Computer-Aided Diagnosis for Liver Patient Classification

Harry Gollakota1\* Erik Wolf2\*

\*DePaul University - 1 E. Jackson Blvd., Chicago, IL 60604 USA

1[hgollakota@gmail.com](mailto:hgollakota@gmail.com)

2<erikgwolf@gmail.com>

<http://www.github.com/hgollakota/research-fds>

**Abstract**

Data science has much potential to benefit the field of Hepatology. We analyzed the Indian Liver Patient Dataset, which is a small, unbalanced dataset of two demographic features, and 8 chemical indicators/test results. Our goal was to classify based on these features whether a patient had liver disease or not. We can show that Principle Component Analysis with Logistic Regression achieves classification accuracy of >72%, while PCA and k-Nearest Neighbor algorithm exceeds previous benchmark accuracy. Our multi-perceptron neural network achieves the greatest accuracy of 75.5%. These conclusions demonstrate the exciting potential Data Science techniques have to aid in diagnostics.

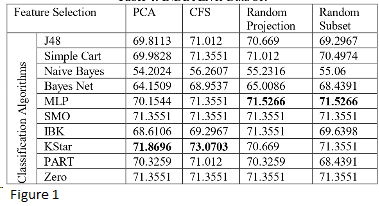
**I. Introduction**

Data Science has enormous potential with regards to application in real-world problems traditionally addressed by human perception. With this increase in awareness of data science techniques comes new chances to apply them in wide-ranging fields. The biomedical field is a prime example, with large sets of patient data providing opportunities to discover new patterns using data mining techniques. Recent research has seen applications for Neural Networks in predicting the onset of Alzheimer’s Disease[1]. In the field of Hepatology, Neural Networks and Deep Learning have been used to help predict the occurrence of liver tumors[2].

The human liver is essential to many bodily functions, such as synthesizing proteins, decomposing red blood cells, detoxification and producing hormones. There are 4.5 million adults in the United States with diagnosed liver disease, of which there are over 100 varieties. Liver disease accounted for 40,545 deaths in 2016[3]. Unfortunately, “...up to 50 percent of individuals with underlying liver disease have no symptoms”[4]. While some risk factors for liver disease are easier to identify, such as obesity or drug-use, other risk factors such as genetics are more subtle[5]. However, the liver has some unique properties among the major organs, such as having the ability to regenerate from some damages. This underscores the need to catch liver-disease early. The question becomes: how do we catch liver disease earlier, quicker or more efficiently so a timely intervention can take place?

The goal of this paper is to demonstrate how effective classification techniques from data science such as Logistic Regression, kNN algorithm and a Multi-Perceptron Neural Network are at identifying whether a patient has liver disease. This classification takes place on a smaller unbalanced dataset, with the hope that a larger, balanced dataset would yield even better results.

**II. Literature Review**

Much of the reviewed literature originated with the creators of the dataset. The original paper[6] attempted to compare this dataset, originating from a few labs in Andhra Pradesh, India, with another well-known dataset known as the BUPA dataset[7]. Unfortunately, it has since been shown that the BUPA dataset was not actually classifying whether a patient had liver disease or not, so the paper’s conclusion that there is a difference in populations between the United States and India are no longer justified. However, there is no reason to doubt the analysis on the new dataset, known as the Indian Liver Patient Dataset (ILPD). 

Further literature from the same authors shows the accuracy of several different classification methods such as tree-based algorithms, statistical based algorithms, neural networks, lazy learners, rule based algorithms and support vector machines[8]. This accuracy provided published benchmarks to attempt to meet or exceed, as seen in Figure 1[8]. The best technique was found to be KStar with Correlation-based Feature Selection.

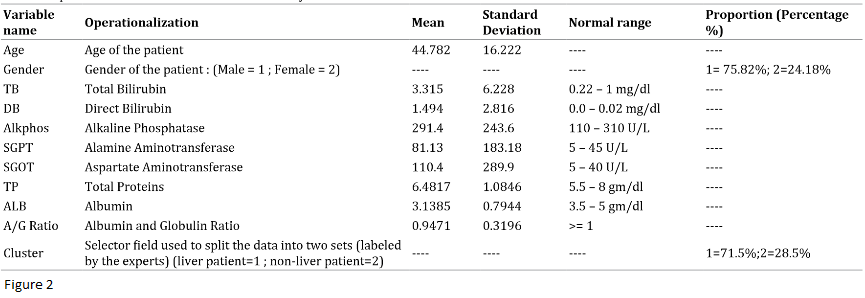
Imbalanced data has long been recognized to affect the accuracy of classifiers[9]. Some of the original authors acknowledged some of the shortcomings of the ILPD, by creating an enhanced ILPD in 2017 that included balancing by gender as well as the label[10]. This dataset is unfortunately not available for public use and the author did not reply to a request by email to get access to the dataset. While the accuracy of classification is not tested, the work shows that there is a difference in several variables dependent on age as well as gender.

A recent 2018 study isolated the variable Alkaline Phosphatase (AlkPhos) which is one of the factors in ILPD. Alkaline Phosphatase is shown to vary based on other factors and is therefore identified as a variable of importance. Further, “...the interaction effects and the variance determinants of alkaline phosphatase are completely new inputs in the liver disease literature”[11]. If these claims are true, the ILPD might yield new information beyond just an exercise in applying ML algorithms. A 2017 paper found that a multi-level perceptron neural network was the most accurate classifier of ILPD[12].

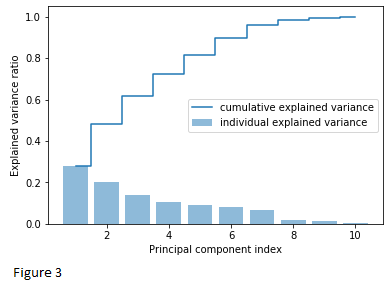
**III. Methodology**

Analysis of ILPD was partially conducted using Python 3 with the popular packages Pandas and Numpy used for data organization and Scikit-Learn used for training and evaluating classification models. IBM SPSS 25 was further used for analysis and the creation of the multi-level perceptron neural network.

The Indian Liver Patient Dataset[13] consists of 583 instances representing the records of individual liver patients taken from labs in Andhra Pradesh, India. The label is whether the patient has liver disease or not, of which 416 were in the positive class (has liver disease) while 167 were not. ILPD contains 10 factors, two of which are demographic information of Age and Gender, while 8 were the results of tests conducted on the patient. These tests were of Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin and the Albumin and Globulin (A/G) Ratio. Analysis of these individual factors is seen in Figure 2[11].



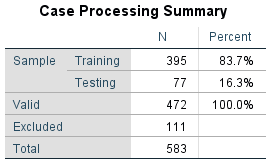
Data Preprocessing was relatively minor. Observation revealed 4 instances that contained NA values, or <1% of ILPD. These instances were dropped. The factor Gender was converted into the binary variable “is\_Female” (1:yes,0:no). The label was also changed to a binary variable “has\_Disease” (1: yes, 0: no). The data types of all the variables were converted from int64 to float64, which was necessary for some Python packages but did not result in any loss of information.

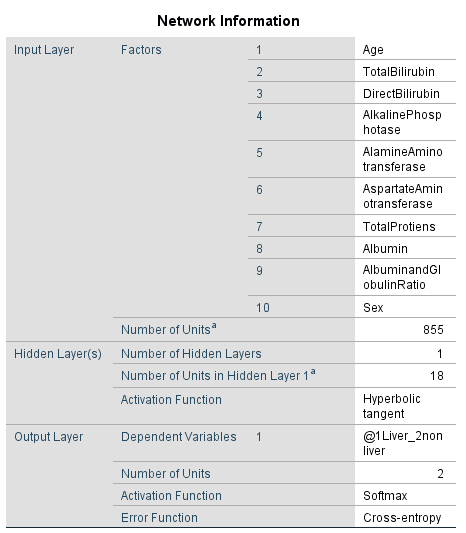
Data Reduction consisted of Principle Component Analysis (PCA). To select the number of Principle Components to consider, the dataset was divided into X (features) and y (label). X was standardized using Scikit-Learn and a correlation matrix was computed on X and eigenvalues graphed, with the purpose of identifying how many components explained an acceptable amount of the variance of the initial data. As seen in Figure 3, the first 7 components explain >95% of the variance, so 7 components were selected. Then X was transformed using PCA from Scikit-Learn’s Decomposition package into X\_pca.

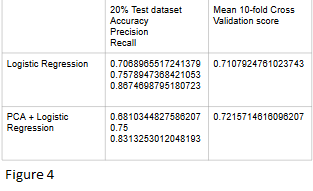
X and y were then divided into train:test in the ratio 80:20 with Scikit-Learn, using stratified sampling based on y in order to balance the positive instances between the two sets. As the minority class was not the positive class, oversampling techniques were not used, though this may be an avenue to explore further.

The first analysis conducted was a Logistic Regression. While many forms of regression are used when the label is a continuous variable, Logistic Regression is applied when the label is a discrete binary variable, as is the case for ILPD, and is therefore an appropriate classification technique. Logistic Regression also has a long history of being used in the medical field and is widely accepted in literature. A logistic regression model from Scikit-Learn was trained using X\_train and y\_train. This model then used X\_test to generate a y\_pred, which was compared against y\_test for accuracy. These steps were repeated with X\_train\_pca and X\_test\_pca for a PCA+Logistic Regression analysis as well. In addition to the train-test validation technique, 10-fold cross validation was also used to measure accuracy because the number of samples were within the range where this is appropriate.

The second analysis conducted was using the K-Nearest-Neighbors algorithm (kNN), as well as PCA+kNN. The data retained the same train/test split as used in Logistic Regression. kNN classifies a point based on the majority of its k nearest neighbor points determined by Euclidean distance. The accuracy of this prediction was measured for 3 to 33 nearest neighbors to determine the optimal number of neighbors for X\_train and X\_train\_pca. This was then the k used with X\_test and X\_test\_pca to generate y\_pred, which was compared against y\_test for accuracy. The analysis was run again using 10-fold cross validation as well.

The third analysis conducted with IBM SPSS 25 was a multi-layer perceptron neural network (MLP-NN). The case processing produced random sampling validation from the total 583 cases. It was further recognized that 111 cases were excluded from the MLP-NN as they were flagged as being bias. The final sample used in the MLP-NN was 472 cases split into 395 (83.7%) Training and 77 (16.3%) Testing. The input layer consisted of 10 factors; Age, TotalBilirubin, DirectBilirubin, AlkalinePhosphotase, AspartateAminotransferase, TotalProtiens, Alumbumin, AlbuminandGlobulinRatio, Sex. There were 855 units after the removal of bias units. Only 1 hidden layer was created, this hidden layer was composed of 18 units. The output layer was the Liver\_nonliver classifier which consisted of two units, 0 = No liver ailment, 1 = liver ailment detected.

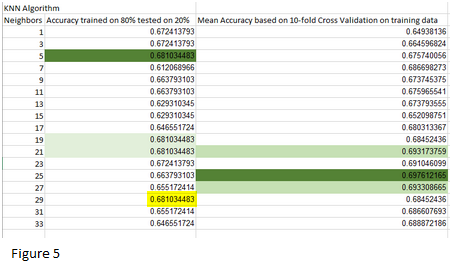


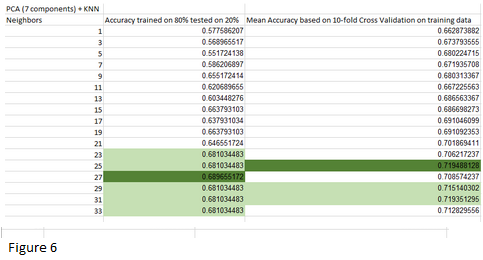


**IV. Results**

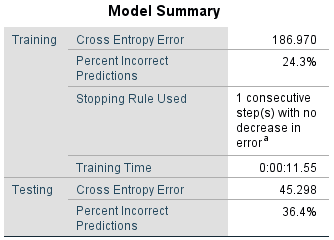
The results for Logistic Regression as well as PCA+Logistic Regression are seen in Figure 4. Precision and Recall were measured in addition to accuracy for the train/test split data, while only accuracy was measured as the Mean 10-fold cross validation score. Mean 10-fold cross validation was higher with regards to both analysis. PCA resulted in a lower accuracy for the selected test data, but resulted in the highest accuracy for 10-fold Cross Validation of 72.16%

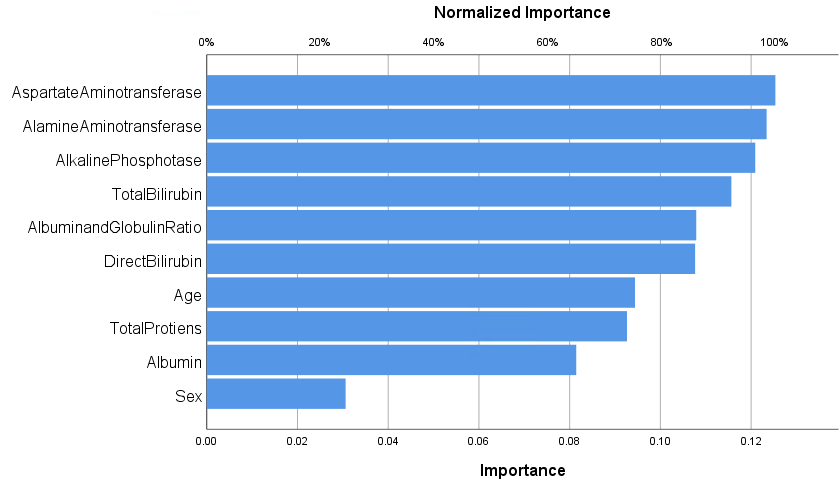
The results for kNN are seen in Figure 5, and the results for PCA+kNN are seen in Figure 6. As stated in Methodology, k=3 to 33 was trained and accuracy measured for both the test data as well as through Mean 10-fold Cross Validation.

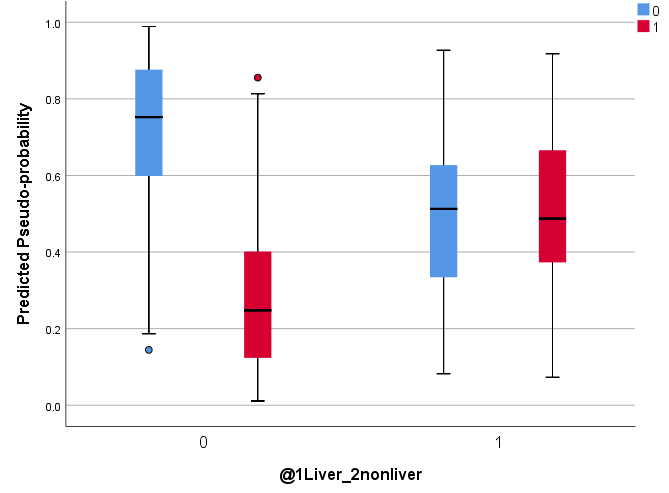




The optimal number of neighbors for kNN on the test data was seen to be either 5, 19, 21 or 29, as they all resulted in the same accuracy. However, of those options, k=5 is much smaller than all the other optimal neighbors for all the other kNN Analysis, so that one seems unlikely. The optimal number of neighbors for kNN using Mean 10-fold Cross Validation was found to be 25 neighbors. PCA+kNN had an optimal 27 and 25 neighbors for the test data and Mean 10-fold Cross Validation respectively. PCA+kNN, as expected, produced the most accurate results of the four, with Mean 10-fold Cross Validation accurately classifying 71.95% of the data.

The multi-perceptron neural network accuracy of classifying the output layer was 75.7% correct. 



Variables of normalized importance are very closely linked, Aspatate Aminotransferase, Alamine Aminotransferase and Alkaline Phosphatase.

**V. Conclusion and Future Work**

Our results for PCA+kNN just barely beat the equivalent benchmark in Figure 1. PCA+Logistic Regression was seen to have higher accuracy than all except one test conducted in the benchmark. However, the best results came from the multi perceptron neural network, and confirmed the results published by Abdar et. al [12].

Our results demonstrate the classification ability of these algorithms on a small, unbalanced dataset. We believe that a larger, balanced dataset constructed from multiple sources, such as one that could be used to train an actual production diagnostic model, would yield much higher accuracy. Future work could consist of building such a comprehensive dataset and training a neural network to be shared amongst medical professionals to aid and confirm in diagnostics.

Other future work might involve analysis of individual variables, such as the analysis conducted on Alkaline Phosphate[11]. To reiterate, our results were generated from a relatively small, unbalanced, publicly available dataset. We predict that the sort of dataset available to private industry or through dedicated sample aggregation by a public entity could yield much more exciting and conclusive results and could be an invaluable tool in liver disease diagnostics.

**VI. References**

1. <https://arxiv.org/ftp/arxiv/papers/1905/1905.00931.pdf>
2. <https://arxiv.org/pdf/1902.07971.pdf>
3. <https://www.cdc.gov/nchs/fastats/liver-disease.htm>
4. <https://uihc.org/health-topics/liver-disease-frequently-asked-questions>
5. <https://liverfoundation.org/5-facts-about-liver-disease/>
6. <https://github.com/hgollakota/Research-fds/blob/master/Papers/ACriticalComparativeStudyofLiverPatientsfromUSAandINDIAAnExploratoryAnalysis_BendiVenkataRamana%2CProf.M.SurendraPrasadBabu%2CProf.N.B.Venkateswarlu.pdf>
7. <https://archive.ics.uci.edu/ml/datasets/liver+disorders>
8. <https://github.com/hgollakota/Research-fds/blob/master/Papers/LiverClassificationUsingModified%20RotationForestBendi_Venkata%20Ramana%2CProf.M.SurendraPrasad%20Babu.pdf>
9. <https://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf>
10. <https://github.com/hgollakota/Research-fds/blob/master/Papers/criticalanalysisindianliverpatient_2017.pdf>
11. <https://github.com/hgollakota/Research-fds/blob/master/Papers/AlkalinePhosphataseDeterminantsofLiverPatients_RabindraNathDas%2CSabyasachiMukherjee%2CIndrajit%20Sharma.pdf>
12. <https://link.springer.com/article/10.1007/s40846-017-0360-z>
13. <https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29>