Predictive Modeling for Hepatitis Disease Detection with Machine Learning

ARAVA JAYA LAKSHMI1, CHILKAPATI SAI SUDHA2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE-IoT, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ The selection of the most effective instrument for the diagnosis and detection of hepatitis, as well as for the estimation of the remaining life expectancy of patients with hepatitis, is the purpose of this work. An investigation into the similarities and differences between various machine learning techniques and neural networks was carried out for the purpose of this work. The accuracy rate and the mean square error are the two components that make up the performance measure. The classification and prediction methods for diagnosing hepatitis disease were thought to be the Machine Learning (ML) algorithms such as Support Vector Machines (SVM). A cursory investigation into the aforementioned algorithms was carried out in order to see how accurately disease diagnosis may be predicted.

1.INTRODUCTION

Hepatitis B, an illness of the liver that is caused by the hepatitis B virus (HBV), has continued to be a problem in terms of public health around the world [1–3]. The transmission of HBV requires only the exchange of fluids between an infected individual or a viral carrier and a person who is not infected with the virus. Recent research has shown that hepatitis B virus (HBV) is 100 times more contagious than HIV/AIDS and is still the principal cause of liver cancer [4]. There are currently two billion people infected with HBV all over the world [1,3]. HBV has infected one person in every three people. Every year, there are approximately 1.5 million people who become infected, and there are over 300 million people who are chronically sick [1,2]. It is estimated that 820,000 people pass away every year as a result of problems related to HBV, which equates to an average of two deaths occurring every minute due to HBV [1,3,5]. Many

countries in Africa and Asia are still considered to be high - endemicity areas, with theprevalence of HPV in West Africa estimated to be 8.83% [6]. Despite the significant advancements that have been made in HPV vaccinations, many of these countries are still regarded as high - endemicity areas. In the meanwhile, researchers believe that poor vaccination rates among adults older than 18 are to blame for the high prevalence of the disease in developing nations [7].

People who are infected with HBV require appropriate treatment in order to maintain their health and keep living. Despite the availability of effective preventative vaccines and diagnostic tools, HBV continues to be a substantial threat to public health, resulting in a large number of deaths [8–10]. In recent years, it has been demonstrated that antivirals can eradicate HBV infection; nevertheless, it has also been shown that discontinuing medication can result in a resurgence of the virus [10]. Early discovery can help limit the consequences of the disease, which can be severe, and can also reduce the chance of transmission through contact with the blood and other bodily fluids of an infected person [11]. When it comes to classifying patients who are infected with HBV, conventional medical procedures use a variety of biochemical and biological variables to estimate the extent of liver

damage and viral activity [12]. Other traditional statistical methods have also shown, according to studies [13–15], that they are capable of predicting acute and chronic HBV with clinical data. Because of the multi collinearity problem inherent in high-dimensional medical data, such predictions made with the usual statistical method could contain an element of bias.

2.LITERATURE SURVEY

[1]   In spite of all the attempts that have been made to standardize the process, medical diagnosis is still considered to be an art. This is due to the fact that medical diagnosis involves a competence in handling ambiguity that is unobtainable with the computational gear available today. Even though the idea of artificial intelligence has been around for some time, recent advances in the field of computer science have led to its recognition as an emerging field of technology. The usage of artificial intelligence can be found in a variety of fields, including education, business, medicine, and manufacturing, to name a few. The purpose of the planned study was to evaluate the possibilities of artificial intelligence techniques, primarily for use in the medical field. It is feasible that the algorithms of neural networks could provide an improved answer for various medical issues. In this work, an

investigation of the use of artificial intelligence in the traditional method of diagnosing hepatitis B was carried out. The approach that was taken in this study consisted of using an intelligent system that was based on the concept of logical inference to make a determination regarding the form of hepatitis that is most likely to manifest itself in a patient, specifically whether or not it is hepatitis B. After that, Kohen's self-organizing map network was used to analyze the hepatitis data in order to make predictions regarding the severity level of the patient's Hepatitis B infection. The results show that a SOM, which is a type of unsupervised network, was employed as a classifier to determine how well Hepatitis B could be predicted. We came to the conclusion that the suggested model provides a quicker and more accurate prediction of hepatitis B, and that it also functions as a potential tool for the routine prediction of hepatitis B based on clinical laboratory data.

[2] Patients who have chronic hepatitis C (CHC) and have liver fibrosis have a poor prognosis for their condition. We investigated the efficacy of non-invasive indicators and liver biopsies in determining the likelihood of morbidity and mortality in CHC patients. Patients suffering from hepatitis are the ones who require ongoing particular medical care in order to bring down the mortality rate. It is

possible to classify patients and provide predictions regarding their life expectancy by making use of the findings of clinical tests in conjunction with machine learning technology such as Support Vector Machines (SVM). However, we are unable to guarantee that every single feature value in the data is correlated to every other value in the set. As a result, we make use of Wrapper Methods in order to get rid of noisy features prior to classification. This study demonstrates that combining feature selection methods before the classification process can boost the accuracy of predictions between sets of data

3.PROPOSED SYSTEM

Diagnosing hepatitis medically is a challenging undertaking due to the large number of variables that must be taken into account. Since the accuracy of an automatic system can be immensely helpful for the detection of hepatitis, it has been used alongside clinic tests to aid in the early diagnosis of hepatitis diseases. Consequently, the majority of researchers have concentrated on extracting and selecting the optimal number of features, which they have then fed into the most thoroughly tried and tested algorithms; while this approach yields promising results in the lab, putting it into practise would require patients to undergo a large number of tests before receiving a definitive diagnosis, which could be

prohibitively expensive and time-consuming. In order to construct a training model with the most relevant features, this work seeks to explain the relationships between the various features included in the dataset. With this data in hand, developers may create a web app for patients to use on their own, with only the tests they need, cutting down on time and money spent on unneeded procedures.

3.1 IMPLEMENTATION

Dataset exploration

The dataset, as well as the data dictionary of the properties involved, are studied in the Pythonenvironment.

Data mutilation

It is necessary to estimate missing values in some factors since inadequate data prohibits most interpretations from being made. This term refers to the estimation of incomplete data in some variables. In the case of a variable, the missing data are substituted with the value of the mean, and in the case of a class label, the missing values are substituted with the value of the mode.

Feature Selection

It is crucial for any type of predictive modelling, and it is carried out with the goals of preventing multi collinearity,

removing redundant attributes that are closely correlated with one another, and improving the model's performance. We eliminated characteristics that aren't necessary for disease identification by employing a method known as backward selection. First, we consider each of the model's attributes, and then, using the p-value as a guide, we delete some of them. This helps in establishing the significance of the data when running a test in statistics to determine whether or not the null hypothesis is true. The characteristics that had a p-value that was more than 0.05 were eliminated from the model, and the model was then re-fitted with the variables that were left over. This method was carried out multiple times until each of the model's pre-existing variables reached a level that could be considered meaningful. The most recent value of R square was recorded at the conclusion of each repetition. to determine the proportion of the total variation that can be accounted for by only those independent variables that have a significant bearing on the prediction of the target variable..

Model fitting and Testing

After feature selection was complete, five different classification algorithms, namely Logistic Regression, Decision Trees, Random Forest, Support Vector Machine

(SVM), and Adaptive Boosting, were applied to the data using the selected feature, and the accuracy of their respective predictions was evaluated utilising the Train/Test split methodology. Because the

test size for the comparison was set at 0.1, this indicates that ninety percent of the dataset was utilised for the training of the classifier, while the remaining ten percent was utilised for testing.

4.RESULTS AND DISCUSSION
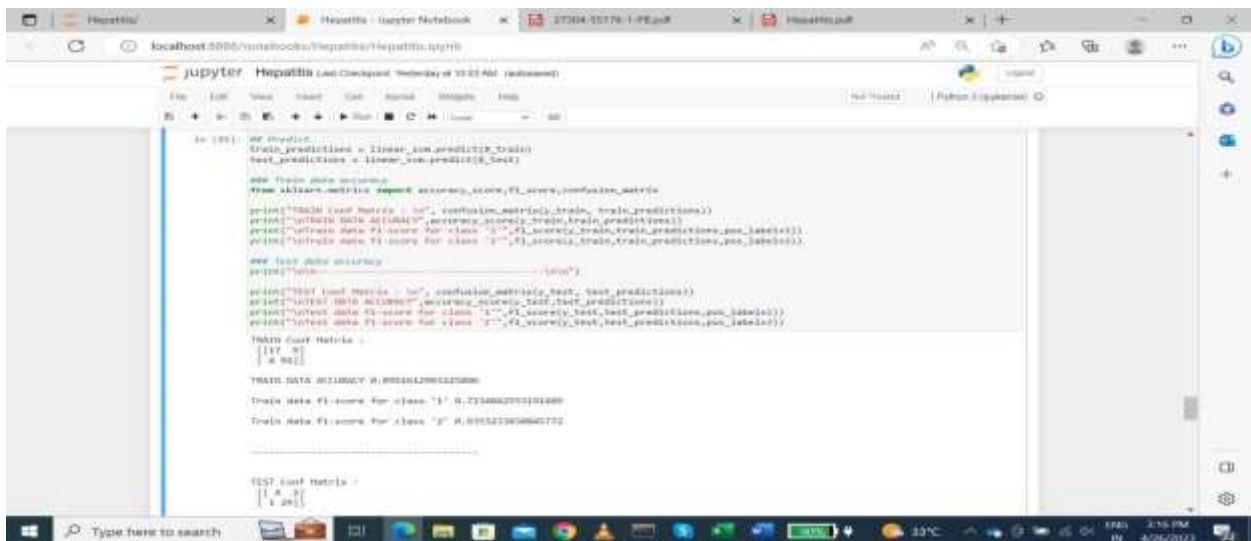
Fig:1 Showing the matrix of train and test data and finding the accuracy of the train and test data
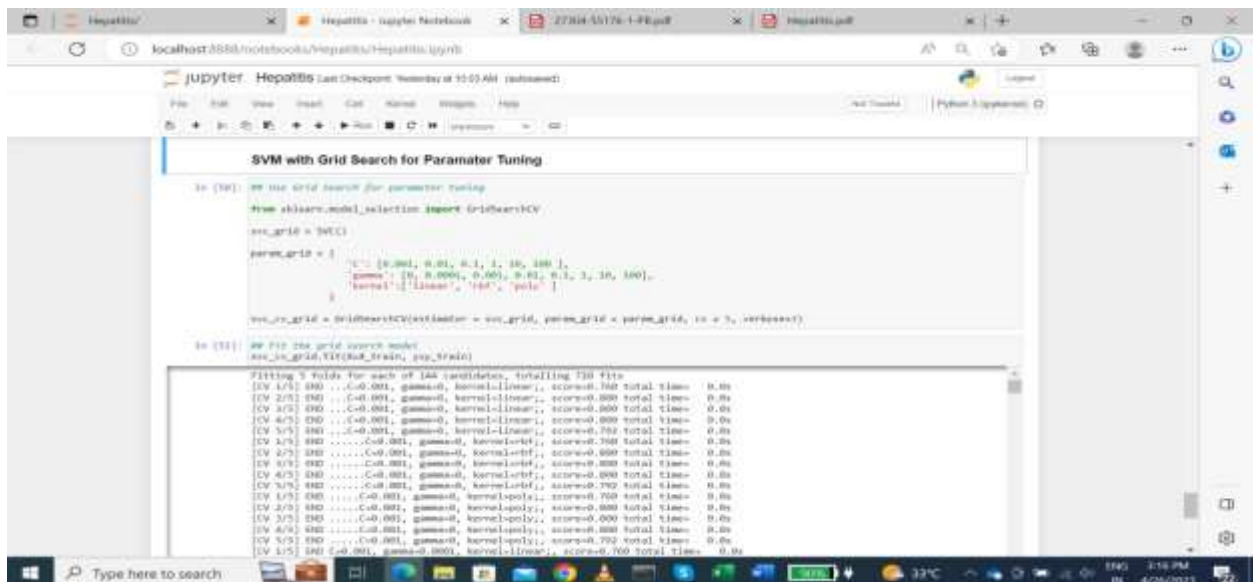
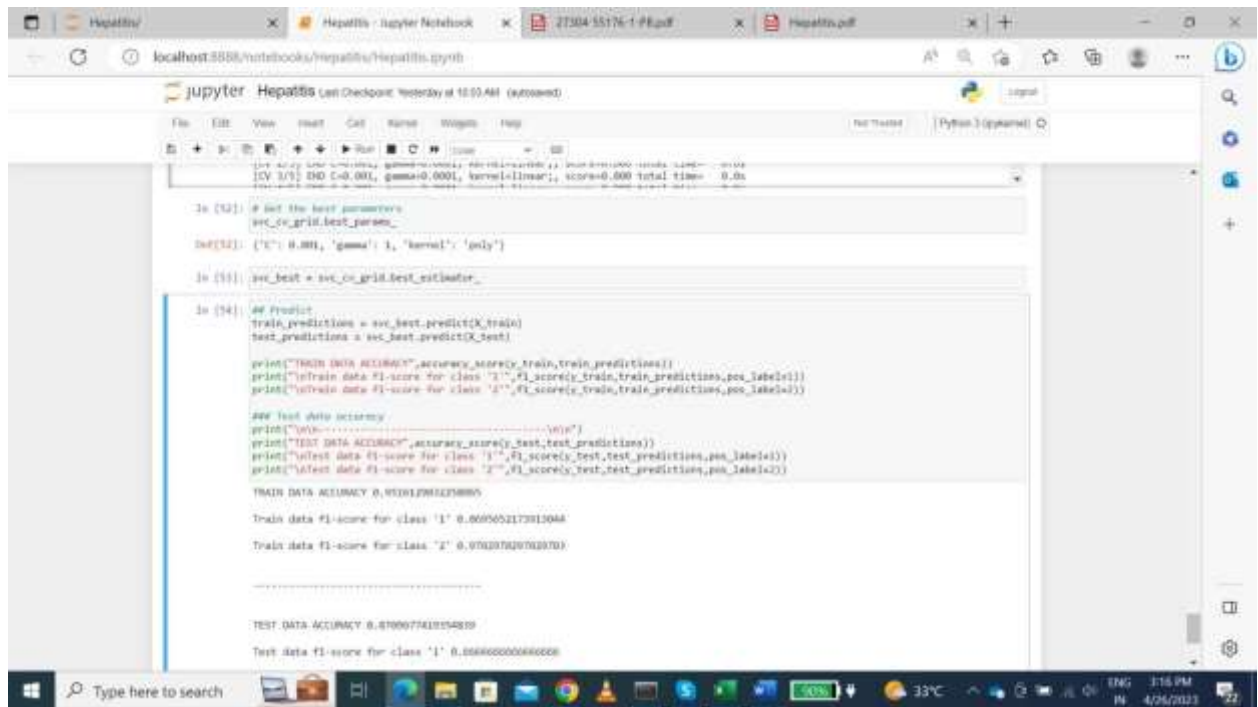**Fig:2 Showing the grid fit of train and test values**

Fig:3 Finding the accuracy of two different classes of training and testing data
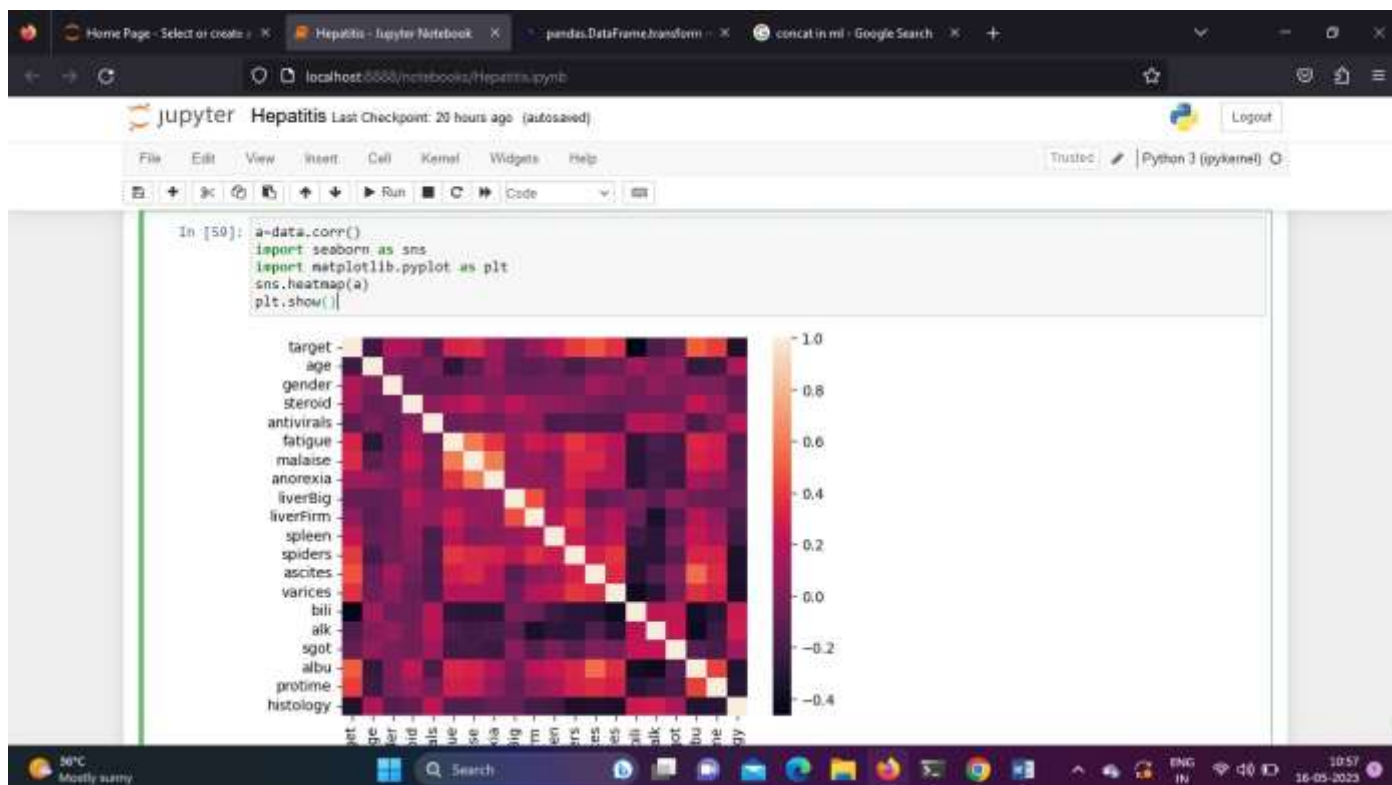

Fig4 : Importing a libraries to representing graph representation of head map

5.CONCLUSION

In this particular research project, the diagnosis of hepatitis was carried out by utilising a wide range of machine learning strategies and neural networks. We examined the accuracy of a number of different ML and SVM methods using the same data set in order to establish which of these approaches was most reliable for diagnosing hepatitis disease. We have used the Support Vector Machine (SVM) approaches in order to make an appropriate prediction regarding the sickness. Based on the findings of this study, it can be deduced that has the highest level of prediction accuracy (96%) and the lowest level of mean square error (MSE) among all the models that were examined. Future research will make use of a concept conceptually similar to this one, namely the utilisation to forecast the emergence of other diseases

REFERENCES

[1]   Ghumbre S. U.; Ghalot A.A, "Hepatitis B Diagnosis using Logical Inference And Self Organizing Map", 2008 ; Journal of Computer Science ISSN 1549-3636.

[2]   M. A. Chinnaratha, G. P. Jeffrey, G. Macquillan, E. Rossi, B. W. D. Boer, D. J. Speers, and L. A. Adams, "Prediction of morbidity and mortality in patients with chronic hepatitis c by non-invasive liver fibrosis models," Liver International, vol. 34, no. 5, pp. 720–727, 2014.

[3]   Roslina, A. H., and A. Noraziah. "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method." In 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, vol. 5, pp. 2209-2211. IEEE, 2010.

[4] G. H. Haydon, R. Jalan, M. Alakorpela, Y. Hiltunen, J. Hanley, L. M. Jarvis, C. A. Ludlum, and P.

C. Hayes, "Prediction of cirrhosis in patients with chronic hepatitis c infection by artificial neural network analysis of virus and clinical factors," Journal of Viral Hepatitis, vol. 5, no. 4, pp. 255–264, 2010.

[5]   Atif Khan, John A. Doucette, Robin Cohen, "Integrating Machine Learning

into a Medical Decision Support System to Address the Problem of Missing Patient Data", 2012 IEEE DOI 10.1109/ICMLA.2012.82.

[6] Uhmn, Saangyong, Dong-Hoi Kim, Young-Woong Ko, Sungwon Cho, Jaeyoun Cheong, and Jin Kim. "A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis." Expert Systems 26, no. 1 (2009): 60- 69.

[7] KayvanJoo, Amir Hossein, Mansour Ebrahimi, and Gholamreza Haqshenas. "Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms." BMC research notes 7, no. 1 (2014): 565. Vijayarani, S., and S. Dhayanand. "Liver disease prediction using SVM and Naïve Bayes algorithms." International Journal of Science, Engineering and Technology Research (IJSETR) 4, no.4 (2015): 816- 820.

[8] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVMSA)." Computer methods and programs in biomedicine 108, no. 2 (2012): 570-579.

[9] Uttreshwar, Ghumbre Shashikant, and A. A. Ghatol. "Hepatitis B diagnosis using logical inference and generalized regression neural networks." In 2009 IEEE International Advance Computing Conference, pp. 1587-1595. IEEE, 2009.

AUTHOR PROFILES