Final Project Report for Methods for Data-Driven Computational Engineering Research

Smrithi AJIT[1]

[1] Department of Civil and Environmental Engineering, Iowa State University; email: sajit@iastate.edu

ABSTRACT

This project aims develop and compare GAM modelling technique with Deep Learning modelling technique to help predict if a patient is a liver patient or not based on parameters available from liver function test results. Their performance was evaluated and compared based on a combination of rmse, AUC and GAM performance plots. Additionally, the author has attempted to understand how data curing techniques can help to improve model and has also made an attempt to develop a representative or index dataset that can be used to represent the complete dataset using GA-NMM method.

INTODUCTION

Liver disease is one of the most life-threatening diseases in the world and affects a large number of people in developing countries. Liver is one the largest organ in the human body and carries out vital function like bile and protein production, storing and releasing glucose, processing hemoglobin, blood cleaning, immune factor production, clearing bilirubin. Liver disease can be caused by numerous factors like viral infections, immune system issues, drugs, poison and alcohol consumption, inherited diseases or even cancer and tumor. The liver failure can be averted if the liver disease is detected at an early stage. Liver Function Test is a test that helps to determine the amount of SGPT, SGOT, albumen and several other components of blood that help determine if a patient is a liver patient or not. Machine learning models trained on real world data can be very useful in helping doctors diagnose and categorize such patients.

SECTION 1. BASIC STATISTICS OF ORIGINAL DATA

The dataset contains 583 records for 11 attributes in which class 1 includes 416 liver patients, and class 2 includes 167 non-liver patients. The dataset used in this study was obtained from the University of California at Irvine (UCI) Machine Learning

Repository (*1*) and has been found used by other studies (*2–7*). In this problem, the variable 'Liverpatient' is the dependent variable and is of factor type. This variable indicates whether a patient is a liver patient or not. It takes on the value 1 if it is a liver patient and 2 otherwise. The other predictor variables include Age, gender, total bilirubin(TB), Alkphos(AAP), SGPT, SGOT, Total proteins(TP), Albumin(ALB) and A/G ratio. Among the remaining variables that act as independent variables, in terms of the datatype of the variables, the variable gender is factor type, age is integer and all other variables are of type float. The mean, median, min, max, first and third quartile values corresponding to the dataset have been summarized in the Fig 1. For the categorical variables gender and Liverpatient the number of instances in each category is provided. The complete range of all the variables, the standard deviation, skew and kurtosis are additional details about each variable is provided in Fig 2 and is obtained using the describe function in R. The independent variables SGPT and SGOT have the maximum skew followed by AAP, DB and TP all having positive skew.

Age	Gender	TB	DB	AAP	SGPT	SGOT
Min. : 4.00	1:441	Min. : 0.400	Min. : 0.100	Min. : 63.0	Min. : 10.00	Min. : 10.0
1st Qu.:33.00	2:142	1st Qu.: 0.800	1st Qu.: 0.200	1st Qu.: 175.5	1st Qu.: 23.00	1st Qu.: 25.0
Median :45.00		Median : 1.000	Median : 0.300	Median : 208.0	Median : 35.00	Median : 42.0
Mean :44.75		Mean : 3.299	Mean : 1.486	Mean : 290.6	Mean : 80.71	Mean : 109.9
3rd Qu.:58.00		3rd Qu.: 2.600	3rd Qu.: 1.300	3rd Qu.: 298.0	3rd Qu.: 60.50	3rd Qu.: 87.0
Max. :90.00		Max. :75.000	Max. :19.700	Max. :2110.0	Max. :2000.00	Max. :4929.0

TP	ALB	AG_ratio	Liverpatient
Min. :2.700	Min. :0.900	Min. :0.3000	1:416
1st Qu.:5.800	1st Qu.:2.600	1st Qu.:0.7000	2:167
Median :6.600	Median :3.100	Median :0.9300	
Mean :6.483	Mean :3.142	Mean :0.9471	
3rd Qu.:7.200	3rd Qu.:3.800	3rd Qu.:1.1000	
Max. :9.600	Max. :5.500	Max. :2.8000	
		NA's :4	

Fig 1: Basic summary of the dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Age	1	583	44.75	16.19	45.00	44.84	17.79	4.0	90.0	86.0	-0.03	-0.57	0.67
Gender*	2	583	1.24	0.43	1.00	1.18	0.00	1.0	2.0	1.0	1.19	-0.58	0.02
TB	3	583	3.30	6.21	1.00	1.74	0.44	0.4	75.0	74.6	4.88	36.70	0.26
DB	4	583	1.49	2.81	0.30	0.74	0.30	0.1	19.7	19.6	3.20	11.20	0.12
AAP	5	583	290.58	242.94	208.00	238.40	74.13	63.0	2110.0	2047.0	3.75	17.52	10.06
SGPT	6	583	80.71	182.62	35.00	43.91	22.24	10.0	2000.0	1990.0	6.52	49.95	7.56
SGOT	7	583	109.91	288.92	42.00	56.79	31.13	10.0	4929.0	4919.0	10.49	149.10	11.97
TP	8	583	6.48	1.09	6.60	6.51	1.04	2.7	9.6	6.9	-0.28	0.21	0.04
ALB	9	583	3.14	0.80	3.10	3.15	0.89	0.9	5.5	4.6	-0.04	-0.40	0.03
AG_ratio	10	579	0.95	0.32	0.93	0.93	0.25	0.3	2.8	2.5	0.99	3.22	0.01
Liverpatient*	11	583	1.29	0.45	1.00	1.23	0.00	1.0	2.0	1.0	0.94	-1.11	0.02

Fig 2: Basic statistics of the dataset

In terms of the correlation between variables, from the correlation plot in Fig 3 it is clear that the correlation is high between DB and TB, SGPT and SGOT, ALB and TP is high, AG_ratio and ALB. Any correlation value above 0.4 has been taken as a high level of correlation between variables. Therefore, we drop TB, SGOT, TP and ALB

from subsequent analysis since they have a very strong correlation between them. From Fig 4, it is observable from the graph that TB and DB have a more or less linear relationship. Age, Gender, DB and TB are negatively correlated with the outcome variable Liverpatient. While TP is positively correlated with the outcome variable, SGPT,AAP and SGOT are all negatively correlated to the outcome as shown in Fig 5. The outcome variable is also heavily unbalanced with a large proportion of the datapoints in 1 compared to 2.

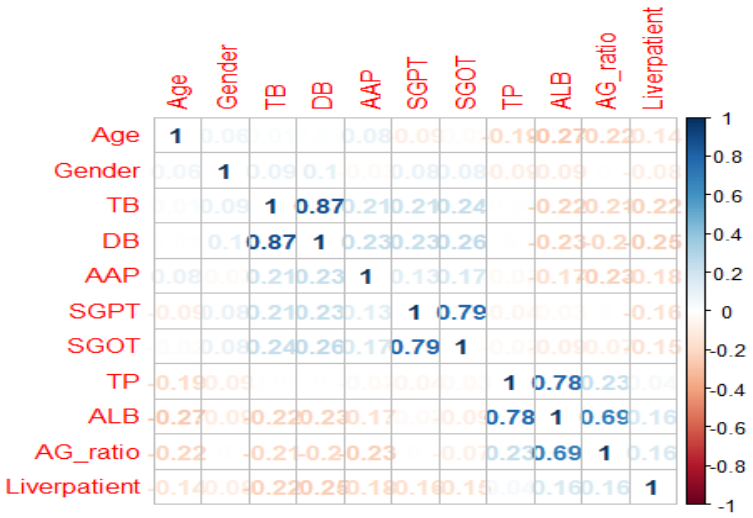


Fig 3: Correlation plot between variables of the dataset

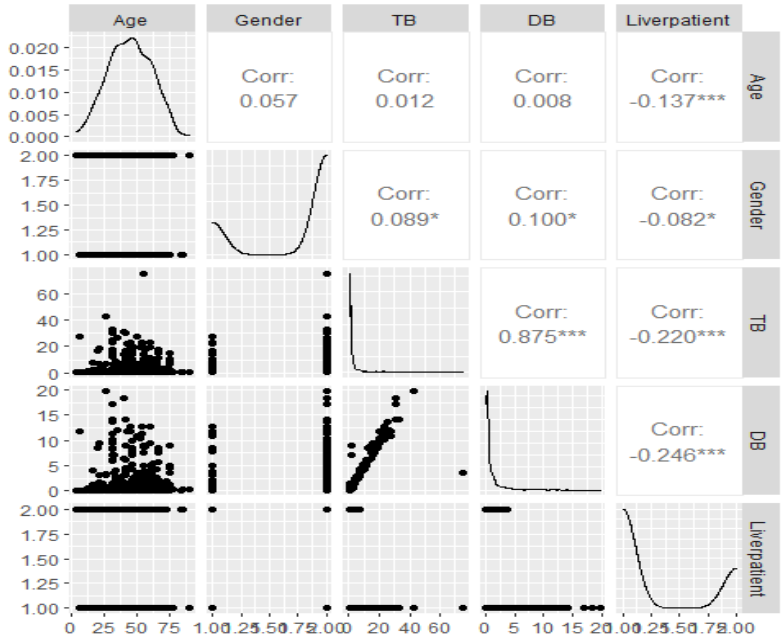


Fig 4: Correlation between variables Age, gender,TB,DB and liverpatient

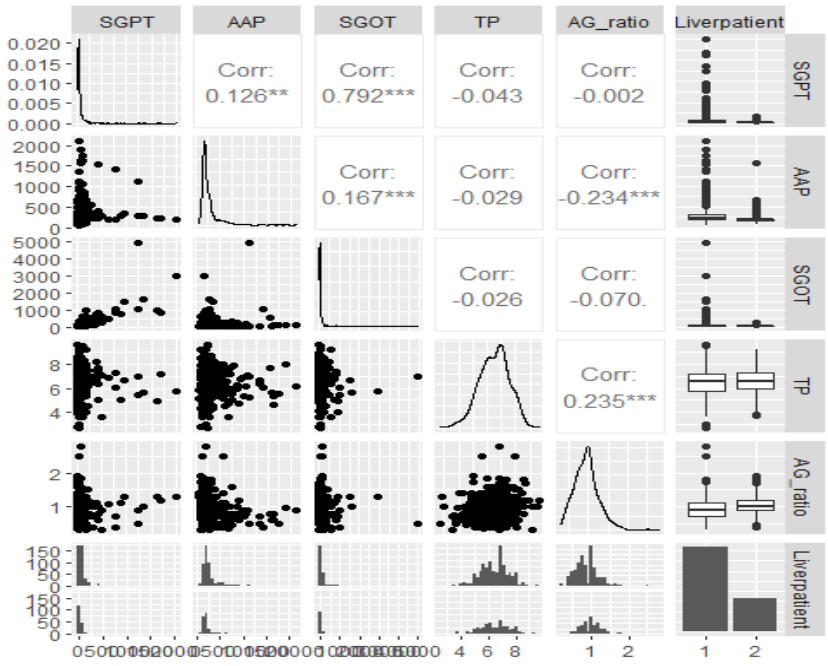


Fig 5: Correlation between variables SGPT,AAP,SGOT,TP,AG_ratio and liverpatient

SECTION 2. DATA SQUASHING OF ORIGINAL DATA

In this section, a moment matching technique called as Genetic Algorithm Moment Matching Technique has been used to squash the original dataset into a smaller representative dataset. The methodology was adopted from (*8*).

As equation defining the relationship between the dependent and independent variables is found by formulating a linear model on the dataset. The linear model function in R package was used to develop the linear model. The results of the linear model have been summarized in Fig 7. The equation formulated from the linear model results was used to determine the outcome of whether a patient is a liver patient or not. An adhoc rule that any value of greater 1.5 is considered as a negative and closer to 1 is considered a positive case since here, 1 stands for liver patient and 2 stands for not liver patient.

$$\mathrm{Y} = 1.368 - 0.003155 * \mathrm{Age} + 0.07281 * \mathrm{Gender} - 0.000244 * \mathrm{AAP} - 0.0003751 * \mathrm{SGPT} - 0.0129 * \mathrm{TP} + 0.1619 * \mathrm{AG\_ratio} \quad \ldots\ldots(1)$$

Equation 1 was used to find the response variable 'Liverpatient' for all the datapoints of the original dataset and then find the response of the index dataset shown in Fig 4. The mean and standard deviation of the responses for the two datasets are shown in Table 1. It can be seen that the index dataset is quite representative of the original dataset. The original dataset has a mean of 1.28 which can be interpreted as a discrete output of 1w hile the index datapoints have a mean of 1.3278. However the variance has increased by approximately 1.2 percent. This is expected since we are reducing a dtaaste from 583 to just 12 representative points. Though the dataset is unbalanced with 416 instances of 1 and 167 cases of 2, this method is doing a good job of generating a representative dataset. The expected value of X, X^2,X^3,X^4,X^5 have been summarized in the Table 2 along with the range of the variables that have been provided to the Genetic Algorithm software.table 3 provides the range of variable and ther moments or expectation that were calculated and fed to the Genetic Algorithm software.

Table 1: Summary of mean and variance for original dataset and index dataset

Dataset	Mean	Variance
Original complete dataset	1.285	0.0178
Index dataset	1.3278	0.0385

Table2: Index location and weight for each variable

AAP			TP			SGPT		
Index	loc	W	Index	loc	W	Index	loc	W

1	1.74148	0.476357	1	4.33264	0.00019083	1	593.535	0.0435
2	1.14899	0.46074	2	0.780014	0.770304	2	1727.79	0.007585
3	2.74343	0.062902	3	1.50493	0.229505	3	44.0407	0.948918
AG_ratio			Age			Gender		
Index	loc	W	Index	loc	W	Index	loc	W
1	0.72471	0.692664	1	44	0.588864	1	1	0.756
2	1.53166	0.24556	2	69	0.212933	2	2	0.244
3	0.790347	0.0733591	3	19	0.198204			

Table 3: Range of variables and expectation provided to the GA-NMM

	AAP	AG_ratio	SGPT	TP	Age
E(x)	290.57633	0.9470639	80.7135	6.48319	44.746141
$E(x^2)$	143352.2367	0.992039	39807.667	43.2079	2263.878216
$E(x^3)$	129601246.2847	1.16333	48269749.6741	295.012698	124592.4065
$E(x^4)$	170876609821.678	1.52962	73053060873.9451	2058.321158	7296761.778
$E(x^5)$	267599196686563	2.285405	120326246530528	14646.02874	448382767.542
Min	63	0.3	10	2.7	4
Max	2110	2.8	2000	9.6	90

	AAP	TP	SGPT	AG_ratio	Age	Gender	Weight	Predictedoutcome
1	1.74148	1.50493	593.535	0.790347	44	1	-3.1344	1.312382791
2	1.14899	1.50493	593.535	0.790347	44	1	0.46074	1.312527359
3	2.74343	1.50493	593.535	0.790347	44	1	0.062902	1.312138315
4	1.74148	0.780014	593.535	0.790347	44	1	0.770304	1.195018891
5	1.74148	4.33264	593.535	0.790347	44	1	0.229505	1.77018904
6	1.74148	1.50493	1727.79	0.790347	44	1	0.007585	0.886923741
7	1.74148	1.50493	44.0407	0.790347	44	1	0.948918	1.518498103
8	1.74148	1.50493	593.535	0.72471	44	1	0.770304	1.313229508
9	1.74148	1.50493	593.535	1.53166	44	1	0.229505	1.302819853
10	1.74148	1.50493	593.535	0.790347	69	1	0.212933	1.233507791
11	1.74148	1.50493	593.535	0.790347	19	1	0.198204	1.391257791
12	1.74148	1.50493	593.535	0.790347	44	2	0.2435	1.385192791

Fig 6: Multi-dimensional indexes with the outcome from linear model

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.6190 -0.3191 -0.2112  0.5615  1.1108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.368e+00  1.489e-01   9.186  < 2e-16 ***
Age         -3.155e-03  1.160e-03  -2.720 0.006726 **
Gender       7.281e-02  4.253e-02   1.712 0.087458 .
AAP         -2.439e-04  7.705e-05  -3.166 0.001627 **
SGPT        -3.751e-04  1.004e-04  -3.737 0.000205 ***
TP          -1.290e-02  1.740e-02  -0.741 0.458891
AG_ratio     1.619e-01  6.065e-02   2.669 0.007828 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4339 on 572 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.08726,   Adjusted R-squared:  0.07769
F-statistic: 9.114 on 6 and 572 DF,  p-value: 1.545e-09
```

Fig 7: Linear model results using original dataset

SECTION 3. DATA CURING OF ORIGINAL DATA

The variable mean estimation and standard error corresponding to 10%, 20% and 30% missing data has been summarized in the figures Fig 6, Fig 7, Fig 8. The variables 1 to 7 correspond to Age, gender, AAP,SGPT,TP, AG_ratio and Liverpatient. It is observed that as the missingness increases from 10 to 30 percent the standard error also increases. The variables 3 and 4 i.e. AAP and SGPT are very sensitive to the missingness and show very high standard error of 10.2 and 9.467 even with 10 % missingness. Age shows a standard error of 0.715 which goes a little above 1 when the missingness goes upto 30%. The first variable, Age can be rounded off to 45 in all three cases since this is an integer. This implies that Age is not affected by the missingness and imputation may work well this variable. However, for other sensitive quantities AAP, SGPT, TP, AG_ratio the impact is felt more. The value of k used is 3 and M=5, where k refers to the categories and M refers to the neighbors being considered.

For the variables Age, gender,TP,AG_ratio and Liverpatient the standard error is not varying much with the value of k. For 10% data missing, k=3 produces the lowest error. For 20% missing data k=3 and k=6 produce lower standard error and for 30% missing data k=4 produces lowest standard error. However, k seems to have a stronger influence over variables AAP and SGPT where, the lowest standard error can be found corresponding to k=3 for 10%,20% and 30% data missing. Table 3 summarizes the

mean values obtained for different values of k and the graphs in Fig 8 and 9 show the impact of k on the standard error.

Table 3: Mean values of the variables corresponding to different values of k

Missingness	k	Age	Gender	AAP	SGPT	TP	AG_ratio	Liverpatient
10%	3	44.83	1.246	288.261	84.442	6.468	0.9548	1.27
10%	4	44.14	1.2407	291.643	99.574	6.454	0.943	1.2899
10%	5	44.718	1.254	287.804	106.5139	6.466	0.94998	1.2777
10%	6	44.573	1.2444	291.3297	96.37148	6.45695	0.94338	1.2827
20%	3	44.717	1.24	288.228	90.48	6.436	0.9404	1.2833
20%	4	44.572	1.244	291.3297	96.3715	6.457	0.9433	1.2827
20%	5	44.1423	1.2407	291.6432	99.5742	6.454	0.9437	1.29
20%	6	44.718	1.254	287.804	106.5138	6.466	0.94998	1.2777
30%	3	44.6567	1.232	293.053	71.435	6.4698	0.9399	1.2589
30%	4	44.27	1.2422	292.13	84.3166	6.4589	0.9519	1.291
30%	5	44.3524	1.237	288.5202	86.46	6.456	0.9514	1.3
30%	6	44.33	1.2424	289.265	84.4517	6.4629	0.65522	1.302

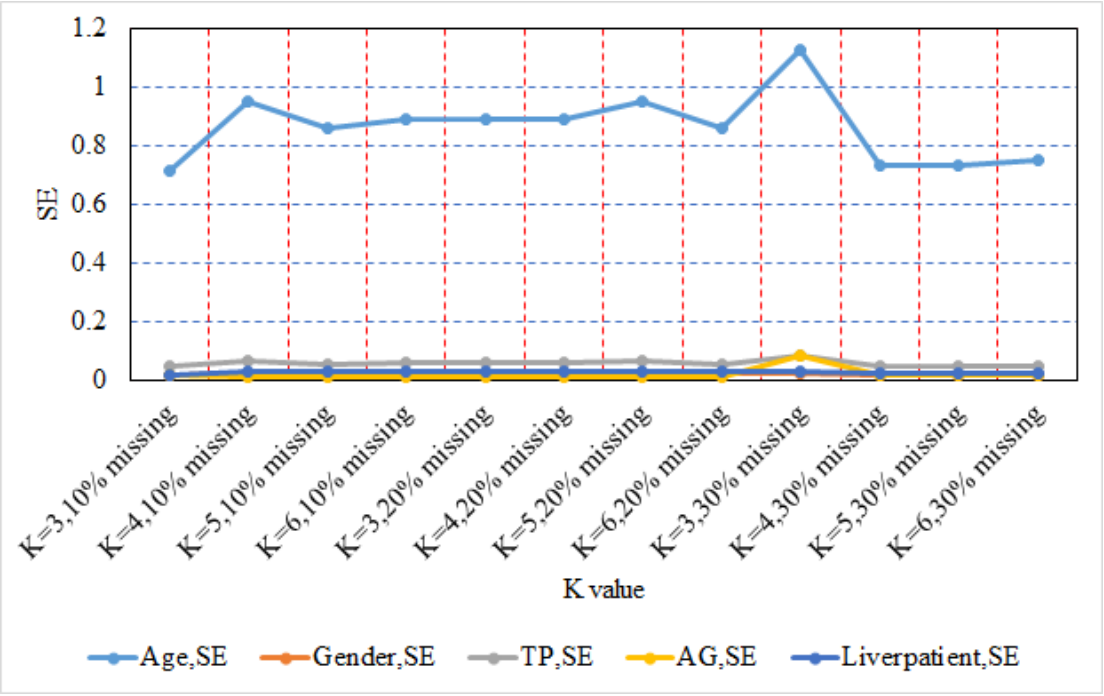


Fig 8: Variation of standard error for different values of k

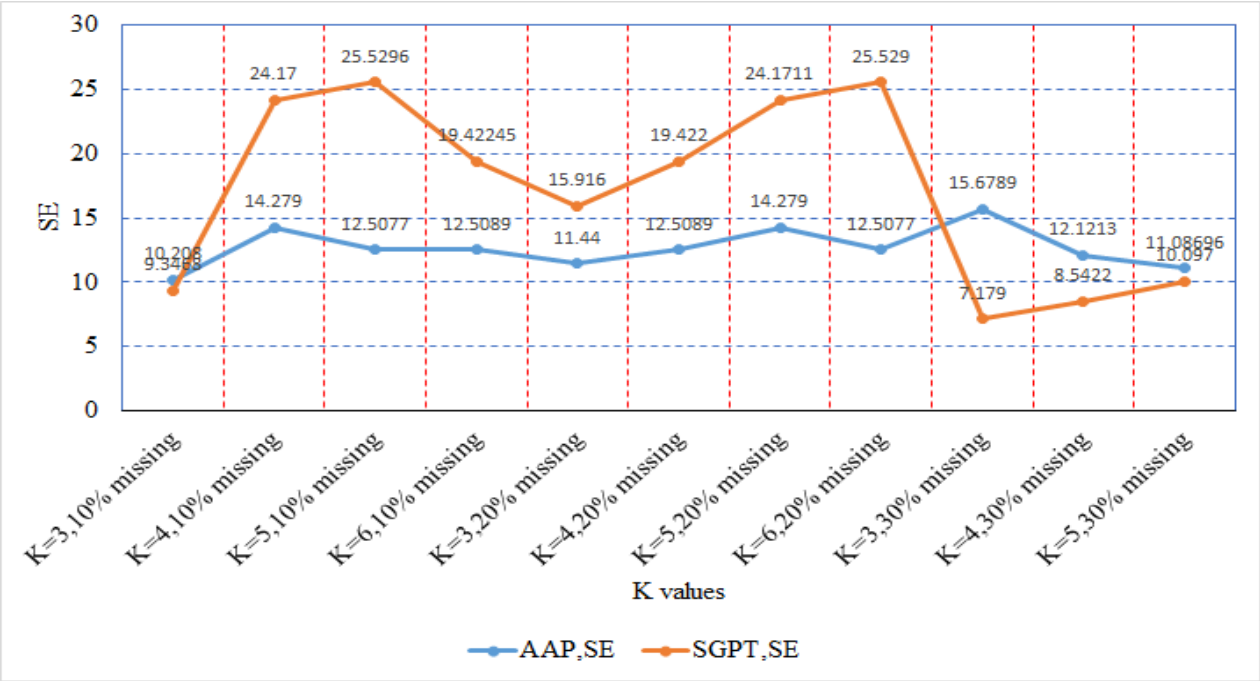


Fig 9: Variation of standard error for different values of k

In the next part, a comparison between Naïve method and FHDI methods of imputation was compared. The influence of each of these methods on the predictor variables and the model performance helped understand if the particular imputation method was helping the variable to contribute better to the model or otherwise. The methodology and inspiration to use these techniques comes from (*9*). In the Tables 4,5,6, we can interpret a reduction of standard error and p value and an increase in the t value as improving the model performance. In the case of 10% missing data condition, we see that the standard error for slope and intercept is lower for FHDI method compared to the Naïve method for all the variables.The same is the case with 20% and 30% data missing conditions. If we compare the t value and p values reported, all the conditions considered report a higher value of t and lower or equal value of p for both slope and intercept under the FHDI imputation method illustrating that it is a better imputation technique for the dataset compared to Naïve method where the data with null values is dropped. Table 7 shows the mean of the fitted values for the overall model including all variables together.

Table 4: Impact of 10% data missingness on each variable in the model

Model	Variable	Intercept	SE	t_value	p_value	Slope	SE	t_value	p_value
Naive	Age	1.4582	0.054647	26.684	<2e-16	-0.003839	0.001149	-3.342	0.00884
FHDI	Age	1.44732	0.037692	38.4	<2e-16	-0.003629	0.000796	-4.56	5.64e-6

Naive	Gender	1.1785	0.05729	20.57	<2e-16	0.08681	0.04355	1.993	0.0467
FHDI	Gender	1.22778	0.03948	31.102	<2e-16	0.04647	0.02995	1.552	0.121
Naive	AAP	1.387	0.02875	48.222	<2e-16	-3.443e-4	7.594e-5	-4.534	7.03e-6
FHDI	AAP	1.386	0.00198	70.046	<2e-16	-3.505e-4	5.95e-5	-6.619	5.39e-11
Naive	SGPT	1.319	0.02	65.194	<2e-16	-0.000405	0.0001	-3.993	7.37e-5
FHDI	SGPT	1.317	0.01388	94.868	<2e-16	-3.654e-4	6.451e-5	-5.664	1.84e-4
Naive	TP	1.19183	0.11361	10.49	<2e-16	0.01459	0.01728	0.844	0.399
FHDI	TP	1.1938	0.07783	15.339	<2e-16	0.01418	0.01186	1.196	0.232
Naive	AG_ratio	1.06657	0.05803	18.379	<2e-16	0.23061	0.05806	3.972	8.04e-5
FHDI	AG_ratio	1.09362	0.03983	27.45	<2e-16	0.20146	0.03958	5.09	4.14e-7

Table 5: Impact of 20% data missingness on each variable in the model

Model	Variable	Intercept	SE	t_value	p_value	Slope	SE	t_value	p_value
Naive	Age	1.4582	0.05465	26.684	<2e-16	-0.00384	0.001149	-3.342	0.000884
FHDI	Age	1.4633	0.0362	40.383	<2e-16	-0.00403	0.0007626	-5.281	1.5e-7
Naive	Gender	1.17850	0.05729	20.57	<2e-16	0.08681	0.04355	1.993	0.0467
FHDI	Gender	1.201	0.0379	31.657	<2e-16	0.06632	0.02891	2.293	0.022
Naive	AAP	1.387	0.02875	48.22	<2e-16	-3.443e-4	7.594e-5	-4.534	7.53e-6
FHDI	AAP	1.407	0.0188	74.836	<2e-16	-0.00043	0.00005	-8.606	<2e-16
Naive	SGPT	1.319	0.02	65.194	<2e-16	-0.0004	0.0001	-3.993	7.37e-5
FHDI	SGPT	1.314	0.0134	99.582	<2e-16	-3.364e-4	5.513e-5	-6.102	1.37e-9
Naive	TP	1.19183	0.11336	10.49	<2e-16	0.0146	0.01728	0.844	<2e-16
FHDI	TP	1.306	0.074	17.577	<2e-16	-0.00349	0.0114	-0.307	0.759
Naive	AG_ratio	1.06657	0.05803	18.379	<2e-16	0.23061	0.05806	3.972	8.04e-5
FHDI	AG_ratio	1.03229	0.04062	25.413	<2e-16	0.26688	0.0412	6.477	1.31e-10

Table 6: Impact of 30% data missingness on each variable in the model

Model	Variable	Intercept	SE	t_value	p_value	Slope	SE	t_value	p_value
Naive	Age	1.45822	0.0546	26.684	<2e-16	-0.000389	0.001149	-3.342	0.000884
FHDI	Age	1.487	0.0357	41.606	<2e-16	-0.0051	0.00075	-6.773	1.91e-11
Naive	Gender	1.1785	0.05729	20.57	<2e-16	0.08681	0.4355	1.993	0.0467
FHDI	Gender	1.36356	0.03732	36.54	<2e-16	-0.0849	0.02864	-2.964	0.00309
Naive	AAP	1.387	0.02875	48.22	<2e-16	-3.443e-4	7.59e-5	-4.534	7.03e-6
FHDI	AAP	1.362	0.01757	77.542	<2e-16	-3.52e-4	4.423e-5	-7.958	3.77e-15
Naive	SGPT	1.319	0.02	6.194	<2e-16	-0.0004	0.0001	-3.993	7.37e-5
FHDI	SGPT	1.284	0.01307	98.24	<2e-16	-3.547e-4	1.307e-2	98.24	<2e-16
Naive	TP	1.19183	0.11361	10.49	<2e-16	0.01459	0.01728	0.844	0.399
FHDI	TP	1.36565	0.07357	18.562	<2e-16	-0.0165	0.01122	-1.471	0.142
Naive	AG_ratio	1.06657	0.05803	18.379	<2e-16	0.23061	0.05806	3.972	8.04e-5
FHDI	AG_ratio	0.992	0.03969	24.99	<2e-16	0.28399	0.04028	7.05	2.89e-12

Table 7: Mean of the fitted values for the overall model including all variables

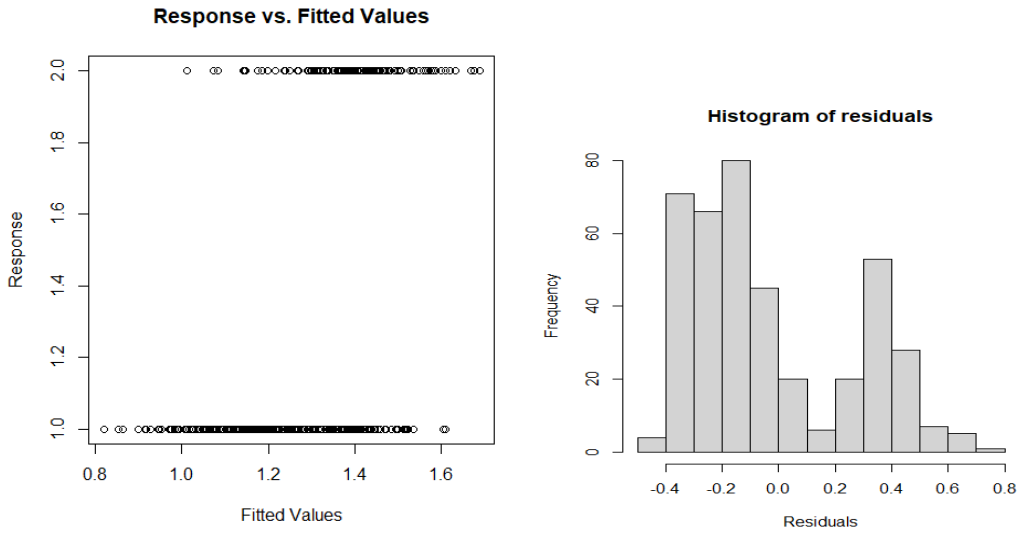
Model	Data missingness	Mean of the fitted values

Naïve	10%	1.285
FHDI	10%	1.282
Naive	20%	1.285
FHDI	20%	1.272
Naive	30%	1.285
FHDI	30%	1.2823

SECTION 4. DATA PREDICTION

(1) Statistical Prediction

In this section GAM model was used to predict the outcome of whether a patient is a liver patient or not. Different combination of variables was used to construct 5 different models. Each one was evaluated using gam.check function which generates a combination of plots that have been summarized in this section along with their interpretation in this section. The default smoothing is set. Here, Gamma(link=log) has been used as the family. Additionally, the family was set to binomial and the model was evaluated. In the cases where Gamma(link=log) is used the outcome is continuous and as per the adhoc rule set in Section 2, any continuous outcome above 1.5 is assumed to predict category 2 and any value below 1.5 is assumed to predict category 1. When all the variables are considered together as shown in Fig 10, the model does a relatively good job of predicting the datapoints at 1 better than at 2. Th histogram looks bimodal which may be expected in this case since it is a binomial outcome problem and if we were to color it based on the outcome we would see the histograms corresponding to the two outcomes. The histogram of residuals also shows almost normally distributed curve for datapoints of category 1 and 2 illustrating good performance.



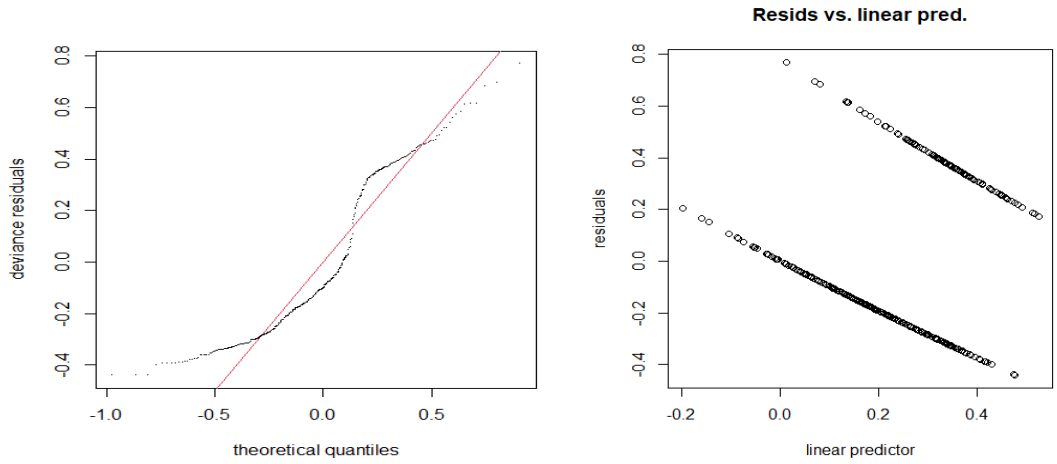
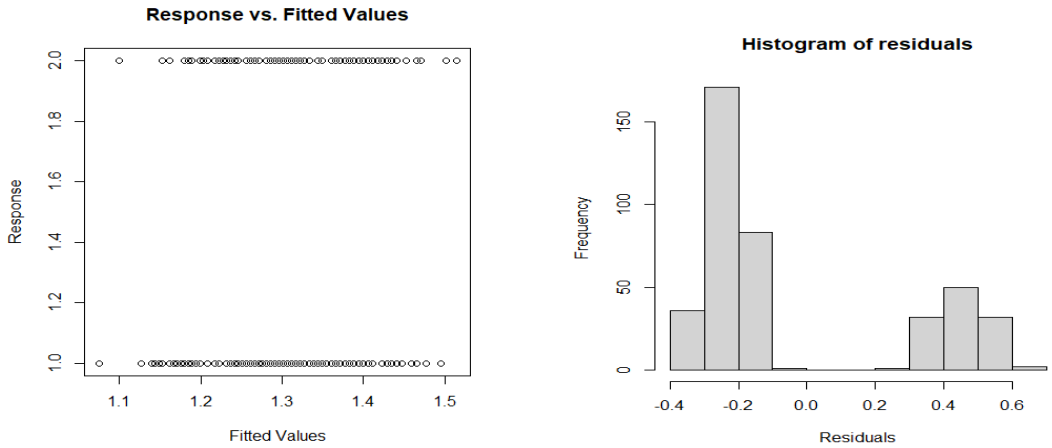


Fig 10: GAM check plot when variables Age, AAP,SGPT,TP, AG_ratio are considered



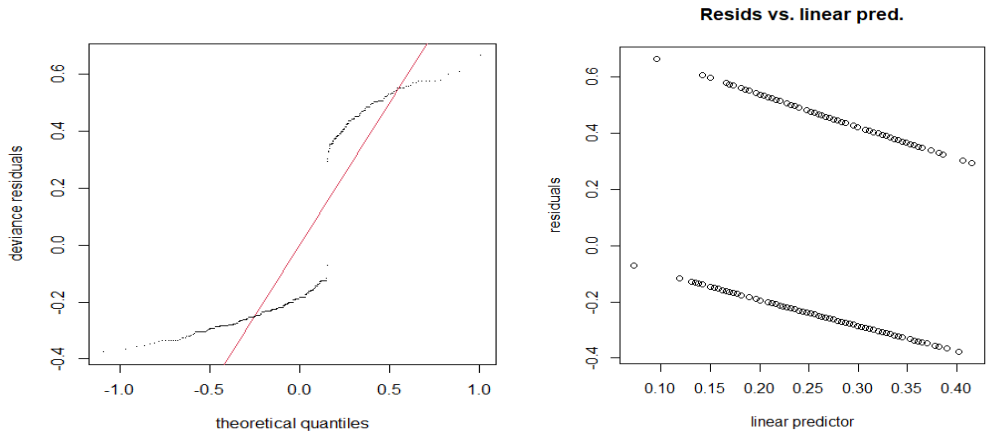
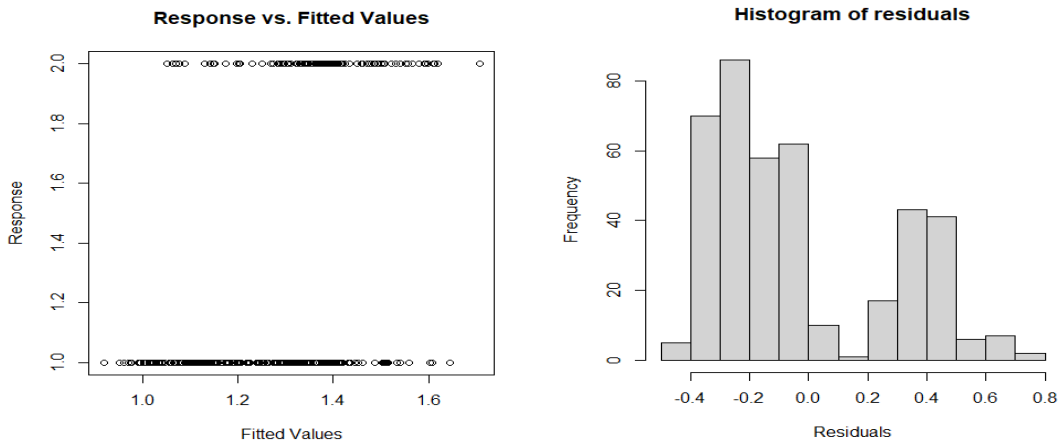


Fig 11: GAM check plot when variables Age considered

When Age alone is considered, the model is performing badly with prediction made randomly as shown in Fig 11. The histogram of residuals also not normally distributed. The deviance of the residuals is not proportional to the theoretical quantiles. The residual versus linear prediction plot should look random. Here, there seems to be a pattern indicating heteroskedasticity.



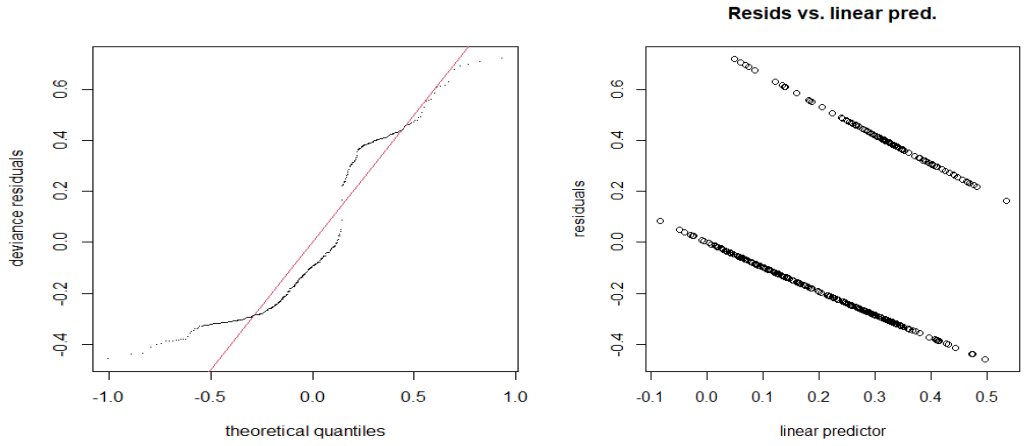
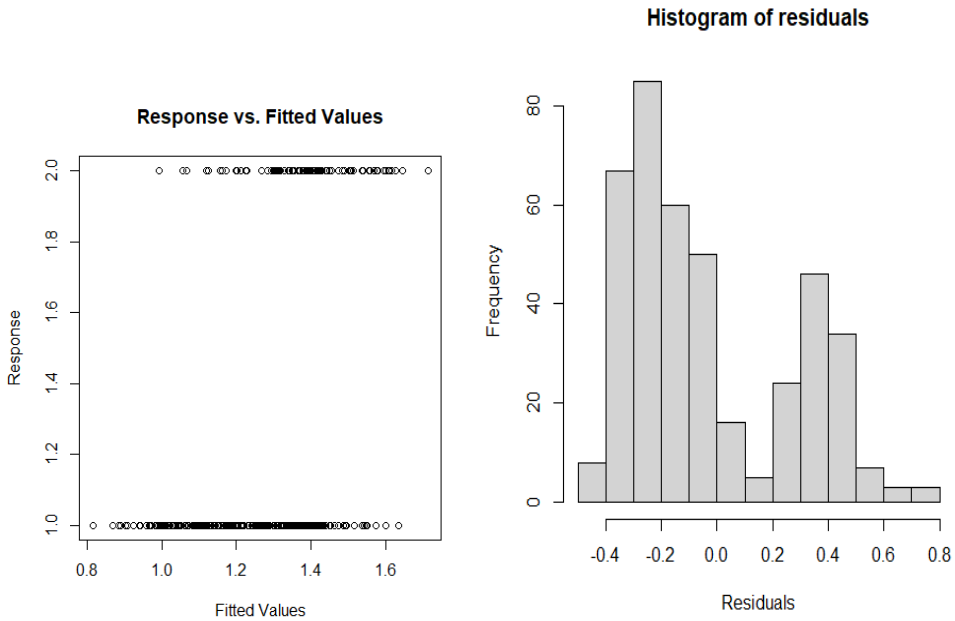


Fig 12: GAM check plot when variables Age, AAP considered

When AAP is considered along with Age, the values corresponding to the outcome 1 seem to be better predicted with more predictions scattered within the range of 1.4 which implies it would be predicted as 1 as shown in Fig 12. The histogram shows two clusters corresponding to the two outcomes of the dependent variable but is more or less normal within each of them. The residual versus linear prediction plot however, is still not very random. In fact it is exhibiting some sort of inverse linearity.



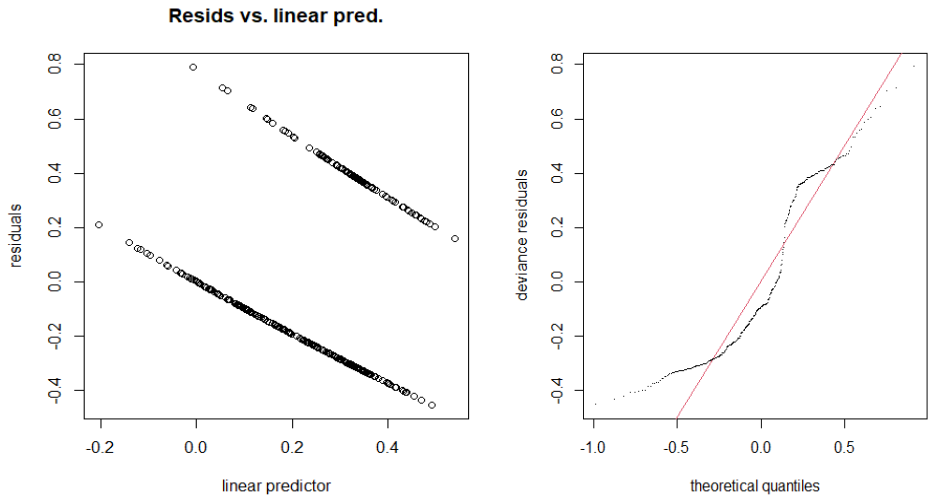
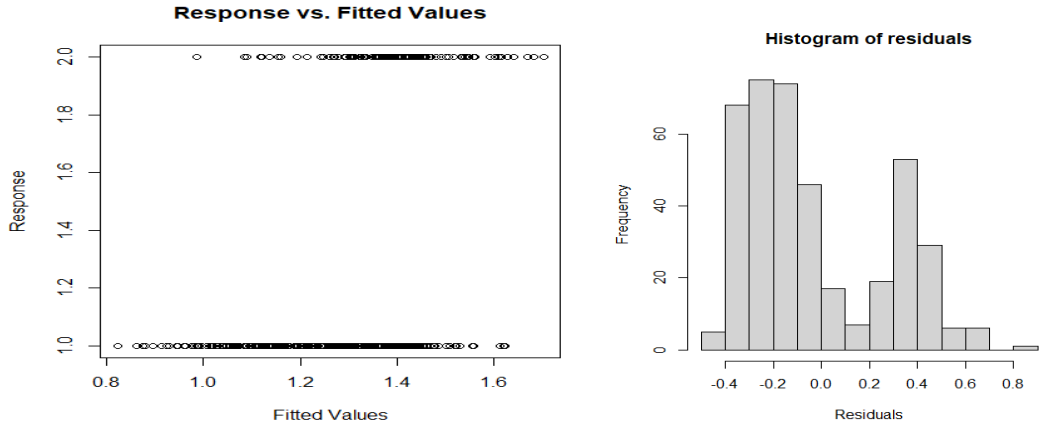


Fig 13: GAM check plot when variables Age, AAP,SGPT considered

From Fig 13 it is evident the model is performing better than the earlier case. However, her the histogram of residuals for category 2 look more normal than category 1.



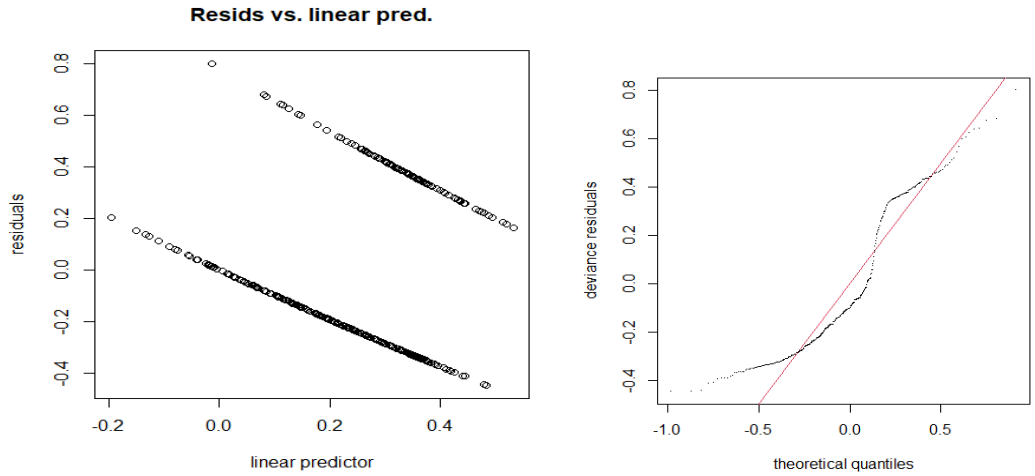


Fig 14: GAM check plot when variables Age, AAP,SGPT, AG_ratio considered

As the number of variables is increased to include Age, AAP, SGPT and AG_ratio the model seems to be performing better with the responses corresponding to outcome'2' or 'not a liver patient' aggregated more towards values beyond 1.5 and the responses corresponding to '1' or 'is liver patient clustered more around the lower range less than 1.5 as shown in Fig 14. It is however, still tending to predict the category 1 better than 2 i.e.the model tends to identify liver patients better than non-liver patients.

After setting the family to binomial the results were obtained and summarized in Table 8. From the Table 8 it is clear that GAM model that uses Age, Gender, AAP, SGPT, TP variables were able to a high AUC value of 0.706 and an RMSE of 1.107. Since this is a classification problem AUC has more relevance than rmse.

Table 8: Result summary when the family is set to binomial

Variables	RMSE	AUC
Age, Gender	1.1	0.5332
Age, Gender, AAP	1.102	0.6144
Age, Gender, AAP, SGPT	1.107	0.707
Age, Gender, AAP, SGPT, TP	1.107	0.7061
Age,Gender, AAP,SGPT,TP,AG_ratio	1.106	0.70496

(2) Deep Learning Prediction

In this section the influence of the number of hidden layers, the number of neurons, the learning rate, epoch number and the activation functions on the model performance have been discussed. In order to arrive at the best model, after the optimal value of the parameter under consideration has been arrived at, that optimal value is kept constant

in subsequent study of other parameters.

(a) Study of different number of hidden layers

As the number of hidden layers changes from 4 to 5, it is evident from the graph that the RMSE increases and then dips for 6 layers, again increase for 7 and decreases and then decreases as the layers change 9 to 13. The lowest value of RMSE of 0.4915 was found corresponding to 4 layers as shown in Fig 15. In the subsequent studies the number of layers is kept constant at 4.

Table 9: Results of study based on different number of hidden layers

Hidden layers	No. of neurons	Learning rate	epochs	Activation function	RMSE	elapsed	user	system
4	**25**	**0.005**	**500**	**Tanh**	**0.4915**	**3.23**	**0.7**	**0.06**
5	25	0.005	500	Tanh	0.5403	3.31	0.79	0.05
6	25	0.005	500	Tanh	0.5316	78.66	1.29	0.06
7	25	0.005	500	Tanh	0.5847	136.39	1.63	0.08
8	25	0.005	500	Tanh	0.56	56	1.02	0.11
9	25	0.005	500	Tanh	0.5434	107.31	1.42	0.12
10	25	0.005	500	Tanh	0.5323	156.73	1.83	0.14

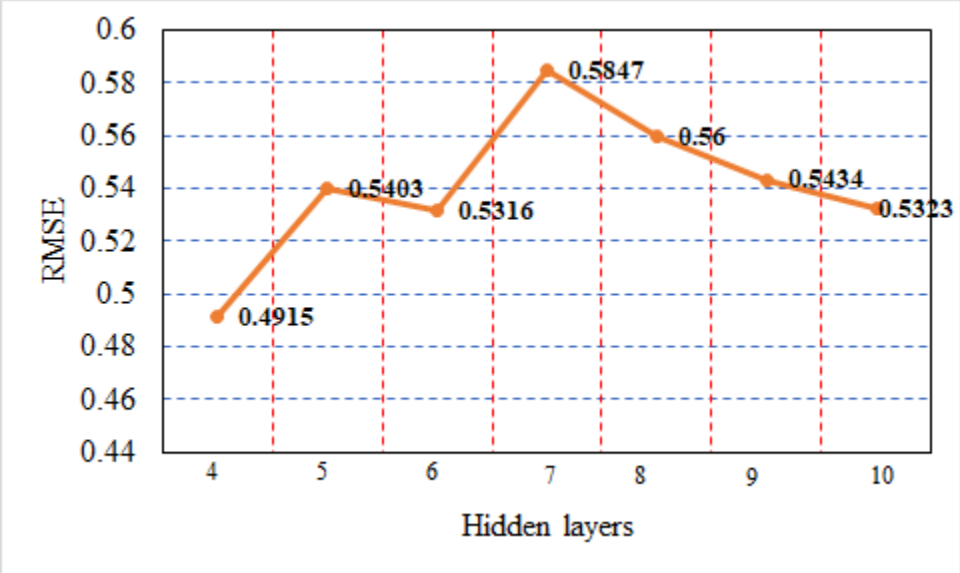


Fig 15: Variation of RMSE with the number of hidden layer

(b) Study of different number of number of neurons in each layer

Table 10: Results of study based on different number of neurons in each layer

Hidden	No. of	Learning	epochs	Activation function	RMSE	elapsed	user	system

layers	neurons	rate						
4	25	0.05	500	TanhWithDropout	0.4756	373.07	2.2	0.23
4	30	0.005	500	TanhWithDropout	0.4728	425.98	2.66	0.26
4	**40**	**0.005**	**500**	**TanhWithDropout**	**0.4498**	**495.29**	**3.11**	**0.31**
4	50	0.005	500	TanhWithDropout	0.4547	569.71	3.53	0.36

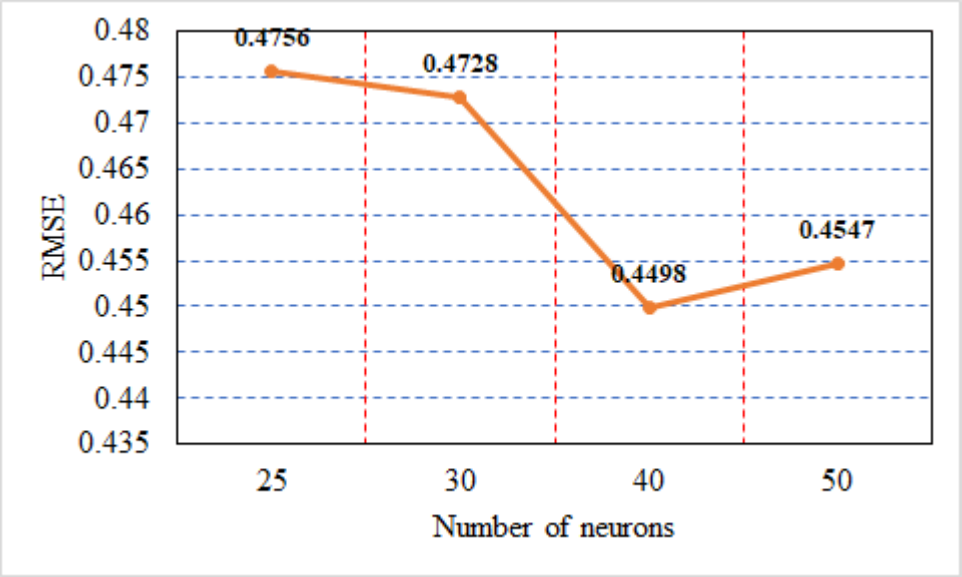


Fig 16: Variation of RMSE with the number of neurons

As the number of neurons was increased from 25 to 30 the RMSE value decreases slightly and as it Is increased to 40 the RMSE decrease by 4.86% to a value of 0.4498. In the subsequent steps of the study the number of neurons is chosen as 40 as seen in Fig 16.

(c) Study of different learning rates

Table 11: Results of study based on different learning rates

Hidden layers	No. of neurons	Learning rate	epochs	Activation function	RMSE	elapsed	user	system
4	40	0.005	500	TanhWithDropout	0.457	666.78	3.84	0.44
4	40	0.004	500	TanhWithDropout	0.454	871.43	4.25	0.53
4	40	0.003	500	TanhWithDropout	0.45	912.84	4.7	0.58
4	40	0.002	500	TanhWithDropout	0.456	948.48	5.02	0.62
4	**40**	**0.001**	**500**	**TanhWithDropout**	**0.439**	**989.37**	**5.52**	**0.69**

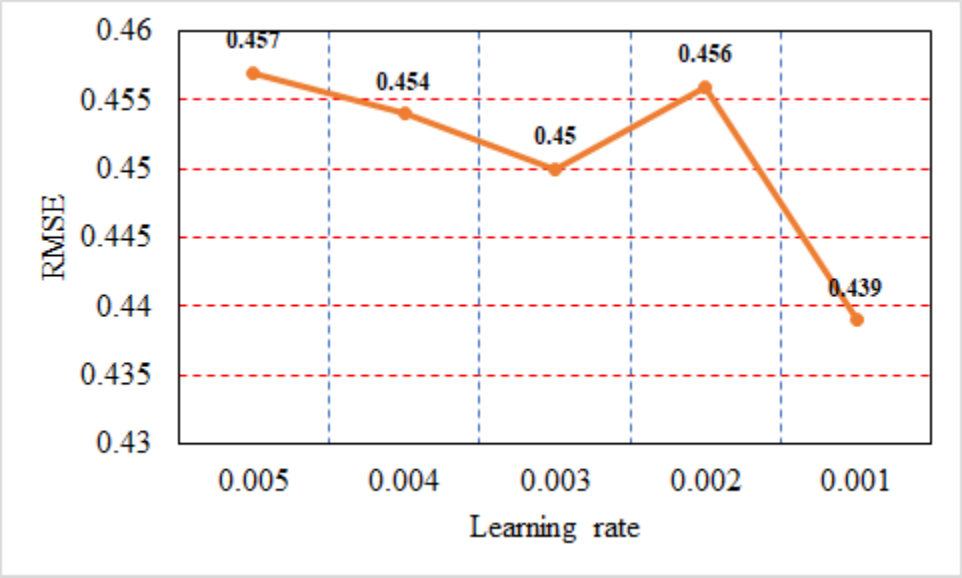


Fig 17: Variation of RMSE with the learning rate

The learning rate was varied from 0.005 to 0.001. The least value of RMSE of 0.439 was obtained corresponding to a learning rate of 0.001.However the change in RMSE when the learning rates were changed was small as seen in Fig 17.

(d) Study of the influence of varying epoch rates

Table 12: Results of study based on different epoch rate

Hidden layers	No. of neurons	Learning rate	epochs	Activation function	RMSE	elapsed	user	system
4	**40**	**0.001**	**500**	**TanhWithDropout**	**0.439**	**989.37**	**5.52**	**0.69**
4	40	0.001	600	TanhWithDropout	0.45	1176.8	6.25	0.76
4	40	0.001	700	TanhWithDropout	0.4412	1216.6	6.63	0.81
4	40	0.001	800	TanhWithDropout	0.4521	1461.5	6.98	0.86
4	40	0.001	900	TanhWithDropout	0.446	1793.9	7.39	0.87
4	40	0.001	1000	TanhWithDropout	0.445	1838.1	7.86	0.9

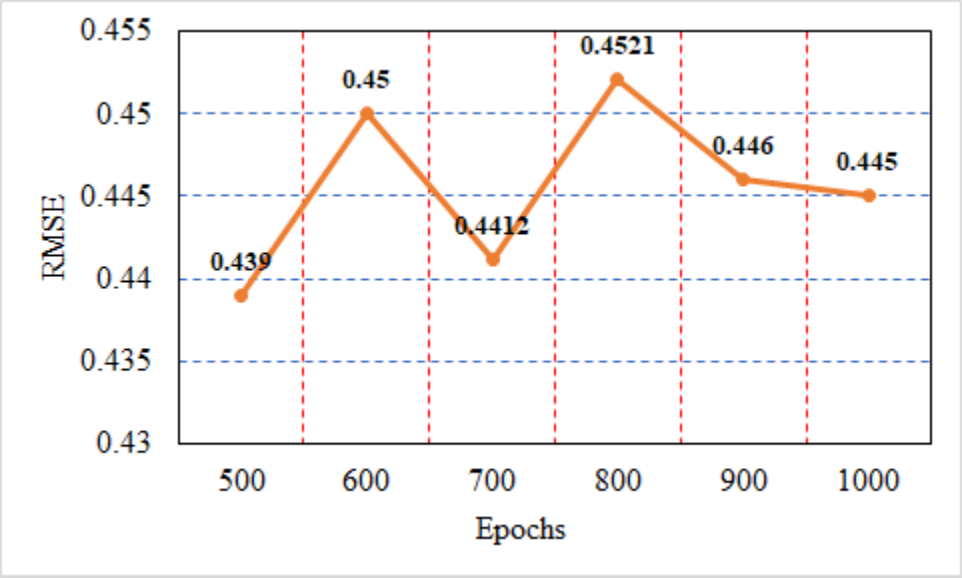


Fig 18: Variation of RMSE with the number epochs

As the number of epochs was increased from 500 to 1000 incrementing by 100 in every run, the least value of RMSE was obtained to be 0.439 corresponding to 500 epochs as seen in fig 18.

(e) Study of the influence of different activation functions

Table 13: Results of study based on different activation functions

Hidden layers	No. of neurons	Learning rate	epochs	elapsed	user	system
4	40	0.001	500	2130.3	8.56	1.01
4	**40**	**0.001**	**500**	**2171.1**	**8.97**	**1.08**
4	40	0.001	500	2208.2	9.34	1.14
4	40	0.001	500	2250.5	9.77	1.17
4	40	0.001	500	2290.5	10.17	1.2
4	**40**	**0.001**	**500**	**2355.9**	**10.66**	**1.26**

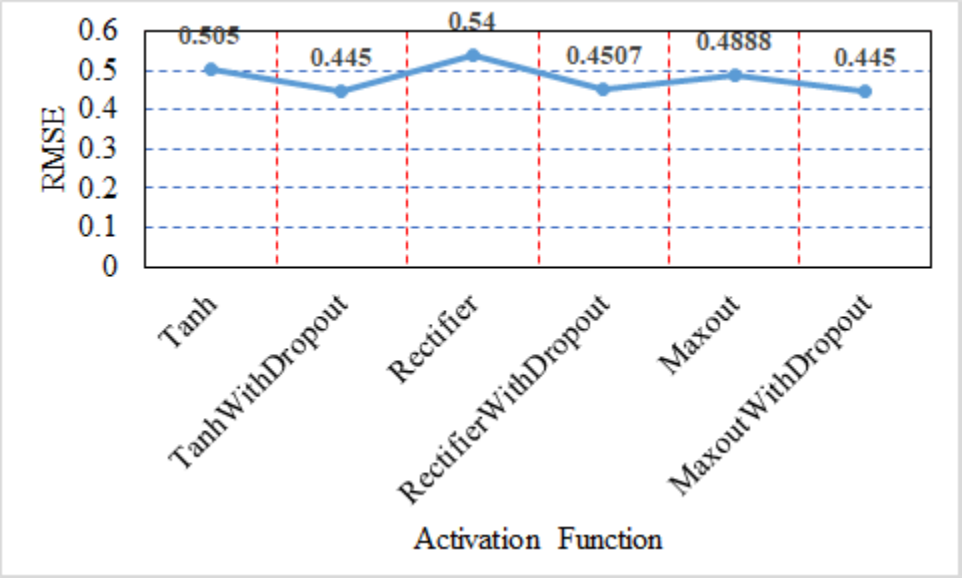


Fig 19: Variation of RMSE with the different activation functions

When trying to understand the influence of different activation functions on the RMSE value it was seen that both TanhWithDropout and MaxoutWithDropout obtained the same RMSE as seen in Fig 19. However, the former had reduced computation time and therefore can be though of as being less memory intensive. So, the best model is chosen to have TanhWithDropout as the activation function, 500 epochs, 4 layers, 40 neurons per layer and learning rate of 0.001. The target versus predicted plot is shown in the Fig below. However, from the plot it does not seem like a very robust model. The lowest RMSE achieved here is 0.45. The RMSE of the deep learning model is lower than that of the GAM model. Compared to the GAM model, the deep learning model seems to be performing badly as seen in the target versus prediction graph. RMSE may not be a good measure of estimation.

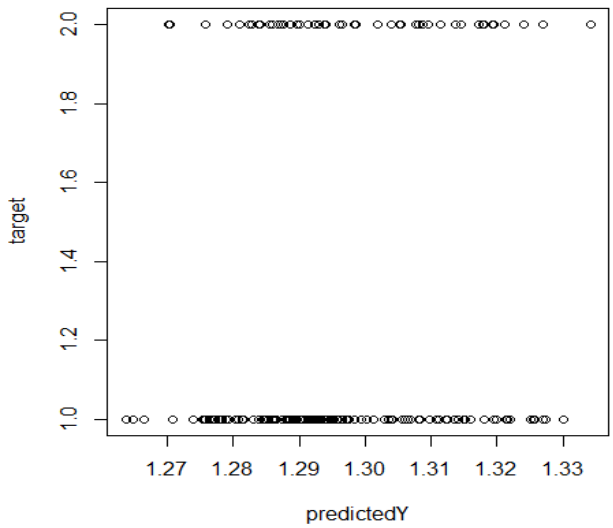


Fig 20: Target versus prediction for the best deep learning model

DISCUSSION AND CONCLUSION

When we have very large datasets, especially of the nature of big data it is a great advantage if we can have a representative smaller dataset that can represent the complete dataset. This will help us develop and deploy machine learning models that are computationally less expensive. In this paper, GA_NMM method was used to help develop the index dataset. It is a very good representation of the original dataset and poses as a promising method for bigger and more balanced datasets that can be collected from the medical field.

With respect to the data curing methods, it was observed that while some variables like Age were not sensitive to data curing, variables like AAP and SGPT were very sensitive to data curing. This information is very valuable in understanding the precision with such repots must be made and collected. Very small changes in these variables or their missingness can cause the patient to be misdiagnosed by the ML model.

Both GAM and Deep Learning models exhibited good performance and were able to classify the liver patients or category 1 better when compared to category 2. Though in terms of accuracy, these models may not seem to be working well, the recall rate is higher in the more severe category. In the field of medical science, the penalty of a liver patient being diagnosed as a non-liver patient is very high. Misclassifying a non-liver patient as a liver patient does not have that much implication with respect to saving

human lives. The problem of diagnosing liver patients is like cancer classification problem where the emphasis is more on classifying the cancer positive cases accurately. One of the limitations of the dataset has been that the number of instances in each category of the output variable is highly unbalanced. The use of synthetic sampling techniques like SMOTE that help enhance the dataset and reduce the imbalance may be able to help better classification accuracy and lies within the future scope of this work.

REFERENCES

1. Https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset). UCI Data Repository.
2. Nagaraj, K., and A. Sridhar. NeuroSVM: A Graphical User Interface for Identification of Liver Patients Kalyan Nagaraj 1* and Amulyashree Sridhar 2 1*. pp. 1–9.
3. Venkata Ramana, B., M. S. P. Babu, and N. . Venkateswarlu. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *International Journal of Database Management Systems*, Vol. 3, No. 2, 2011, pp. 101–114. https://doi.org/10.5121/ijdms.2011.3207.
4. Ramana, B. V., P. M. Surendra, P. Babu, and P. N. B. Venkateswarlu. A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis. *International Journal of Computer Science Issues*, Vol. 9, No. 3, 2012, pp. 506–516.
5. Abdar, M., M. Zomorodi-Moghadam, R. Das, and I. H. Ting. Performance Analysis of Classification Algorithms on Early Detection of Liver Disease. *Expert Systems with Applications*, Vol. 67, 2017, pp. 239–251. https://doi.org/10.1016/j.eswa.2016.08.065.
6. Abdar, M. A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms. *Cumhuriyet Science Journal*, Vol. 36, No. 3, 2015, pp. 3230–3241.
7. Hassoon, M., M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar. Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver Disease Prediction. *2017 International Conference on Computer and Applications, ICCA 2017*, 2017, pp. 299–305. https://doi.org/10.1109/COMAPP.2017.8079783.
8. Cho, I. H., and K. Porter. Modeling Building Classes Using Moment Matching. *Earthquake Spectra*, Vol. 32, No. 1, 2016, pp. 285–301. https://doi.org/10.1193/071712EQS239M.

9. Song, I., Y. Yang, J. Im, T. Tong, H. Ceylan, and I. H. Cho. Impacts of Fractional Hot-Deck Imputation on Learning and Prediction of Engineering Data. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 32, No. 12, 2020, pp. 2363–2373. https://doi.org/10.1109/TKDE.2019.2922638.