# Extreme Gradient Boosting (XGBoost) and Support Vector Machine (SVM) models for Hepatitis C Prediction

**Alber S. Aziz*, Haitham Rizk Fadlallah**

Information System Department; October 6 University; Cairo, Egypt

Emails: Albershawky.csis@o6u.edu.eg; Haitham.rizk.csis@o6u.edu.eg

**Abstract**

Hepatitis C Virus (HCV) is a worldwide epidemic. The World Health Organization estimates that annually between 3 and 4 million instances of HCV are recorded. People with HCV would benefit from knowing their illness stage earlier thanks to accurate and timely prognoses. Different noninvasive blood biochemical indicators and patient clinical data have been utilized to determine the disease phase. As a substitute for the invasive and sometimes harmful liver biopsy, machine learning approaches have shown useful in diagnosing each phase of this chronic liver disease. To accurately estimate HCV using sparse weather information, this work offers two machine learning (ML) methods: The Support Vector Machine (SVM) and a simple tree-based ensemble approach called Extreme Gradient Boosting (XGBoost). The two models are applied to real-world data on HCV. The dataset contains 13 variables and 615 cases. The results showed the SVM achieved more accuracy than the XGBoost. The SVM gets 93.5% accuracy and XGBoost gets 90.23% accuracy.

**Keywords:** Machine Learning (ML) Models; Hepatitis C; Prediction; Support Vector Machine (SVM); XGBoost.

## 1. Introduction

Chronic hepatitis, liver cirrhosis, and hepatocellular cancer are all linked to hepatitis C, an infection of the liver triggered by the hepatitis C virus (HCV). Anti-HCV antibody frequency varies throughout nations, with reports of particularly high levels in Egypt. Individuals diagnosed with HCV must have liver fibrosis staging evaluation as part of their ongoing care. It is crucial for tracking the disease's forecasting, pinpointing when medication should begin, planning for its course, and anticipating the patient's reaction to any given intervention[1]–[3].

When it came to determining the severity of liver fibrosis, a liver biopsy was the method of choice. Nevertheless, the drawbacks of liver biopsies, such as their susceptibility to sample mistakes and their invasive its very nature, as well as their extremely expensive for most patients, particularly when done on a periodic basis for tracking the progression of disorders, pose a potential concern[4]–[6].

Because of the risks associated with biopsies, non-invasive approaches have become more popular in recent years as an option for assessing chronic liver disorders[7], [8].

The use of AI in the medical field is expanding quickly. More focus has been paid in the past few years to initiatives involving AI and health than a variety of others from the global economy. Artificial intelligence (AI) is used in the medical field for robotic diagnostic procedures and patient monitoring. If AI were used more often in prescribing medicine, it might automate a large portion of the procedure, freeing up time for doctors to focus on more complex cases[9], [10].



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

Data analysis and ML rely heavily on categorization to infer the categories into which various data objects may fall. Data mining is the process of using analytical tools to mine databases for useful insights and patterns that can then be used to make predictions[11]–[13].

Data is collected and then used by ML models to construct algorithms from which intelligent decisions may be made. In order to build models for forecasting from clinical files, for example, to ascertain whether or not a patient has Hepatitis C through clinical and biochemical data, numerous researchers have explored ML techniques in many fields during the past many years[14]–[16].

This study used ML models to predict the HCV. This study used two common ML models such as SVM and XGBoost. After applying these models there are some preprocessing data are applied such as deal with missing valued and encoding dataset. The missing values are filled with the mean of data. Then applying the two models to predict the HCV.

## 2. Hepatitis C virus

The Hepatitis C virus (HCV) is a pandemic virus that has affected people all over the world. Direct contact with blood or other bodily fluids from an infected individual is all that's needed for the virus to spread. Hepatitis C is a worldwide epidemic, as stated by the WHO. About 1.5 million individuals are infected with HCV every year, and the World Health Organization estimates that 58 million people worldwide have chronic HCV infection. When comparing the prevalence of this virus in wealthy nations like Europe and North America, poor developing countries like Asia and Africa have the highest rates. Moreover, the prevalence of chronic illnesses is greater in nations like Pakistan, China, and Egypt.

More research is required to determine the rate of HCV transmission among Egyptian healthcare providers. Healthcare workers (HCWs) are routinely monitored because of the high risk of exposure to various biological agents. Thus, in Egypt, those who work in healthcare and have frequent contact with patients are at a higher risk of contracting HCV and other blood-borne viruses [1, 2].

As a result, noninvasive technologies for HCV diagnosis are urgently needed. When it comes to analyzing medical phenomena, machine learning (ML) algorithms shine because of their ability to capture complicated and nonlinear correlations in clinical data. Machine learning (ML) methods, including as classification approaches, may be used to create a model for the diagnosis of HCV by determining which individuals are infected. However, the performance of the classifier might be hampered by the presence of incorrect qualities in the attribute set. In order to acquire a more condensed and important representation of the given information while ignoring any other superfluous or unnecessary aspects, feature selection specifies a subset of features or variables that characterize the data. Selecting the right features to include in a classifier is a strong approach to improve its performance and cut down on the time it takes to train a model.

The goal of ensemble learning, a kind of meta-machine learning, is to improve forecast performance by combining the outputs of many models. It's a versatile algorithm that makes predictions using supervised learning after being trained with a large number of separate models or inexperienced students. The most well-known ensemble techniques are bagging and boosting, although there are many more to choose from. Bagging techniques include random forest and additional tree algorithm, while boosting methods include gradient boosting, adaboost, and extreme gradient boosting. Although certain regular decision tree algorithms may generate better ensembles than purposeful ones, in principle ensembles are supposed to deliver more effective outcomes if the models are modified significantly.

## 3. Machine Learning Models

This section presented the two models of machine learning. Figure 1 shows the methodology of this work.

### 3.1 Support Vector Machine (SVM)

Vapnik's support vector machine (SVM) technique is a popular supervised AI model for a variety of tasks including data analysis, pattern recognition, statistical regression, and prediction. Estimation of the regression in the SVM model is based on a set of kernel functions that may implicitly transform the original, lower-dimensional input information



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

into a higher-dimensional set of features. In contrast to traditional ANN models, which often exhibit several regional minima, the SVM provides a single solution due to the convexity of the optimality issue[17]–[19].

We can compute the function of approximation by:

$$f(y) = wh(y) + t \quad (1)$$

Where $h(y)$ refers to the high-D feature, t and w refer to weights of threshold and vector.

The minimization hazard function can be computed as:

$$M(F) = F\frac{1}{n}\sum_{i=1}^{n} L(E_i, x_i) + \frac{1}{2}\|w\|^2 \quad (2)$$

Where f is a penalty variable, $E_i$ refers to desired variable.

The approximation function can be computed by including constraints of optimality and Lagrange multipliers as:

$$F(y, r_i, r_i^*) = \sum_{i=1}^{n}(r_i - r_i^*)K(y, y_i) + t \quad (3)$$

$K(y, y_i)$ refers to function of kernel

The RBF can be computed as:

$$K_{rbf}(y, y_i) = \exp\left[\frac{-(y-y_i)^2}{2\sigma^2}\right] \quad (4)$$

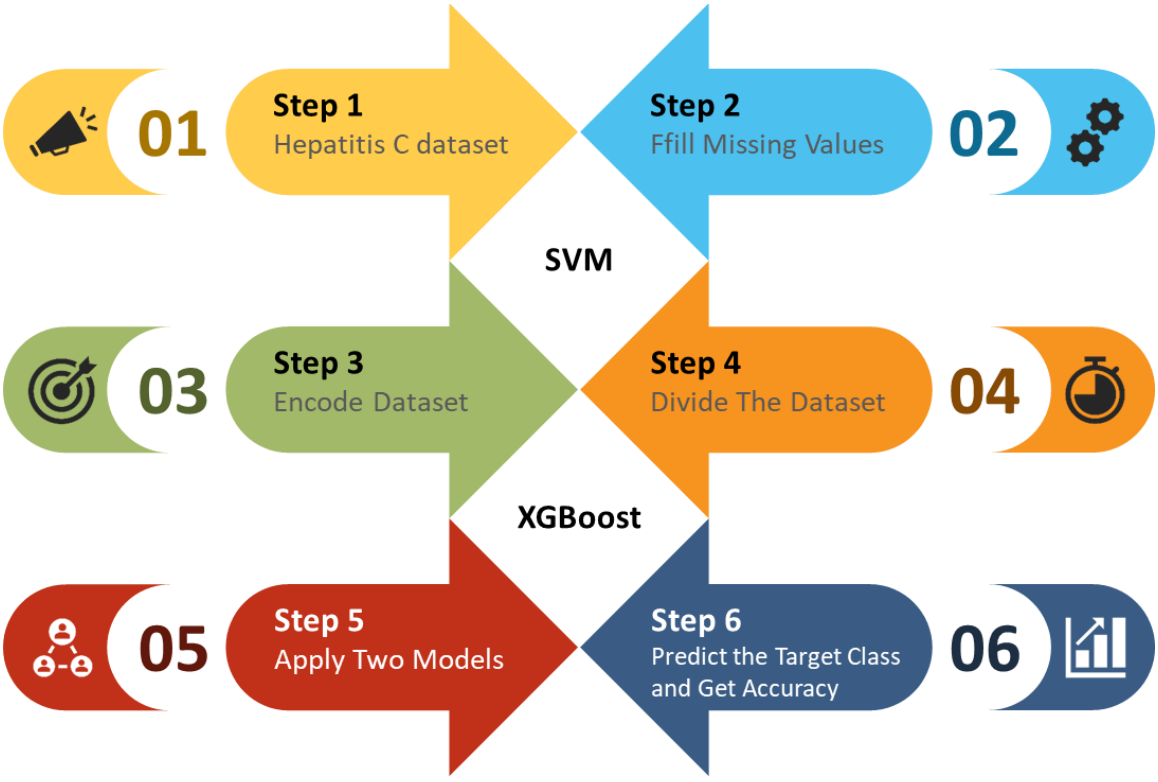


Figure 1: The steps of the two models.

**3.2 XGBoost**

When it comes to implementing Gradient Boosting Machines, especially for K The categorization and Regression Trees, the Extreme Gradient Boosting (XGBoost) method introduced by Chen and Guestrin offers an innovative approach. In order to train a "powerful" learner, the method takes inspiration from the concept of "boosting," which involves combining the predictions of several "weak" learners. Over-fitting is avoided and computational resources



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

are maximized thanks to XGBoost. This is achieved by minimizing the complexity of the goal functions that combine predictive and regularization terms while preserving the best possible computing efficiency. Additionally, XGBoost's functions automatically run parallel computations while training[20]–[22].

What follows is a description of how XGBoost uses additive training. To combat the shortcomings of a poor learner, a second model is assigned to the residuals obtained after the initial learner is applied to the entire set of data entered. This procedure for achieving a good fit is performed many times until the termination criteria are attained. When all the predictions from the many learners are added together, we get the model's final forecast.

The function of prediction stage can be computed as:

$$P_i^{(s)} = \sum_{k=1}^{s} P_k(y_i) = P_i^{(s-1)} + P_s(y_i) \quad (5)$$

$P_s(y_i)$ refers to the step t in learner.

The original function can be modified to prevent overfitting

$$O^{(s)} = \sum_{k=1}^{n} l(y - y_i) + \sum_{k=1}^{s} \theta(f_i) \quad (6)$$

$$\theta(f_i) = \varphi T + 0.5\lambda \|w\|^2 \quad (7)$$

**4. Results**

This section presented results in applying two methods in the HC dataset. Table 1 shows the sample of the dataset. The dataset contains 13 variables and 615 cases. Figure 2 shows the boxplot of the dataset.

Table 1: Sample of HC dataset

| | Category | Age | Sex | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **HCV1** | 0=Blood Donor | 32 | m | 38.5 | 52.5 | 7.7 | 22.1 | 7.5 | 6.93 | 3.23 | 106 | 12.1 | 69 |
| **HCV2** | 0=Blood Donor | 32 | m | 38.5 | 70.3 | 18 | 24.7 | 3.9 | 11.17 | 4.8 | 74 | 15.6 | 76.5 |
| **HCV3** | 0=Blood Donor | 32 | m | 46.9 | 74.7 | 36.2 | 52.6 | 6.1 | 8.84 | 5.2 | 86 | 33.2 | 79.3 |
| **HCV4** | 0=Blood Donor | 32 | m | 43.2 | 52 | 30.6 | 22.6 | 18.9 | 7.33 | 4.74 | 80 | 33.8 | 75.7 |
| **HCV5** | 0=Blood Donor | 32 | m | 39.2 | 74.1 | 32.6 | 24.8 | 9.6 | 9.15 | 4.32 | 76 | 29.9 | 68.7 |
| **HCV6** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **HCV7** | 3=Cirrhosis | 62 | f | 32 | 416.6 | 5.9 | 110.3 | 50 | 5.57 | 6.3 | 55.7 | 650.9 | 68.5 |
| **HCV8** | 3=Cirrhosis | 64 | f | 24 | 102.8 | 2.9 | 44.4 | 20 | 1.54 | 3.02 | 63 | 35.9 | 71.3 |
| **HCV9** | 3=Cirrhosis | 64 | f | 29 | 87.3 | 3.5 | 99 | 48 | 1.66 | 3.63 | 66.7 | 64.2 | 82 |
| **HCV10** | 3=Cirrhosis | 46 | f | 33 | NaN | 39 | 62 | 20 | 3.56 | 4.2 | 52 | 50 | 71 |
| **HCV11** | 3=Cirrhosis | 59 | f | 36 | NaN | 100 | 80 | 12 | 9.07 | 5.3 | 67 | 34 | 68 |



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

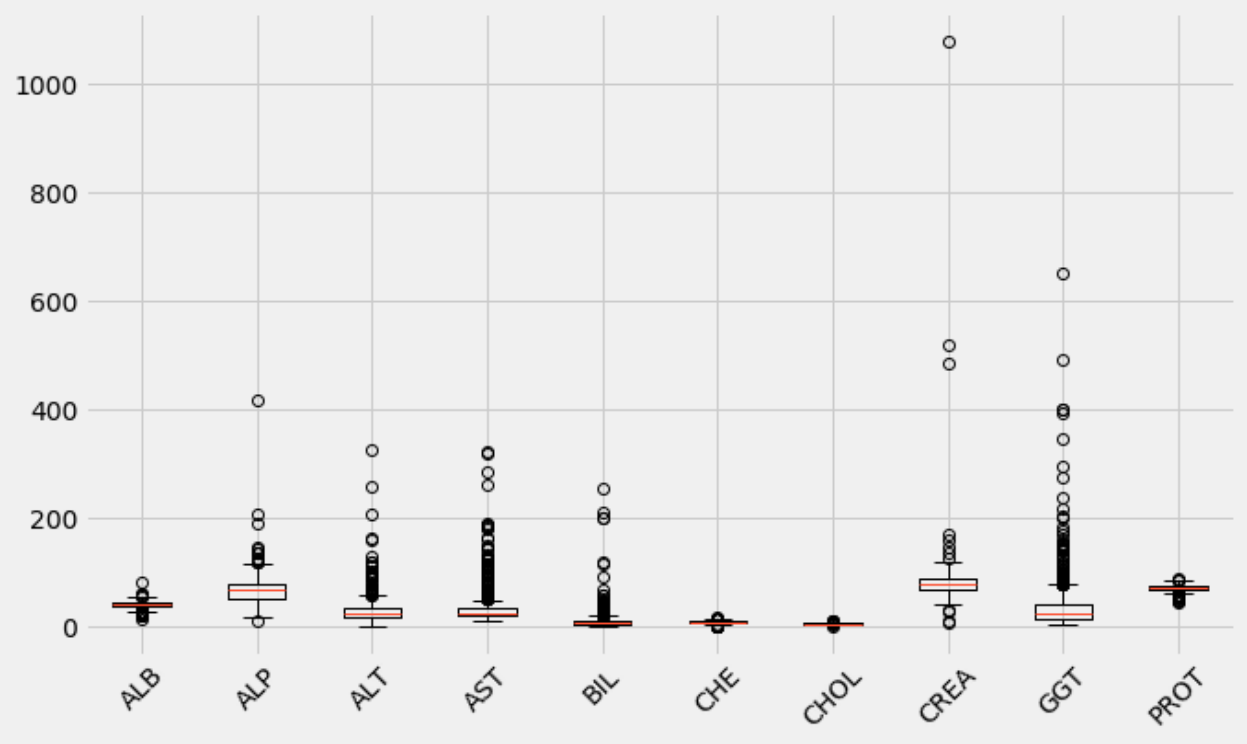


Figure 2: The boxplot of the dataset.

From Table 1 there are missing values in the dataset. So this paper filled the missing valued by the mean value. Table 2 shows the sample of dataset after filling missing values. Then the category column is encoded into numeric values.

Table 2: Sample of HC dataset after filling missing values.

| | Category | Age | Sex | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **HCV1** | 0 | 32 | 0 | 38.5 | 52.5 | 7.7 | 22.1 | 7.5 | 6.93 | 3.23 | 106 | 12.1 | 69 |
| **HCV2** | 0 | 32 | 0 | 38.5 | 70.3 | 18 | 24.7 | 3.9 | 11.17 | 4.8 | 74 | 15.6 | 76.5 |
| **HCV3** | 0 | 32 | 0 | 46.9 | 74.7 | 36.2 | 52.6 | 6.1 | 8.84 | 5.2 | 86 | 33.2 | 79.3 |
| **HCV4** | 0 | 32 | 0 | 43.2 | 52 | 30.6 | 22.6 | 18.9 | 7.33 | 4.74 | 80 | 33.8 | 75.7 |
| **HCV5** | 0 | 32 | 0 | 39.2 | 74.1 | 32.6 | 24.8 | 9.6 | 9.15 | 4.32 | 76 | 29.9 | 68.7 |
| **HCV6** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **HCV7** | 1 | 62 | 1 | 32 | 416.6 | 5.9 | 110.3 | 50 | 5.57 | 6.3 | 55.7 | 650.9 | 68.5 |
| **HCV8** | 1 | 64 | 1 | 24 | 102.8 | 2.9 | 44.4 | 20 | 1.54 | 3.02 | 63 | 35.9 | 71.3 |
| **HCV9** | 1 | 64 | 1 | 29 | 87.3 | 3.5 | 99 | 48 | 1.66 | 3.63 | 66.7 | 64.2 | 82 |
| **HCV10** | 1 | 46 | 1 | 33 | 68.28392 | 39 | 62 | 20 | 3.56 | 4.2 | 52 | 50 | 71 |
| **HCV11** | 1 | 59 | 1 | 36 | 68.28392 | 100 | 80 | 12 | 9.07 | 5.3 | 67 | 34 | 68 |



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

Figures 3 and 4 show the amount of patient and healthy persons, and size of male and female. From Figure 3 the amount of healthy persons greater than the amount of suspected patients. From Figure 4. The amount of male greater than the amount of female.

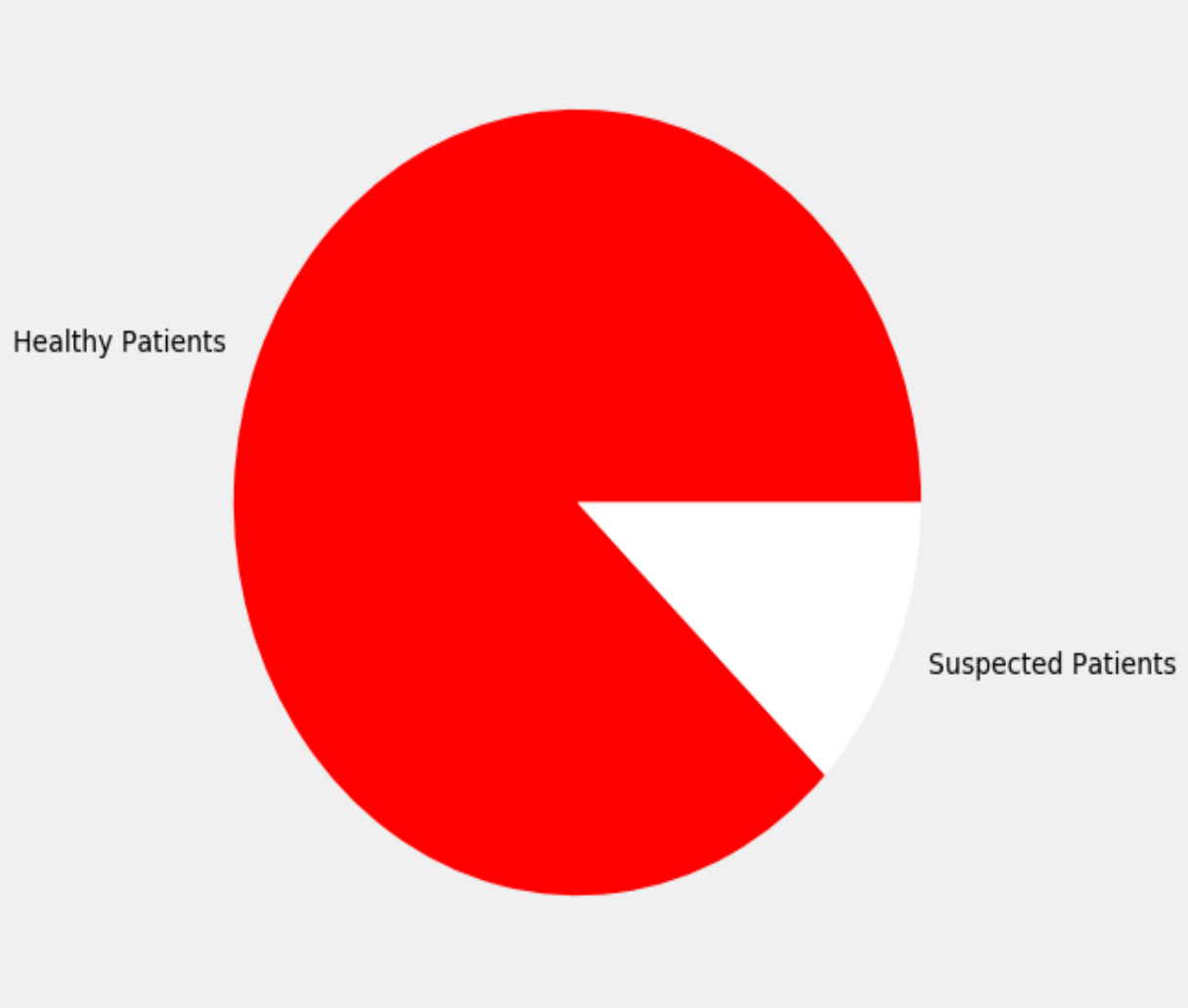


Figure 3: The size of health persons and non-healthy persons.



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

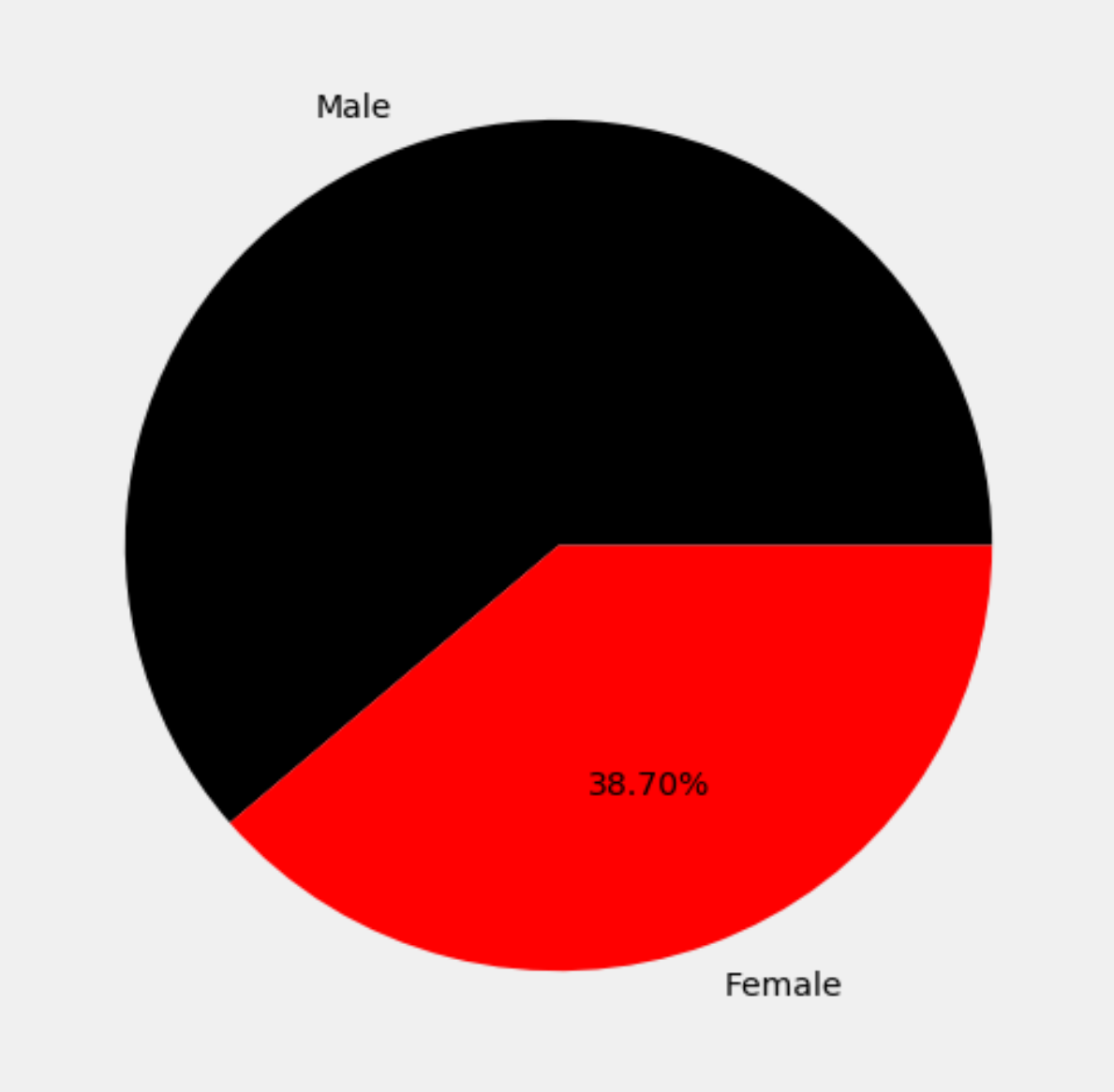


Figure 4: The number of male and female.

Figure 5 shows the heatmap between 13 features in the dataset. There is strong correlation between category variable and AST variable.



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

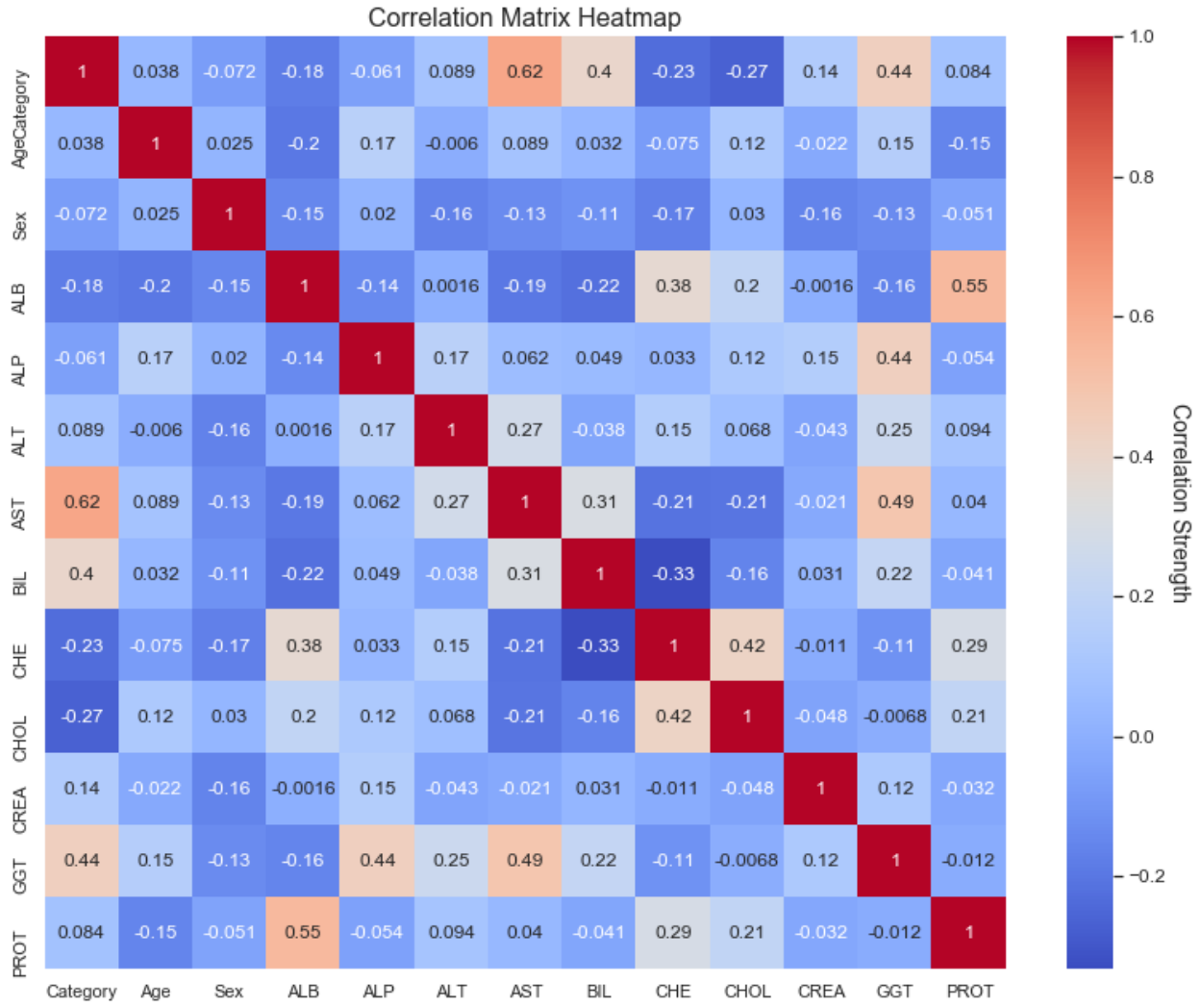


Figure 5: The heatmap of the collected dataset.

Then the dataset is divided into two groups train and test. The amount of train is 80% and the amount of test is 20%. Then fit the SVM and XGBoost models into the train and test dataset.

The confusion matrix of two models are shown in Figure 6. Figure 7 shows the accuracy score of two models. The XGBoost has greater accuracy then the SVM model.



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

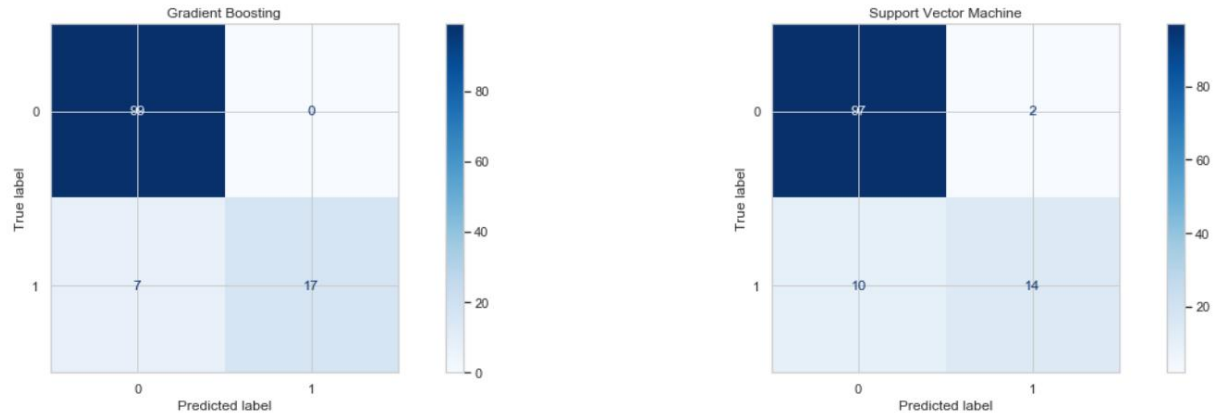


Figure 6: The confusion matrix.

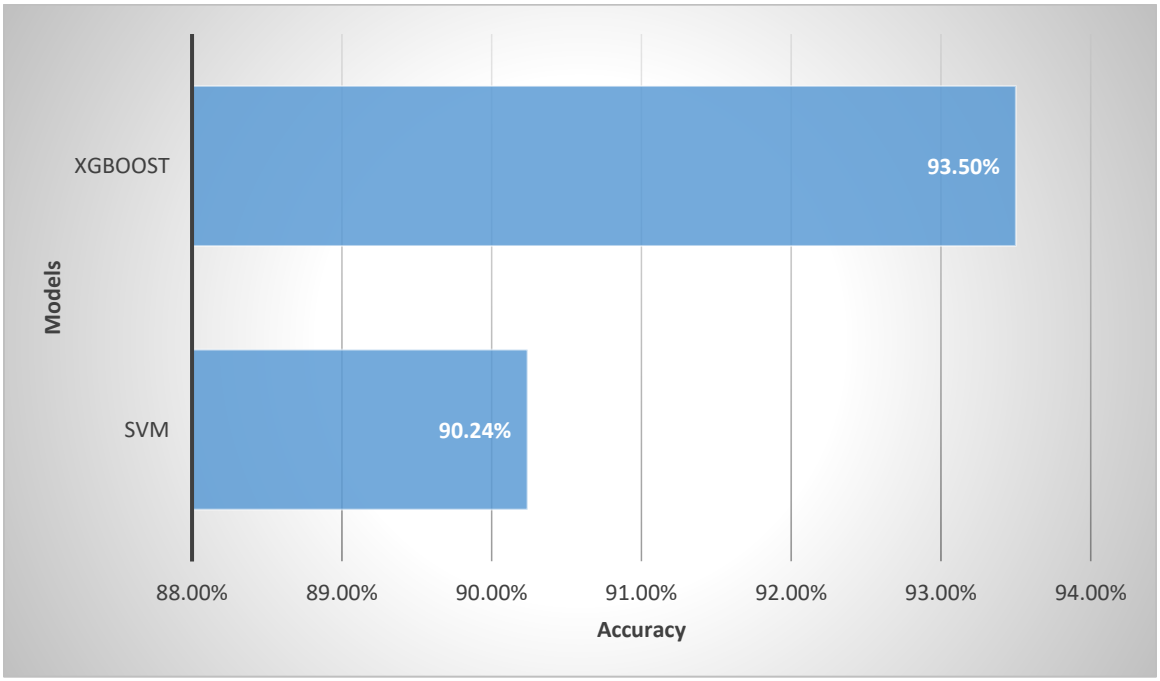


Figure 7: The comparison between SVM and XGBoost.

## 5. Conclusion

Before the development of ML in the healthcare sector, diagnosing a patient's Hepatitis C Level required a battery of diagnostics. In order to pinpoint the exact phase of the illness, an individual must endure a battery of unneeded diagnostic procedures. The efforts and costs involved are too much for the patient. To conserve time and effort for everyone involved, including patients and clinicians, a diagnostic procedure that can determine how far along the disease's progression a given patient is would be very helpful. Forecasting of illnesses and categorization according to health information are two applications of ML methods and their numerous models. This paper applied two ML models SVM and XGBoost on the HCV dataset. This study perfumed some preprocessing steps such as filling in missing values with the mean data and encoding the category feature. The results showed the XGBoost model has more accuracy than the SVM model. The XGBoost gets 93.5% accuracy and SVM achieves 90.23% accuracy.



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

**References**


[1] S. M. Abd El-Salam *et al.*, "Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients," *Informatics Med. Unlocked*, vol. 17, p. 100267, 2019.

[2] A. Akella and S. Akella, "Applying machine learning to evaluate for fibrosis in chronic hepatitis c," *MedRxiv*, pp. 2011–2020, 2020.

[3] D. Chicco and G. Jurman, "An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis," *IEEE Access*, vol. 9, pp. 24485–24498, 2021.

[4] S. C. R. Nandipati, C. XinYing, and K. K. Wah, "Hepatitis C virus (HCV) prediction by machine learning techniques," *Appl. Model. Simul.*, vol. 4, pp. 89–100, 2020.

[5] K. S. Bhargav, D. Thota, T. D. Kumari, and B. Vikas, "Application of machine learning classification algorithms on hepatitis dataset," *Int. J. Appl. Eng. Res.*, vol. 13, no. 16, pp. 12732–12737, 2018.

[6] S. Hashem *et al.*, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 3, pp. 861–868, 2017.

[7] A. A. Malik, W. Chotpatiwetchkul, C. Phanus-Umporn, C. Nantasenamat, P. Charoenkwan, and W. Shoombuatong, "StackHCV: A web-based integrative machine-learning framework for large-scale identification of hepatitis C virus NS5B inhibitors," *J. Comput. Aided. Mol. Des.*, vol. 35, pp. 1037–1053, 2021.

[8] I. E. Weidlich *et al.*, "Inhibitors for the hepatitis C virus RNA polymerase explored by SAR with advanced machine learning methods," *Bioorg. Med. Chem.*, vol. 21, no. 11, pp. 3127–3137, 2013.

[9] T. M. Ghazal, "Hep-pred: hepatitis c staging prediction using fine gaussian svm," *Comput. Mater. Contin.*, vol. 69, no. 1, pp. 191–203, 2021.

[10] M. A. Konerman, Y. Zhang, J. Zhu, P. D. R. Higgins, A. S. F. Lok, and A. K. Waljee, "Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data," *Hepatology*, vol. 61, no. 6, pp. 1832–1841, 2015.

[11] T. C. Feldman, J. L. Dienstag, K. D. Mandl, and Y.-J. Tseng, "Machine-learning-based predictions of direct-acting antiviral therapy duration for patients with hepatitis C," *Int. J. Med. Inform.*, vol. 154, p. 104562, 2021.

[12] A. Orooji and F. Kermani, "Machine learning based methods for handling imbalanced data in hepatitis diagnosis," *Front. Heal. Informatics*, vol. 10, no. 1, p. 57, 2021.

[13] M. ElHefnawi *et al.*, "Accurate prediction of response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, IEEE, 2012, pp. 771–778.

[14] M. B. Butt *et al.*, "Diagnosing the stage of hepatitis C using machine learning," *J. Healthc. Eng.*, vol. 2021, 2021.

[15] K. Ahammed, M. S. Satu, M. I. Khan, and M. Whaiduzzaman, "Predicting infectious state of hepatitis C virus affected patient's applying machine learning methods," in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 1371–1374.

[16] A. A. Kashif, B. Bakhtawar, A. Akhtar, S. Akhtar, N. Aziz, and M. S. Javeid, "Treatment response prediction in hepatitis C patients using machine learning techniques," *Int. J. Technol. Innov. Manag.*, vol. 1, no. 2, pp. 79–89, 2021.

[17] Y. Li *et al.*, "A county-level soybean yield prediction framework coupled with XGBoost and




Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023

multidimensional feature engineering," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 118, p. 103269, 2023.

[18] Z. Li, T. Lu, X. He, J.-P. Montillet, and R. Tao, "An improved cyclic multi model-eXtreme gradient boosting (CMM-XGBoost) forecasting algorithm on the GNSS vertical time series," *Adv. Sp. Res.*, vol. 71, no. 1, pp. 912–935, 2023.

[19] B. Pan, "Application of XGBoost algorithm in hourly PM2. 5 concentration prediction," in *IOP conference series: earth and environmental science*, IOP publishing, 2018, p. 12127.

[20] L. Guo, Z. Li, Q. Tian, L. Guo, and Q. Wang, "Prediction of CSG Splitting Tensile Strength Based on XGBoost-RF Model," *Mater. Today Commun.*, p. 105350, 2023.

[21] Rose Aljanada,Ghadeer W. Abukhalil,Aseel M. Alfaisal,Raghad M. Alfaisal, Adoption of Google Glass technology: PLS-SEM and machine learning analysis, International Journal of Advances in Applied Computational Intelligence, Vol. 1 , No. 1 , (2022) : 08-22 (Doi : https://doi.org/10.54216/IJAACI.010101)

[22] H. Ma, P. Yang, F. Wang, X. Wang, D. Yang, and B. Feng, "Short-Term Heavy Overload Forecasting of Public Transformers Based on Combined LSTM-XGBoost Model," *Energies*, vol. 16, no. 3, p. 1507, 2023.



Doi: https://doi.org/10.54216/IJAACI.030202
Received: August 15, 2022 Revised: December 09, 2022 Accepted: February 09, 2023