM) and ANN to diagnose hepatitis B and hepatitis C. The authors considered 6 classes: hepatitis B (two phases), hepatitis C (two phases), non-viral hepatitis and no-hepatitis. Therefore, they designed various networks such as RBF, GRNN, SVM, and two more. [11] report that the RBF network outperforms the rest of the networks, yielding an overall accuracy of over 96.4% for train and test data. This reinforces [12] findings and further supports their points of the efficiency of an RBF network in hepatitis diagnosis and the high potential that a hybrid neural network may have. Indeed, we can see from previous literature that a hybrid neural network that extracts the best features from various networks and builds further on those shortcomings can yield the best results.

In [12], the authors report on the efficiency of using a hybrid network in the diagnosis of hepatitis. Multilayer Perceptron (MLP) architecture showed poor performance due to the low classification accuracy. Moreover, despite the Radial Basis Function (RBF) network showing promising results, but [12] emphasis that the use of a Conic Section Function Neural Network (CSFNN) that combines RBF and MLP is more reliable for an accurate for the diagnosis. This work suggests that the use of hybrid networks is more efficient and successful for diagnosing hepatitis, which was an inspiration for our own

approach in this project. All in all, all previous literature reviewed suggest that the use of an ANN architecture, in specific a hybrid model, yields the best results as they build upon each other. Henceforth, increasing accuracy, reliability, and its potential to be integrated a real-world application. This can be assessed by comparing the performance of different models and architectures and improving upon them, as we have done and demonstrated further down in this report.

3. Data Exploration

The dataset we are using is the HCV Hepatitis patients dataset donated by Lichtinghagen, Klawonn and Hoffman. It is accessible through https://archive.ics.uci.edu/ml/datasets/HCV+data. The dataset contains data related to a hepatitis C case study of 614 patients. The patients have been categorized into 5 different groups: Blood donors, Suspected blood donors, Hepatitis C patients, Fibrosis patients, and finally, Cirrhosis patients. The dataset contains the following predictor variables, which we will be using in order to predict whether a patient is healthy or unhealthy, and more specifically, which liver tissue disease they are suffering from (if any). The variables are as follows:

- Age
- Sex
- ALB (Albumin Blood Test)
- ALP (Alkaline phosphatase)
- ALT (Alkaline Transaminase)
- BIL (Bilirubin)
- CHE (Acetylcholinesterase)
- CHOL (Cholesterol)
- CREA (Creatinine)
- GGT (Gamma-Glutamyl Transferase)
- PROT (Proteins)

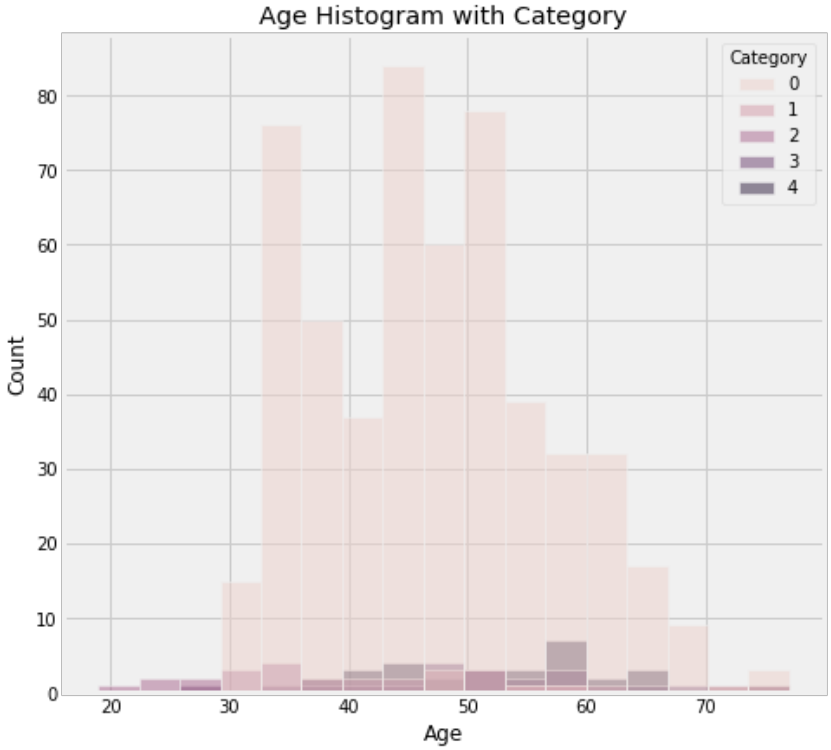
The data notably consists of laboratory test results for each of the patients, which is what we will primarily be using to diagnose the type of liver tissue damage a patient is suffering from. The following table is an indicative summary of the data as a whole:

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
0	0	32	1	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
1	0	32	1	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
2	0	32	1	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
3	0	32	1	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
4	0	32	1	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7

More information regarding the dataset (including the standard deviations of each of the variables, etc.) can be found in the table below. Note that the table has the N/A values omitted for the sake of providing accurate statistical measures.

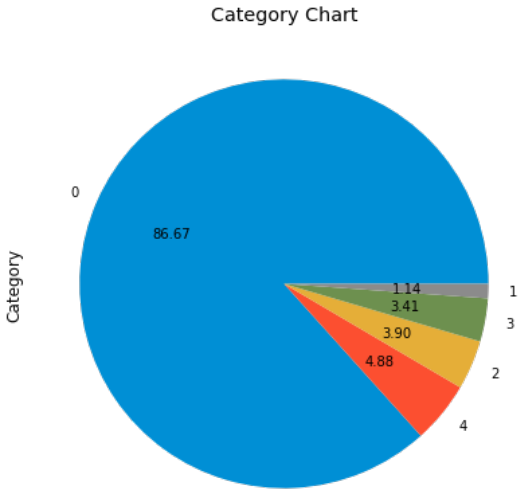
	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
count	615.000000	614.000000	597.000000	614.000000	615.000000	615.000000	615.000000	605.000000	615.000000	615.000000	614.000000
mean	47.408130	41.620195	68.283920	28.450814	34.786341	11.396748	8.196634	5.368099	81.287805	39.533171	72.044137
std	10.055105	5.780629	26.028315	25.469689	33.090690	19.673150	2.205657	1.132728	49.756166	54.661071	5.402636
min	19.000000	14.900000	11.300000	0.900000	10.600000	0.800000	1.420000	1.430000	8.000000	4.500000	44.800000
25%	39.000000	38.800000	52.500000	16.400000	21.600000	5.300000	6.935000	4.610000	67.000000	15.700000	69.300000
50%	47.000000	41.950000	66.200000	23.000000	25.900000	7.300000	8.260000	5.300000	77.000000	23.300000	72.200000
75%	54.000000	45.200000	80.100000	33.075000	32.900000	11.200000	9.590000	6.060000	88.000000	40.200000	75.400000
max	77.000000	82.200000	416.600000	325.300000	324.000000	254.000000	16.410000	9.670000	1079.100000	650.900000	90.000000

The main purpose behind looking at such a summary is to gain a better understanding of the scale and type of data that we are working with. This will allow us to have some sort of indication as to what direction we could take the data and the prediction models is. The first area that we decided to look into further was whether or not there was some degree of correlation between the age of a patient and their healthiness. This is based purely on initial hypotheses that we formed and a simple observation of the dataset values. The following histogram can be used to see whether this correlation is in-fact something to look further into:



Given that the age takes somewhat of a normal form, it appears as though the existence of liver tissue damage follows in the same normalized fashion. However, there could be some discrepancy in this observation, given that this is formed mainly on the basis that the majority of the patients are patients that fall into the middle third of the data in terms of age. If we take this into consideration, there does not seem to be much of a correlation between the age of a patient and their healthiness in this specific application. This is something that we will look further into when building the neural network models (removing the age as a predictor could potentially increase the accuracy, or at least provide the same accuracy with less inputs).

Accounting for the patients that fall into each of the 5 categories available in this dataset, we can count the number of instances for each type of damage (or non-damage):



It is evident that the data is, in fact, incredibly unbalanced. Almost 87% of the data is patients that fall into the "Blood Donor" category, giving us little to no data left for patients that suffer from each of the possible diseases. We will need to account for this imbalance when building the models and cleaning the data, which will be done in the "Data Cleaning" section of this report.

4. Data Transformation

Since the data outcome is categorical, we first needed to code this into the program such that each index ranged from 0 to 4 represents one of the categories. We thus came up with the following conversion table:

- 0: Blood donor
- 1: Suspected Blood donor
- 2: Hepatitis C patient
- 3: Fibrosis patient
- 4: Cirrhosis patient

This allows for the categorical implementation of classification-based neural network models – the objective of this project.

The only other transformation that needed to be conducted for the data is another encoding problem: coding the gender to be 1 and 0 in accordance with male and female attributes. This is once again used for the sake of implementing the gender of the patient as part of the modeling.

- 0: Female, count = 238
- 1: Male, count = 377

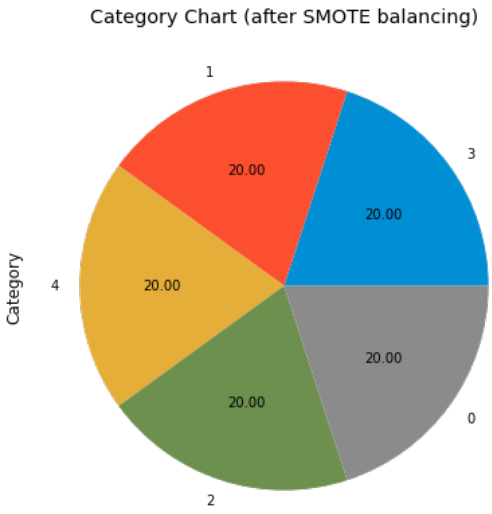
5. Data Cleaning

The first step in cleaning our data is to remove datapoints that are not provided and instead replace them with the average of that specific predictor. We decided to use this rather than simply omitting the variables because we would lose too much data – especially for the categories that have much less data in comparison to the rest. We can then get the following description table, now with no values omitted.

	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
count	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000	615.000000
mean	47.408130	41.620195	68.283920	28.450814	34.786341	11.396748	8.196634	5.368099	81.287805	39.533171	72.044137
std	10.055105	5.775920	25.643955	25.448940	33.090690	19.673150	2.205657	1.123466	49.756166	54.661071	5.398234
min	19.000000	14.900000	11.300000	0.900000	10.600000	0.800000	1.420000	1.430000	8.000000	4.500000	44.800000
25%	39.000000	38.800000	52.950000	16.400000	21.600000	5.300000	6.935000	4.620000	67.000000	15.700000	69.300000
50%	47.000000	41.900000	66.700000	23.000000	25.900000	7.300000	8.260000	5.310000	77.000000	23.300000	72.200000
75%	54.000000	45.200000	79.300000	33.050000	32.900000	11.200000	9.590000	6.055000	88.000000	40.200000	75.400000
max	77.000000	82.200000	416.600000	325.300000	324.000000	254.000000	16.410000	9.670000	1079.100000	650.900000	90.000000

As previously mentioned, the data is also highly unbalanced, which will significantly affect our finalized models if not dealt with. For this task, we decided to oversample the categories with little data so as to match the quantity of the largest category (healthy patients / blood donors). The reason why we chose to oversample the smaller categories is to preserve the data contained in the larger data categories. This allows us to retain as much information as possible while also balancing the data as per our requirements.

The specific method used for oversampling our data is the Synthetic Minority Oversampling Technique (SMOTE). This is a popular oversampling method used in classification problems and works on the basis of the K-nearest neighbors' algorithm. Essentially, SMOTE allows us to create synthetic and / or artificial data points in between sets of K datapoints, which considers the distance between the center of these sets of data as the metric for distinguishing between them. The artificial datapoints are added to the minority categories until they are equal in size to the largest category. Upon implementing SMOTE for our data, we get the rebalanced dataset indicated by the below pie chart:



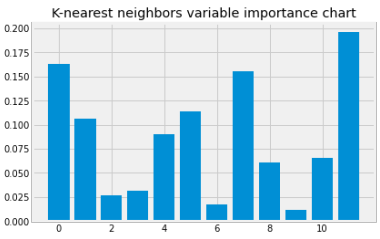
We can now see that each of the categories contributes to exactly 20% of the total data, with each containing a total of 533 datapoints.

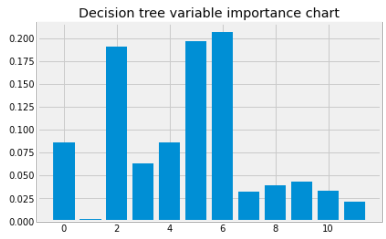
With this, our data is pre-processed and ready to be implemented in the neural network models that we will be using. The following section considers the methodology in conducting the tests, along with how we obtained the different variables that we will be using in the neural network modeling.

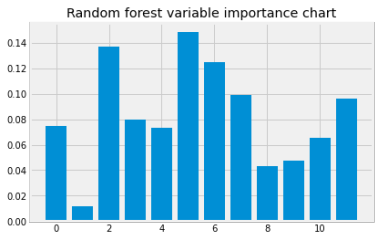
6. Methodology

In the scope of data mining, neural networks (or artificial neural networks) are collections of algorithms that closely follow the structure of the human brain in the way it comprehends pattern recognition, deduction, and beyond. In this specific application, we use neural networks to deduce relationships between the predictors mentioned in the data exploration section and the outcome (whether a patient is healthy, has hepatitis, fibrosis, or cirrhosis). This could also be converted into a binary problem exploring whether a patient is healthy or unhealthy in the context of liver tissue damage. Since neural networks are completely unsupervised, the main objective of the project is to allow the program to deduce said-relationships between predictors and the outcome. We will be choosing five (5) different models to run using neural networks, comparing and contrasting between each to eventually decide which model is best in predicting the existence of live tissue damage within a patient. This can then be applied in the field and beyond for potential early recognition of liver tissue damage based solely on factors identified within a patient's blood samples. The challenge of the project, however, was the feature selection that was associated with each of the five models.

The models were chosen by considering the variables which contribute most to the prediction outcome of pre-established, well-known data mining techniques. Specifically, given the categorical nature of our dataset, we decided to use the following three data-mining techniques: K-nearest neighbors, decision tree modeling, and random forest modeling. For each of these data mining techniques, the variable importance metric was considered, which is a metric indicating the "importance" of a variable in providing the prediction outcome in the given model. For example, a variable having an importance of 0.15 means that that specific variable contributes a maximum of 15% in the actual prediction outcome. Below are the bar charts that show the importance ranking of each of the aforementioned methods:



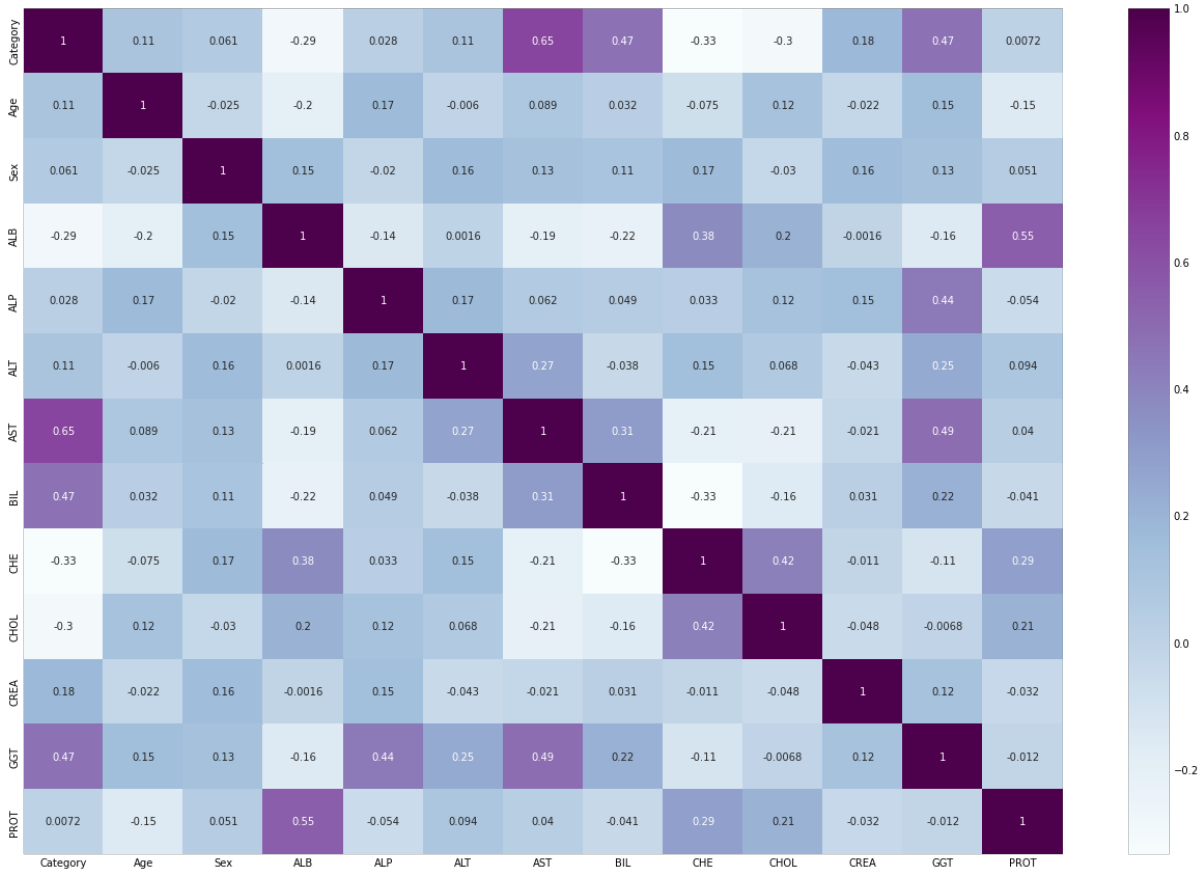




Given these values, we can attribute a score to each of the variables and their importance, calculated based on the average of each. The table below describes the averages for each of the variables.

Variable	Importance Score
AGE (VAR 0)	0.10770947635466771
SEX (VAR 1)	0.03989632374876976
ALB (VAR 2)	0.11816277579941716
ALP (VAR 3)	0.05790419266882033
ALT (VAR 4)	0.08338147526141615
AST (VAR 5)	0.15291713605933596
BIL (VAR 6)	0.11608543460118827
CHE (VAR 7)	0.09546399895371337
CHOL (VAR 8)	0.04772722579894203
CREA (VAR 9)	0.033844343496945296
GGT (VAR 10)	0.054781737475094495
PROT (VAR 11)	0.1045298118639057

In addition, we wanted to create a potential "simplest model" (least number of input variables), which was done by considering the correlation matrix of the variables. We can see the correlation matrix below:

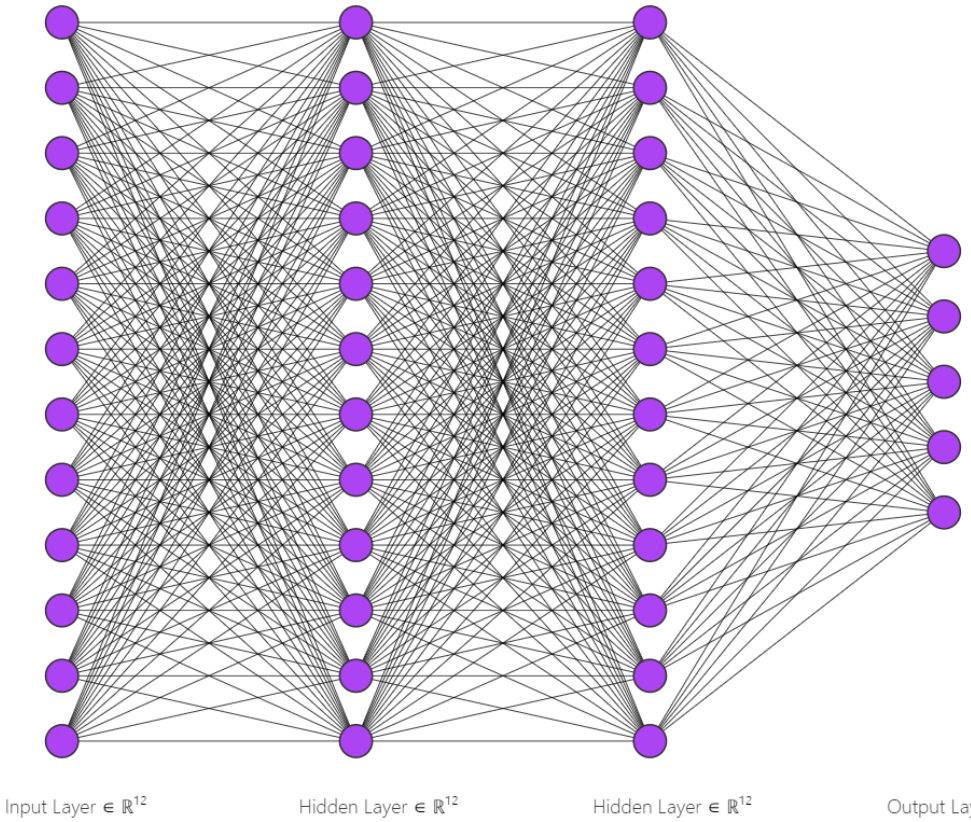


Based on this correlation matrix, we can see that the variables that have the highest correlation with the outcome variable "Category" are the following: AST, BIL, and GGT. This will serve as our simplest model.

By taking into account both the "importance scores" and the correlation matrix, we can come up with the following 5 models, which we will use in our neural network predictions:

- Model 1: Full model (all 12 variables are included)
- Model 2: "CREA," "CHOL" and "GGT" are dropped from the model
- Model 3: Simplest model (only the variables "AST," "BIL," and "GGT" are included)
- Model 4: Age is dropped as an input variable (all predictors except "AGE")
- Model 5: "ALP," "ALT," "GGT" and "CHOL" are dropped from the model

Finally, the neural network architecture that we will be using is going to be is a simple, feedforward neural network with 2 dense layers (other than the input layer and the output layer [5 neurons representing the 5 possible outputs ranged 0 to 4]). We decided on this architecture experimentally – anything beyond this would not contribute to increasing the accuracy. Given the number of input variables, the network architecture is provided in the figure below, along with a summary of the base model architecture (full, 12 input variables used).




```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 12)                156

 dense_1 (Dense)             (None, 12)                156

 dense_2 (Dense)             (None, 5)                 65

=================================================================
Total params: 377
Trainable params: 377
Non-trainable params: 0
_________________________________________________________________
```

The remainder of the network architectures are very similar to the one visualized above – except the input layer will have varying dimensions based on the model itself (model 2 will have 9 input neurons, model 3 will have only 3, and so on…).

7. Results

For each model, the dataset was split into the training and the testing dataset, with the neural network model running on the training set to then be evaluated in the testing set. When considering how good a model is, there are a few important metrics that we need to look into. The bulk of these metrics originate from the classification matrix, a commonly used statistical analysis tool. From the classification matrix, we can obtain results such as the F1 score, the precision and recall scores, the accuracy of the model, and so on. These are the primary metrics that we can use in evaluating how "good" a model is. However, it is important to note that the most important metric is ultimately the accuracy of the model, which is

a percentage basis indication of whether the model is accurately pairing the datapoints with their respective category. Furthermore, another area worth noting in the evaluation of a neural network model is the loss and accuracy curves per epoch (or per unit training run of the model). This gives us an indication of whether the model is overfitting or underfitting the data, allowing us to tune it further for the best possible results. Therefore, we have the following areas of consideration for the results:

- Model accuracy (and model loss) per unit training run;
- Training data prediction (using classification matrix)
- Testing (unseen) data prediction (using classification matrix)
- Classification report (precision, recall & F1 score)

Note that the training data classification matrix is used more as a supplementary tool to see whether the model works well on the seen data. On the other hand, the testing data classification matrix is where we draw the majority of our conclusions on how good the model is. Additionally, the following formulae can be useful in understanding the precision, recall and F1 score metrics:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

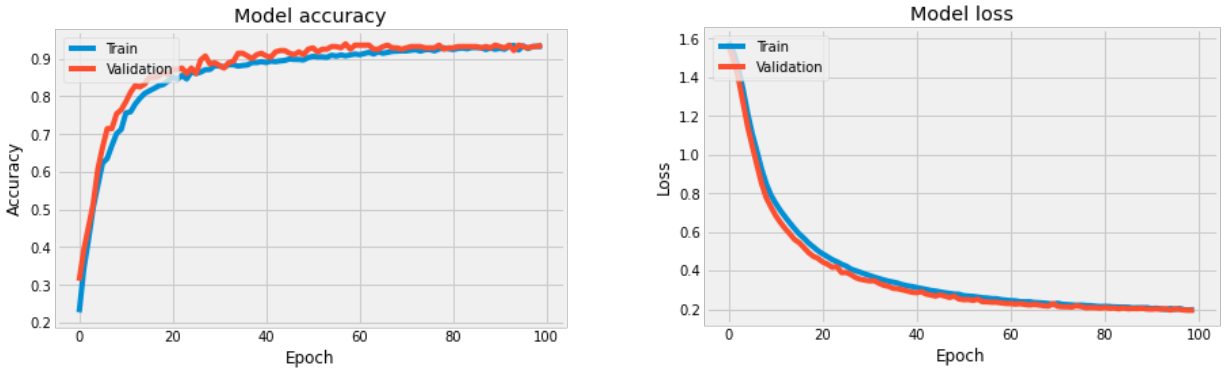
$$\text{F1} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}}$$

In the grander scheme, the F1 score ends up being the best indicator as to how well a model is performing, although these are all once again extensions of how accurate the model is.

Finally, it is also important to highlight the metrics that weren't considered in this project due to limitations on time and resources. Primarily these are the K-nearest neighbors metrics and the ROC curves. When considering the application of the K-nearest neighbors metric in the results of a neural network model, we are talking about taking different samples from within the testing and training sets and resampling (with replacement) to obtain more sets to test and train the data on. Furthermore, the ROC curves (and by extension, the area under curve (AUC) metrics) are also useful in providing insight for our data since the curves plot the true positive rate against the false positive rate.

a. Model 1 (Full model, no variables dropped)

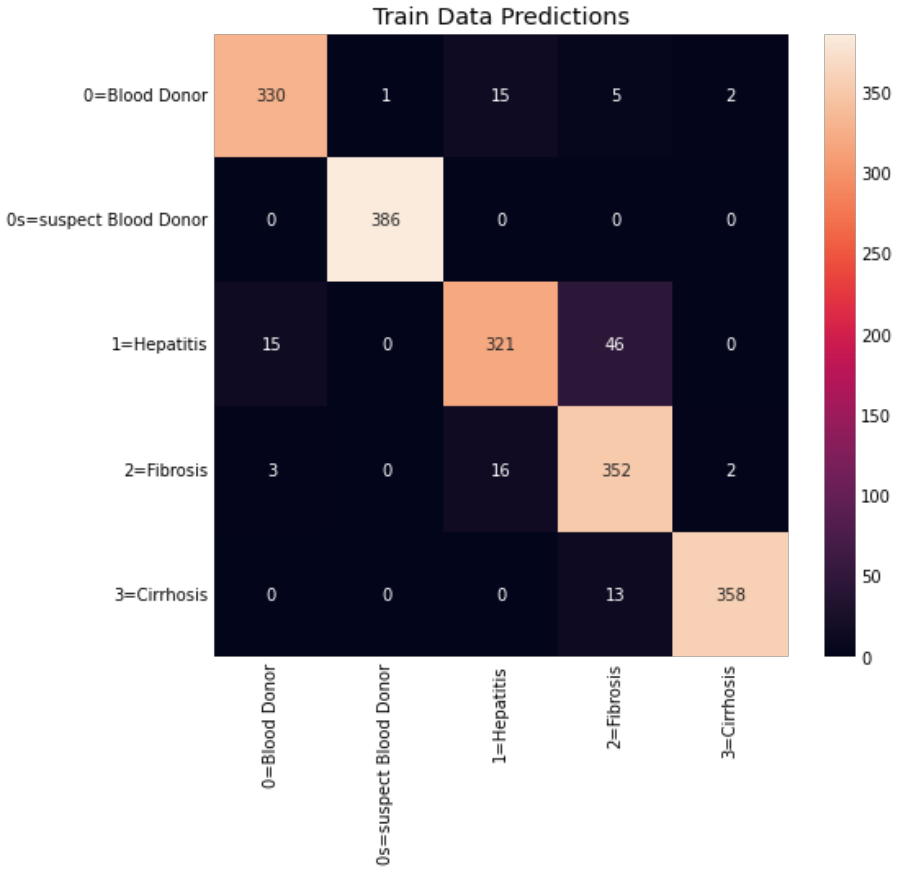
The model's accuracy and loss (for training and validation are shown in the figures below:



We can see that the accuracy increases (tangential to a value of around 0.9) as we increase the epoch, and by plotting it alongside the validation set (15% of the training dataset), we can see that there is no evidence of the model overfitting or underfitting the data. This is a good indication for validating the

accuracy of the model. The same can be said for the model loss: the loss decreases as we increase the epoch (training run). The value of the loss is tangential to 0.2. Once again, we have little to no evidence of over or underfitting in the model itself.

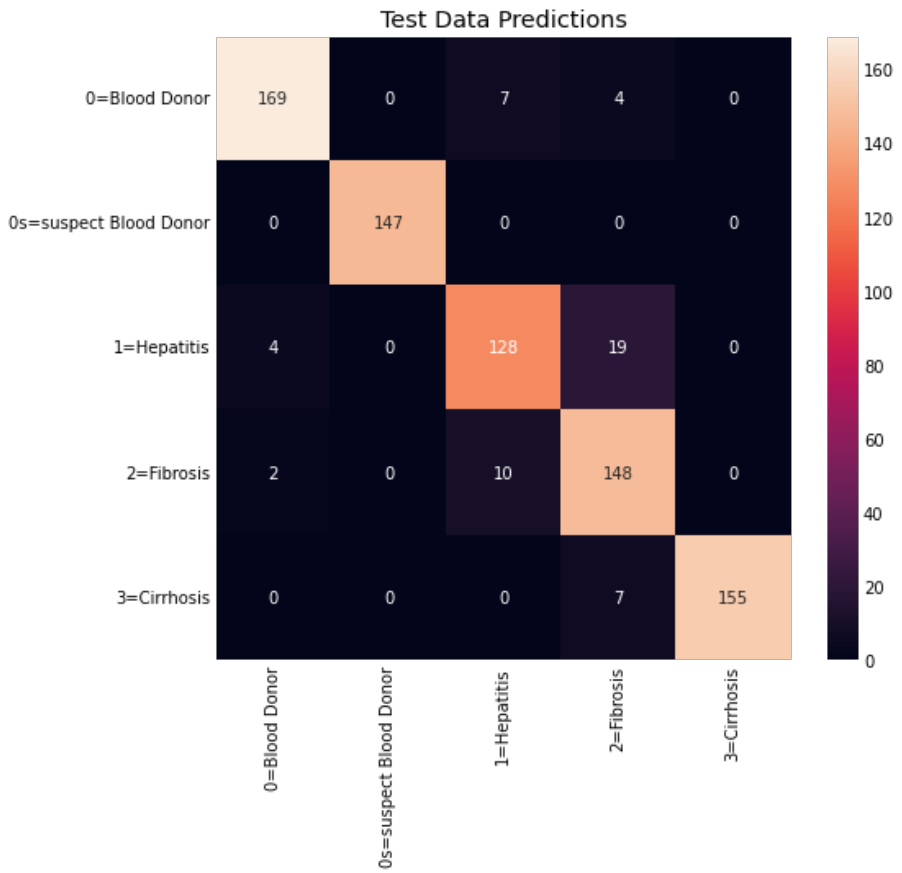
The model's training classification matrix is shown below:



The heatmap shows that the majority of the datapoints are along the main diagonal – an indication of the accuracy of the model. We only have a few outliers in each category. Although this is a good visual indication of the model's accuracy, we can get more valuable information out of the classification report associated with the above matrix:

```
Classification Report
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       353
           1       1.00      1.00      1.00       386
           2       0.91      0.84      0.87       382
           3       0.85      0.94      0.89       373
           4       0.99      0.96      0.98       371

    accuracy                           0.94      1865
   macro avg       0.94      0.94      0.94      1865
weighted avg       0.94      0.94      0.94      1865
```

The classification report indicates an F1 score of 94% for the training data, which is considerably high. However, we need to report on the results for the testing dataset:



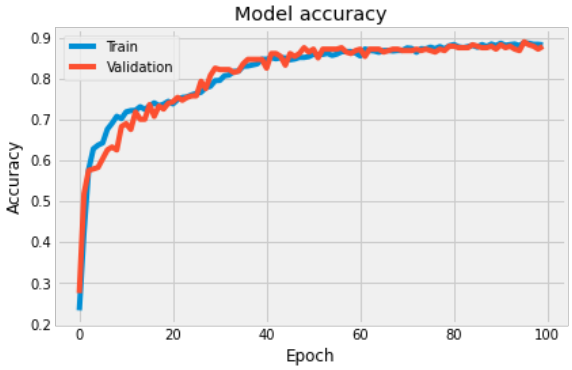
We can once again visually see that the majority of the unseen (test) datapoints are along the main diagonal. Looking at the classification report:

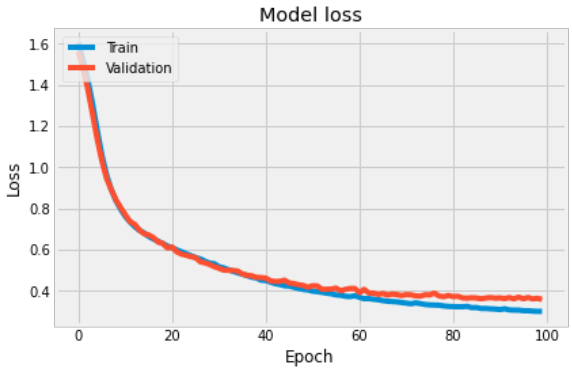
```
Classification Report
              precision    recall  f1-score   support

           0       0.97      0.94      0.95       180
           1       1.00      1.00      1.00       147
           2       0.88      0.85      0.86       151
           3       0.83      0.93      0.88       160
           4       1.00      0.96      0.98       162

    accuracy                           0.93       800
   macro avg       0.94      0.93      0.93       800
weighted avg       0.94      0.93      0.93       800
```

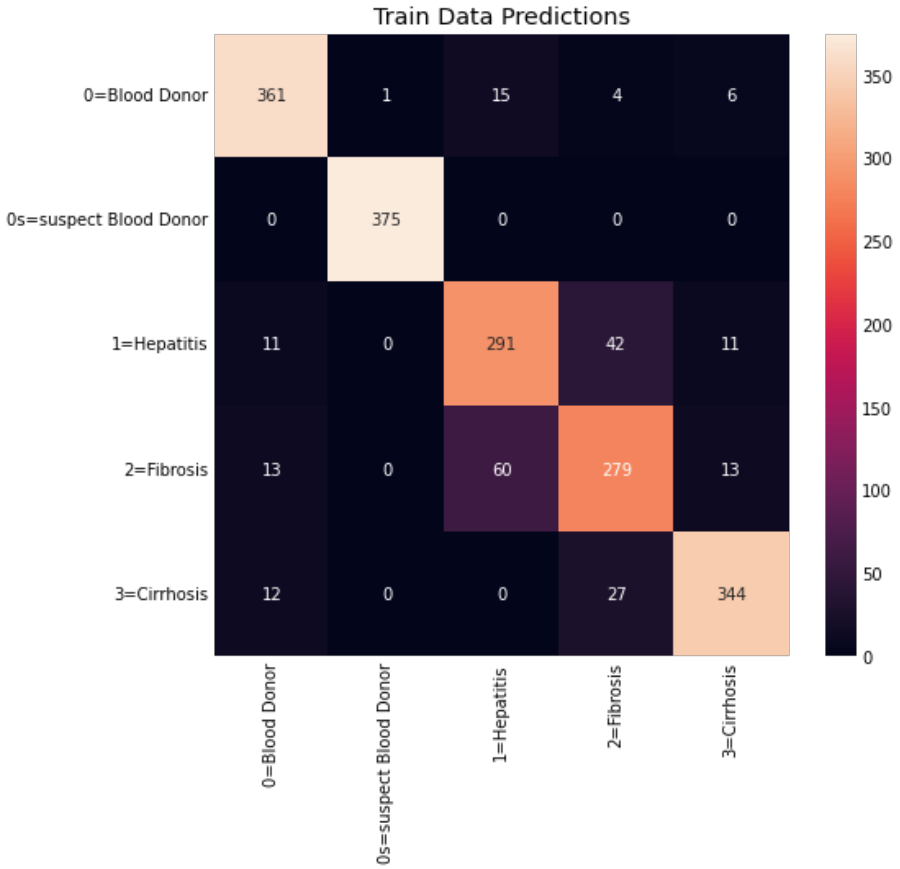
Evidently, the F1 score accuracy of this model, when evaluated against unseen data, is 93%. This is lower than the training data F1 accuracy – which is to be expected.

Therefore, the final evaluation of this model, which contains **all** variables included in the dataset, is an F1 accuracy of **93%**.

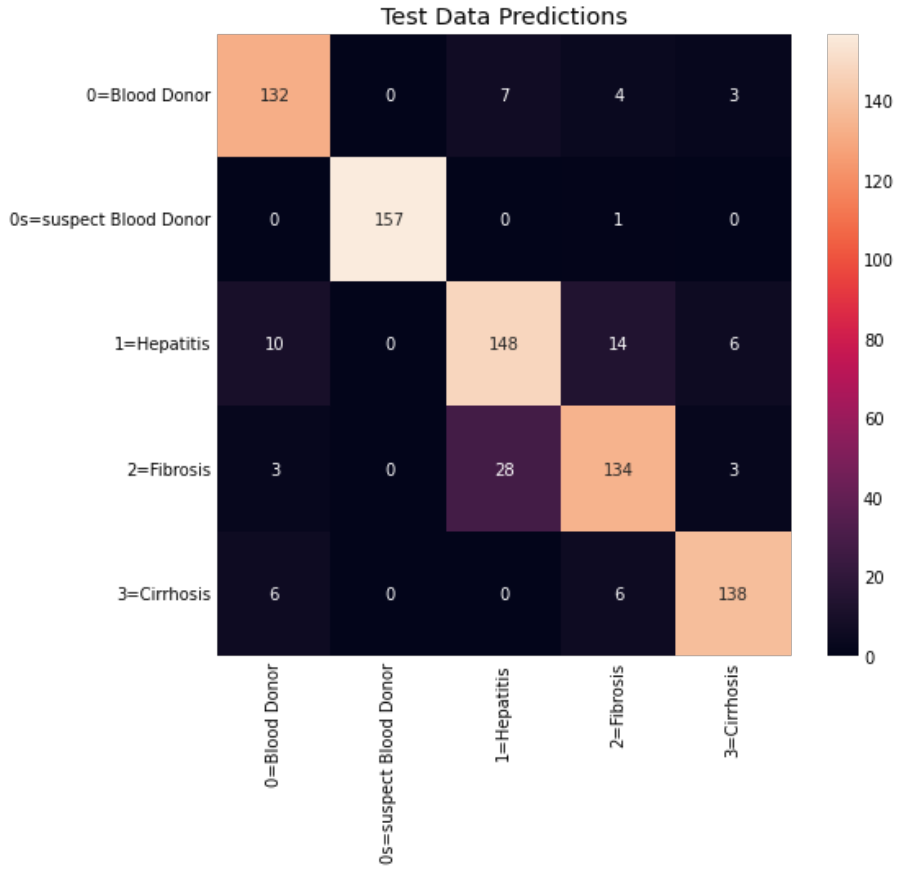
b. Model 2 (“CREA,” “CHOL” and “GGT” are dropped)







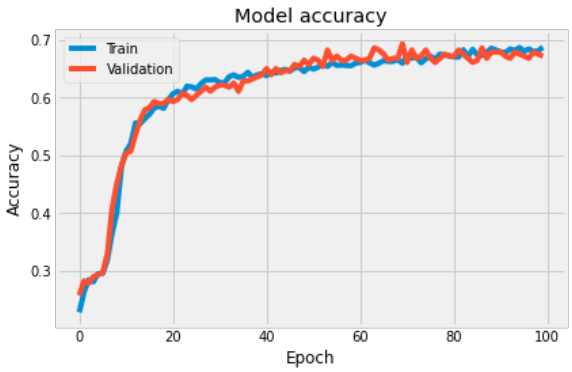

```
Classification Report
              precision    recall  f1-score   support

           0       0.91      0.93      0.92       387
           1       1.00      1.00      1.00       375
           2       0.80      0.82      0.81       355
           3       0.79      0.76      0.78       365
           4       0.92      0.90      0.91       383

    accuracy                           0.88      1865
   macro avg       0.88      0.88      0.88      1865
weighted avg       0.88      0.88      0.88      1865
```

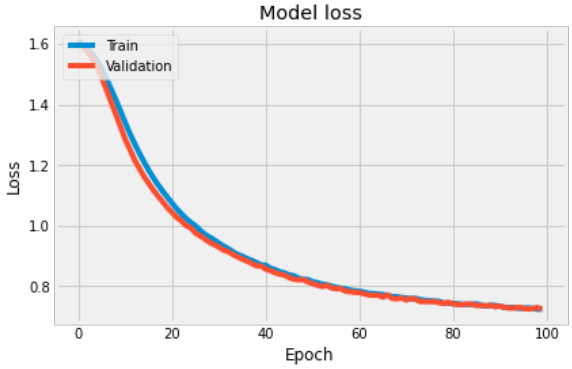


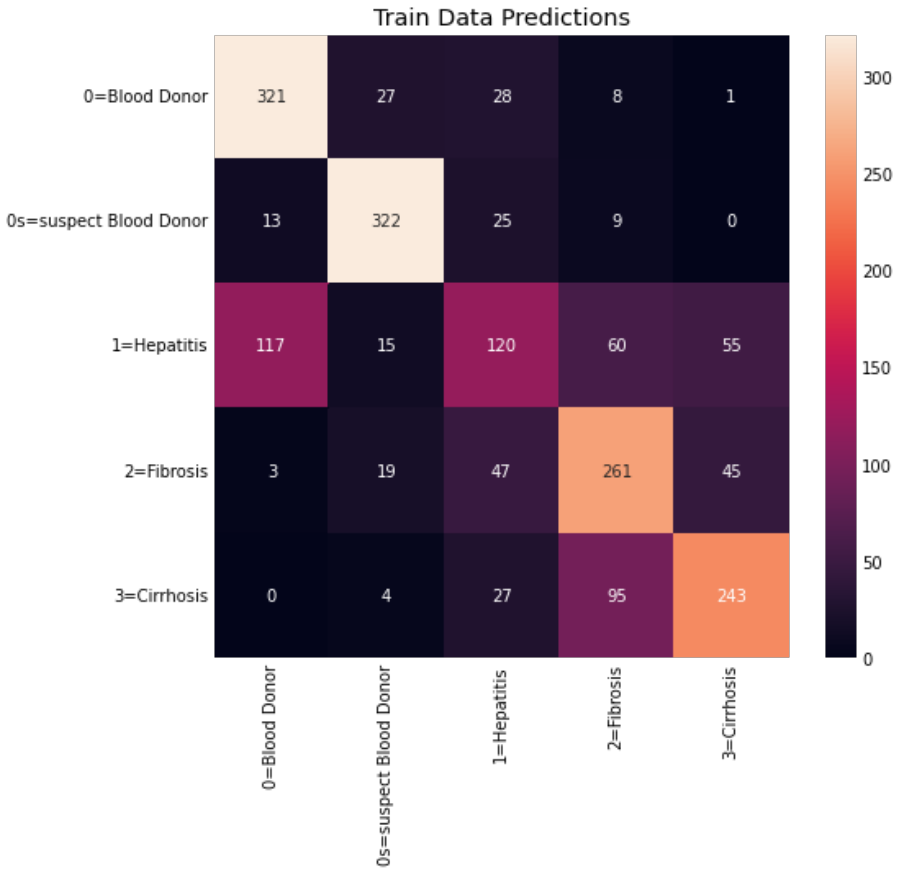

```
Classification Report
              precision    recall  f1-score   support

           0       0.87      0.90      0.89       146
           1       1.00      0.99      1.00       158
           2       0.81      0.83      0.82       178
           3       0.84      0.80      0.82       168
           4       0.92      0.92      0.92       150

    accuracy                           0.89       800
   macro avg       0.89      0.89      0.89       800
weighted avg       0.89      0.89      0.89       800
```

Therefore, for the second model, which includes all the variables except `CREA`, `CHOL` and `GGT`, we can report an F1 accuracy of **89%**.

c. Model 3 (All variables dropped <u>except</u> "AST," "GGT" and "BIL")

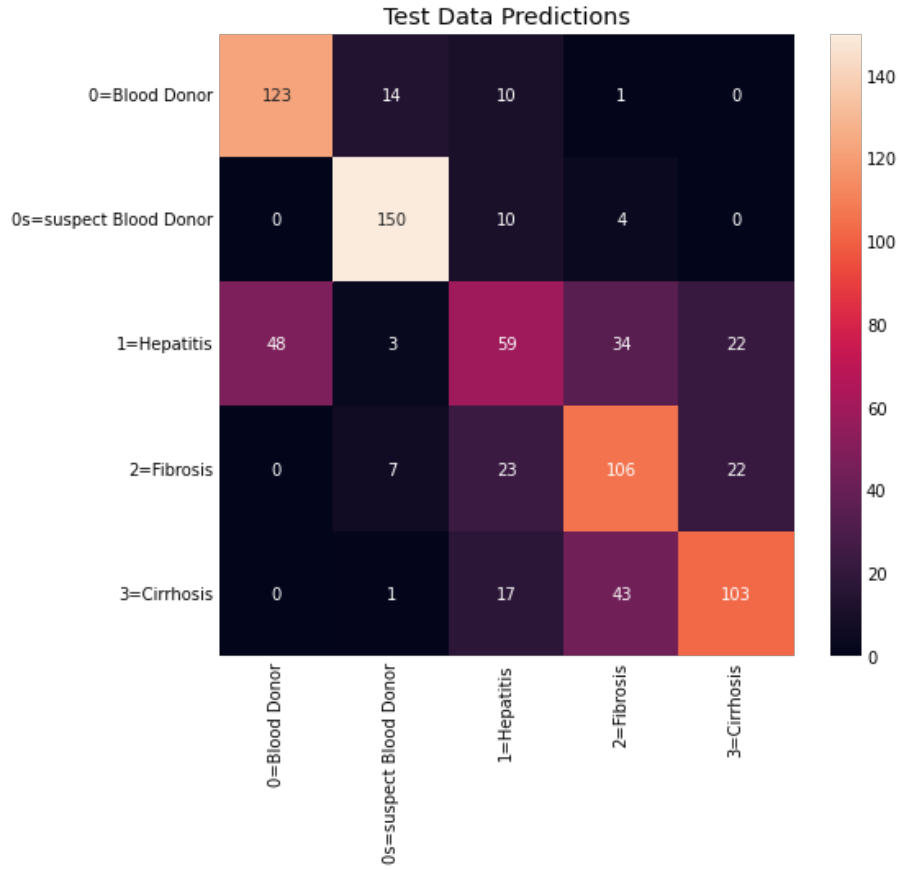






Classification Report

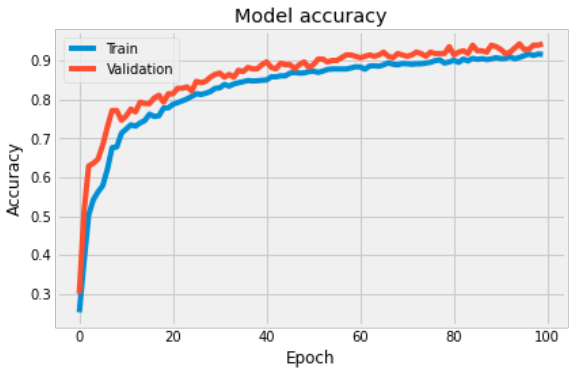
	precision	recall	f1-score	support
0	0.71	0.83	0.77	385
1	0.83	0.87	0.85	369
2	0.49	0.33	0.39	367
3	0.60	0.70	0.65	375
4	0.71	0.66	0.68	369
accuracy			0.68	1865
macro avg	0.67	0.68	0.67	1865
weighted avg	0.67	0.68	0.67	1865

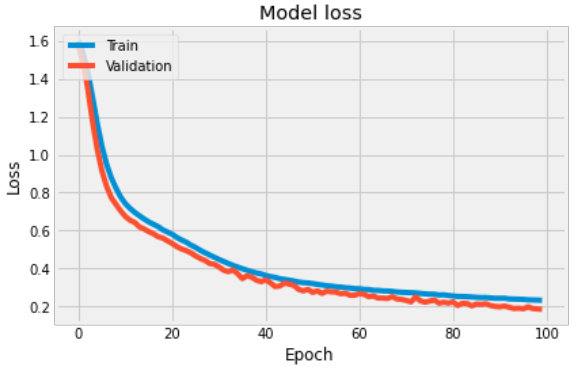


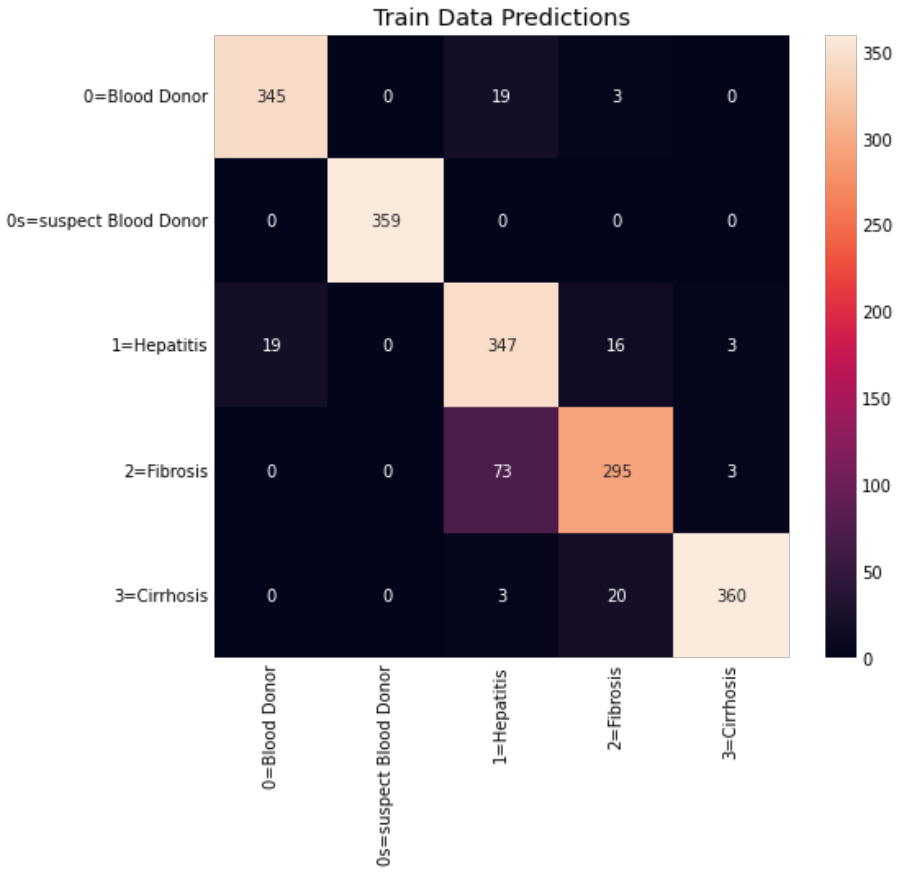

```
Classification Report
              precision    recall  f1-score   support

           0       0.72      0.83      0.77       148
           1       0.86      0.91      0.88       164
           2       0.50      0.36      0.41       166
           3       0.56      0.67      0.61       158
           4       0.70      0.63      0.66       164

    accuracy                           0.68       800
   macro avg       0.67      0.68      0.67       800
weighted avg       0.67      0.68      0.67       800
```

For the simplest model, which only includes three of the original twelve variables (`AST, GGT` and `BIL`), we can report an F1 accuracy of **68%**.

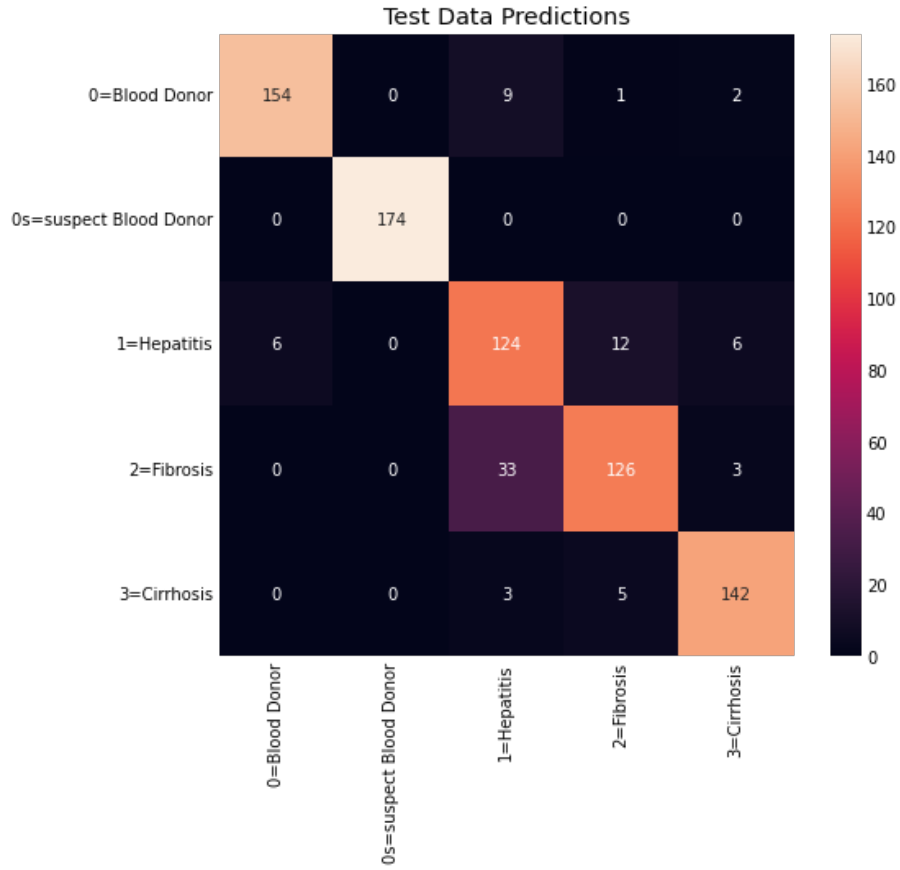
d. Model 4 (All variables <u>included</u> except "AGE")







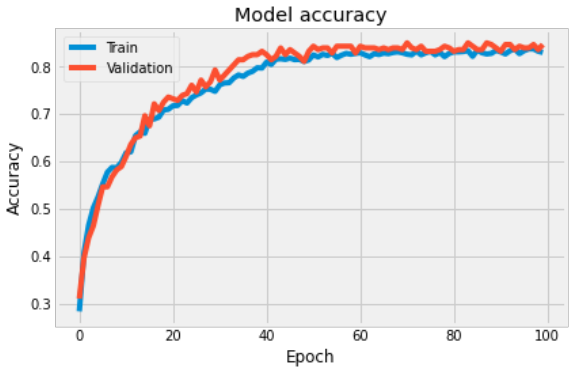

```
Classification Report
              precision    recall  f1-score   support

           0       0.95      0.94      0.94       367
           1       1.00      1.00      1.00       359
           2       0.79      0.90      0.84       385
           3       0.88      0.80      0.84       371
           4       0.98      0.94      0.96       383

    accuracy                           0.91      1865
   macro avg       0.92      0.92      0.92      1865
weighted avg       0.92      0.91      0.92      1865
```

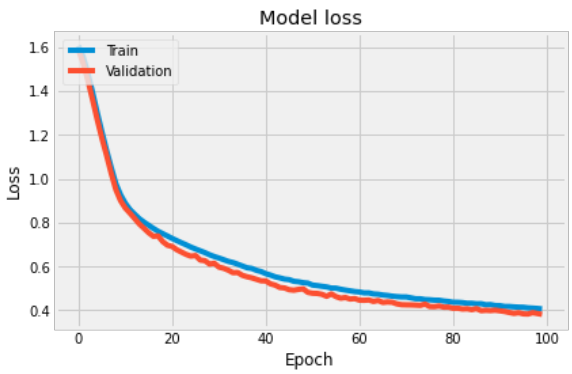


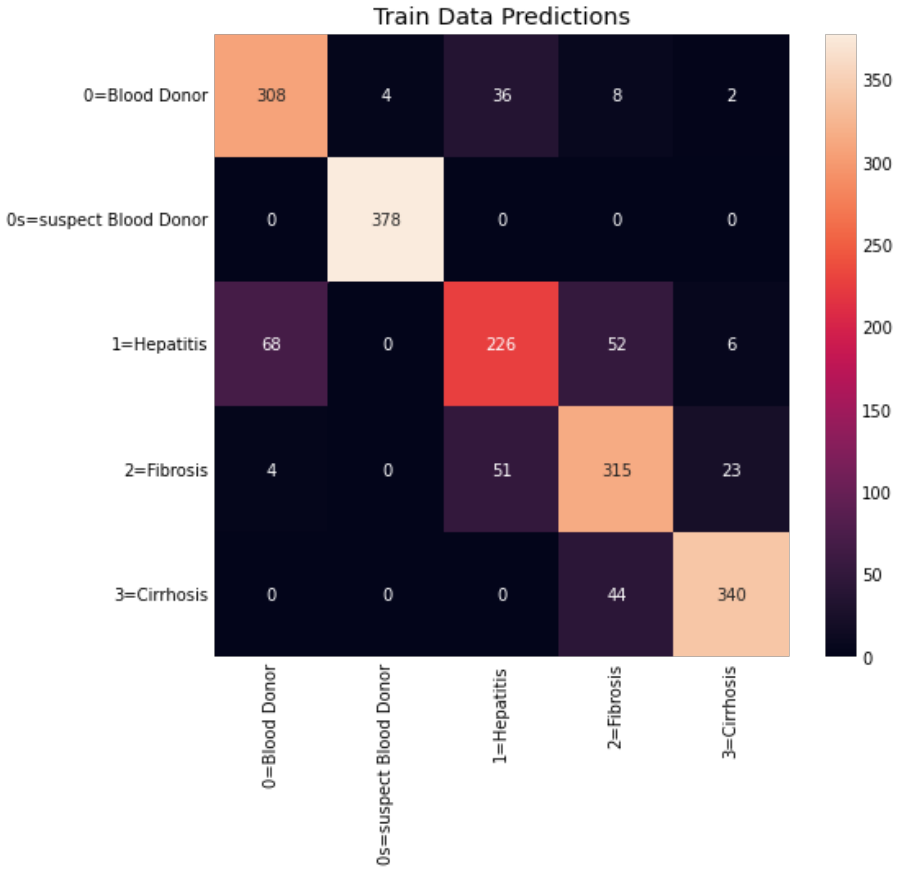

```
Classification Report
              precision    recall  f1-score   support

           0       0.96      0.93      0.94       166
           1       1.00      1.00      1.00       174
           2       0.73      0.84      0.78       148
           3       0.88      0.78      0.82       162
           4       0.93      0.95      0.94       150

    accuracy                           0.90       800
   macro avg       0.90      0.90      0.90       800
weighted avg       0.90      0.90      0.90       800
```

For the model that includes all variables except `AGE`, we can report an F1 accuracy score of **90%**.

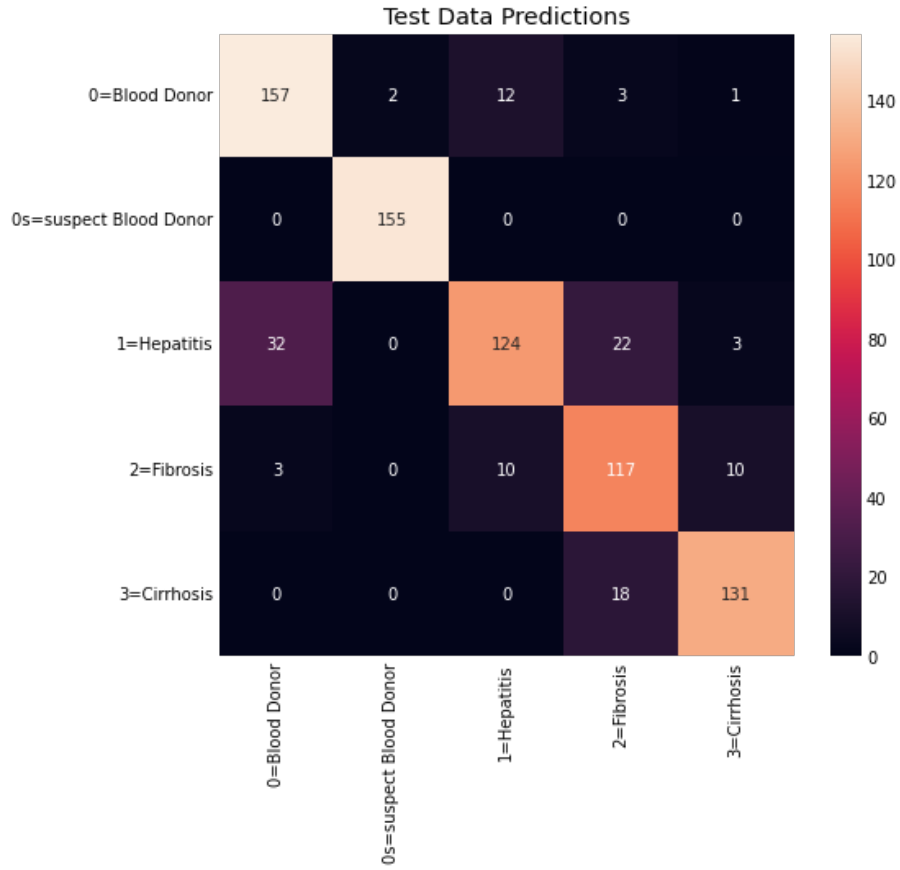
e. Model 5 (“ALP,” “ALT,” “GGT” and “CHOL” dropped from input variables)








```
Classification Report
              precision    recall  f1-score   support

           0       0.81      0.86      0.83       358
           1       0.99      1.00      0.99       378
           2       0.72      0.64      0.68       352
           3       0.75      0.80      0.78       393
           4       0.92      0.89      0.90       384

    accuracy                           0.84      1865
   macro avg       0.84      0.84      0.84      1865
weighted avg       0.84      0.84      0.84      1865
```

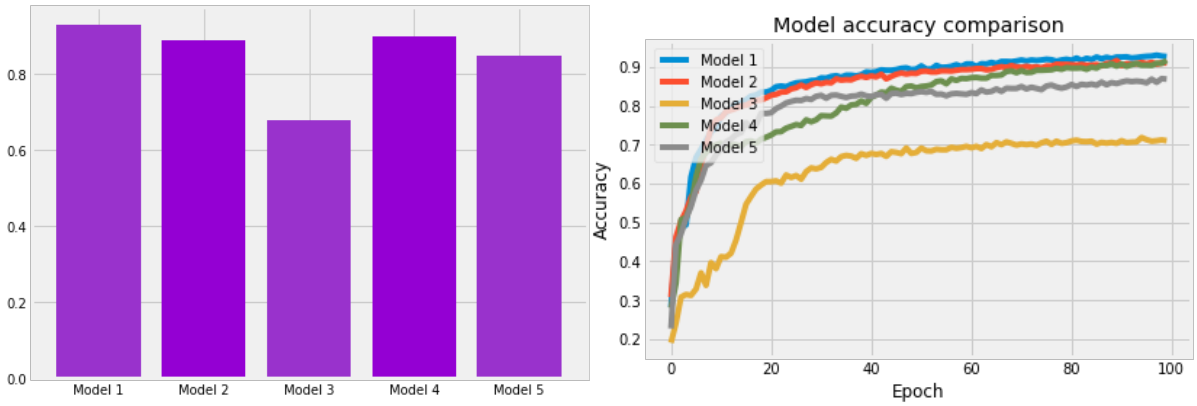



```
Classification Report
              precision    recall  f1-score   support

           0       0.82      0.90      0.86       175
           1       0.99      1.00      0.99       155
           2       0.85      0.69      0.76       181
           3       0.73      0.84      0.78       140
           4       0.90      0.88      0.89       149

    accuracy                           0.85       800
   macro avg       0.86      0.86      0.86       800
weighted avg       0.86      0.85      0.85       800
```

In the final model, which contains all variables except `ALP, ALT, GGT` and `CHOL`, we report an F1 Accuracy of **85%**.

f. Model Comparisons

Our final comparison metric is going to be based solely on the F1 score, which gives us the best overview of how accurate the model is. We can see that for the most part, the models are almost entirely identical (except the one outlier, model 3, which only gives us an F1 score of 68%). The remainder of the models are very similar to each other in terms of accuracy, all boasting F1 scores of 90% and above. Although the actual differences in the model F1 scores tend to be marginal, it is clear that the model with all of the variables included is the best model, which has an F1 score of 93%. Below is a bar chart comparing the different models with each other, along with the accuracy vs. epoch curve for each of the models plotted on the same graph to give a better visual understanding of the performance of the models.



8. Main Findings

- The full model which includes all 12 predictors provides the best accuracy and F1 score in predicting hepatitis C, fibrosis and cirrhosis in patients (F1 score accuracy: 93%). This means that the best prediction model is the full model – the more predictors we have in our training data, the better the accuracy will be when we test the model against unseen data.
- Removal of the age variable provides little to no difference in the model accuracy – only dropping the F1 accuracy from 93% to 90%. The 3% drop is relatively insignificant by our standards.
- Using only 3 out of the 12 variables (**"AST," "GGT"** and **"BIL"**) based off a correlation matrix variable selection method will still result in 68% F1-score accuracy(!). This gives us enough evidence to suggest that selecting variables in order of correlation with **"Category"** can provide high levels of accuracy in neural network models.
- Combining variable importance values from pre-existing machine learning techniques such as decision trees, random forests and K-nearest neighbors can accurately aid in the variable selection process for neural network models (89% F1 score and 85% F1 score using average of 3 ML technique values)
- Artificial Neural Networks of only 2 hidden layers are enough to obtain accuracies of over 90% for problems of this nature (prediction of hepatitis C disease in patients). The addition of more layers beyond that do not provide increases in accuracy as effectively.

9. Conclusion

Based on the research conducted for this project, along with the experimental outputs that we got throughout the process, it is evident that using a combination of neural networks with pre-existing ML techniques is the best way in both optimizing a neural network and reducing the number of variables required in prediction. Although the full neural network model with all 12 variables still proved to be the best model with an F1 accuracy of 93%, the models with the reduced variables (through the hybrid

variable selection method) were only marginally worse in terms of the accuracy metric (89%, 85%). This could easily be attributed to other issues within the neural network model, such as not having enough layers within the model itself, not training the model for enough epochs, and other errors that were left unaccounted. To the best of our knowledge, a hybrid system used in the variable selection process for the neural networks is completely novel – although the literature indicates the use of hybrid neural networks such as CSFNNs, these methods do not consider the selection of variables as much as they do the actual architecture of the neural networks themselves. Our approach is more or less taking a "step back" into pre-processing before we conduct the actual predictions.

This does not, however, mean that our models were perfect, regardless of their novelty. Firstly, when it comes to the metrics used for assessing the "goodness" of the models, we could have considered other appropriate methods, such as the KNN approach for the training / testing data splits, ROC curves, and beyond. This would give us more information to consider rather than "just" the F1-score accuracy. Additionally, it is important to highlight that our hybrid variable selection technique was simply an average. The actual variable importance score was simply the sum of the three ML techniques' importance metrics divided by 3. We could improve this significantly through scaling, associating weights, and other methods. One leap that could be taken is to consider the accuracy of each ML technique, and based on that accuracy, associate a weight to each of the variable importance measures. For example, if using the Random Forest method results in an accuracy of 100%, then the variable importance metrics will all be weighted 1, and so on.

In conclusion, this paper presents a novel approach in pre-processing data for neural network model building – specifically by considering a variable selection method based on widely used ML techniques. This novel approach was then applied to the problem of predicting Hepatitis C in patients, which boasted accuracies of up to 90%. We conclude that despite the improvements that could be made to this method, it is potentially very useful for neural network modeling.

10. Sources

[1] “Hepatitis,” *World Health Organization*. [Online]. Available: https://www.who.int/health-topics/hepatitis#tab=tab_1.

[2] “What is viral hepatitis?,” *Centers for Disease Control and Prevention*, 28-Jul-2020. [Online]. Available: https://www.cdc.gov/hepatitis/abc/index.htm.

[3] “Global progress report on HIV, viral hepatitis and sexually transmitted infections, 2021,” *World Health Organization*, 15-Jul-2021. [Online]. Available: https://www.who.int/publications/i/item/9789240027077.

[4] “Who urges countries to invest in eliminating hepatitis,” *World Health Organization*, 26-Jul-2019. [Online]. Available: https://www.who.int/news/item/26-07-2019-who-urges-countries-to-invest-in-eliminating-hepatitis.

[5] K. M. Robinson, “How to get help with hepatitis C treatment costs,” *WebMD*, 26-Jun-2020. [Online]. Available: https://www.webmd.com/hepatitis/help-hepatitis-c-treatment-costs.

[6] M. Medicine, “Early detection, early treatment for hepatitis C,” *University of Michigan*, 14-Mar-2016. [Online]. Available: https://labblog.uofmhealth.org/rounds/early-detection-early-treatment-for-hepatitis-c.

[7] N. Shahid, T. Rappon, and W. Berta, “Applications of artificial neural networks in health care organizational decision-making: A scoping review,” *PLOS ONE*, vol. 14, no. 2, 2019.

[8] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, “Using machine learning techniques to generate laboratory diagnostic pathways—A case study,” *Journal of Laboratory and Precision Medicine*, vol. 3, pp. 58–58, Jun. 2018.

[9] S. Ansari, I. Shafi, A. Ansari, J. Ahmad, and S. I. Shah, “Diagnosis of liver disease induced by hepatitis virus using Artificial Neural Networks,” *2011 IEEE 14th International Multitopic Conference*, 2011.

[10] D. Mital, S. Haque, and S. Srinivasan, “Prediction of hepatitis C using Artificial Neural Network,” *7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002.*, Oct. 2003.

[11] M. Rouhani and M. M. Haghighi, “The diagnosis of hepatitis diseases by support vector machines and Artificial Neural Networks,” *2009 International Association of Computer Science and Information Technology - Spring Conference*, Jul. 2009.

[12] L. Ozyilmaz and T. Yildirim, “Artificial neural networks for diagnosis of hepatitis disease,” *Proceedings of the International Joint Conference on Neural Networks, 2003.*, Jul. 2003.