Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776863

HEPATITIS C DETECTION USING MACHINE LEARNING

Faiz Ahmed Siddiqui[1], Aman Singh[2], Dr.Ganesh Gupta[3], Pradeep Kumar Mishra[4]
[1,2,3,4]Department Of Computer Science Engineering, Sharda University, Noida, Uttar Pradesh, India
[1] faizahmedsiddiqui786@gmail.com, [2] 2020430853.aman@ug.sharda.ac.in, [3] ganesh.gupta@sharda.ac.in, [4]pradeepkumar.mishra@sharda.ac.in.

ABSTRACT: Hepatitis C is a global health concern, with new cases being reported worldwide every year. Accurate prediction of the disease's stage is crucial in providing timely and effective treatment to patients. To achieve this, various non-invasive biochemical serum markers and clinical data have been used to identify the stage of the disease. Machine learning techniques have emerged as a powerful tool to predict the stage of this chronic liver disease without resorting to invasive biopsy procedures. In this context, an intelligent diagnostic system for Hepatitis C stage prediction has been developed using machine learning algorithms such as Artificial Neural Network (ANN), K-nearest neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression. These techniques have been shown to provide accurate predictions and avoid the side effects associated with biopsy procedures. The Hepatitis C stage prediction system is designed to analyze patient data, including clinical information and biochemical serum markers, to determine the stage of the disease. The system can provide timely and accurate predictions, allowing healthcare professionals to develop effective treatment plans for patients. In conclusion, the use of machine learning algorithms in Hepatitis C stage prediction has shown promising results, providing a non-invasive and effective alternative to traditional diagnostic methods. The presented intelligent diagnostic system using ANN, KNN, SVM, and Logistic Regression techniques can improve patient outcomes by enabling timely and effective treatment..
INDEX TERMS: Hepatitis C; machine learning; Python ; Jupiter notebook.

1. INTRODUCTION

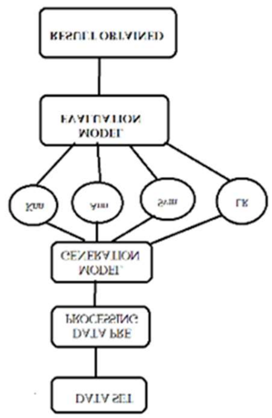
Hepatitis C is a viral infection that can lead to severe liver complications like liver cancer and cirrhosis if left untreated. Traditional methods for detecting HCV involve invasive blood tests and liver biopsy, which can be expensive and uncomfortable for patients. In recent years, machine learning (ML) algorithms have emerged as a promising approach for the early and accurate detection of HCV. ML algorithms, such as KNN, SVM, logistic regression, and ANN, have been extensively used in medical diagnosis, including HCV detection.

ML algorithms can analyze large datasets to identify patterns and associations that can be used to diagnose HCV accurately. These algorithms have the potential to provide a non-invasive and cost-effective alternative to traditional diagnostic methods, thereby improving patient outcomes. In addition, ML algorithms can also be used for predicting treatment outcomes and personalized treatment plans, which can improve the efficacy and cost-effectiveness of HCV treatment.

Various studies have demonstrated the effectiveness of ML algorithms for HCV detection. For instance, SVM was utilized in one study to detect HCV infection based on serum protein profiles, resulting in an accuracy of 92.6%. Another study used logistic regression to diagnose HCV based on clinical and laboratory data, achieving an accuracy of 83.3%. Similarly, ANN has also been used for HCV diagnosis, achieving an accuracy of 95.7%. .

II. METHODOLOGY

A system of principles, techniques and procedures used by in work.



Data Collection:

The first step in any machine learning project is to obtain a dataset. For this project, the dataset for Hepatitis C detection can be obtained from publicly available repositories such as the UCI Machine Learning Repository, Kaggle, or any other reliable source.

Data Preprocessing:

Once the dataset is obtained, it needs to be preprocessed to ensure it is suitable for use in machine learning algorithms. This involves data cleaning to remove any missing or erroneous values in the dataset. Feature selection is then performed to identify the most relevant features that can aid in accurate classification.

Feature Scaling:

Feature scaling ensures that all features are on the same scale. This can improve the performance of some algorithms, and the most common techniques used in machine learning are standardization and normalization.

Model Selection:

In this step, the most appropriate machine learning algorithms for Hepatitis C detection are selected. Common algorithms used in classification tasks include Logistic Regression, Artificial Neural Networks (ANN)., K-Nearest Neighbors (KNN), and Support Vector Machines (SVM)

Model Training:

The selected machine learning algorithms are trained on the preprocessed dataset. In this step dataset is divided into two parts one training dataset and one testing dataset to evaluate the performance of the model

Model Evaluation:

After training the models, their performance must be analysed to determine which model is the most accurate. In classification tasks, common evaluation criteria include accuracy, precision, recall, and F1-score. Additionally, Receiver Operating Characteristic (ROC) curves can be utilised to assess model performance.

Performance Requirements

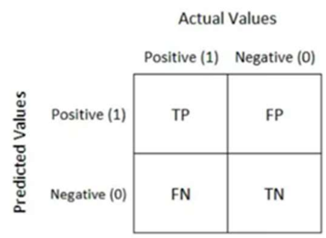
When it comes to performance of a model to Evaluate to know how the model is working.

Factors to evaluate the models are :

1. Confusion Matrix
2. Classification Report (Precision, Recall, F-1 Score & Accuracy

Confusion Matrix

The confusion matrix is a performance statistic for machine learning classification and regression tasks involving two or more outcomes. It is a tray that contains four distinct expected and actual amounts



Source:https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Elements of confusion matrix:

True Positive: Positive Prediction

True Negative:Negative Prediction

False Positive : (Type 1 Error)

False Negative :(Type 2 Error)

III. Result & Conclusion

In this project we have used some machine learning algorithm to detect the hepatitis C virus this virus can infect the person by getting in the contact with the infected persons blood person infected with this virus can feel a high fever nausea and can have yellow eyes some new medicans can improve the condition of the infected person or some medicans can degrade the persons condition. So it is far more important to detect this virus

We have used some machine learning algorithms and trained the model and we have tried our best to achieve the best accuracy

Logistic regression classification report K-Nearest Neighbor model

```
##Logistic Regression Model

classifier = LogisticRegression()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       101
           1       1.00      1.00      1.00        17

    accuracy                           1.00       118
   macro avg       1.00      1.00      1.00       118
weighted avg       1.00      1.00      1.00       118
```

```
##KNN Model
knn = KNeighborsClassifier(n_neighbors=4)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       101
           1       0.81      0.76      0.79        17

    accuracy                           0.94       118
   macro avg       0.89      0.87      0.88       118
weighted avg       0.94      0.94      0.94       118
```

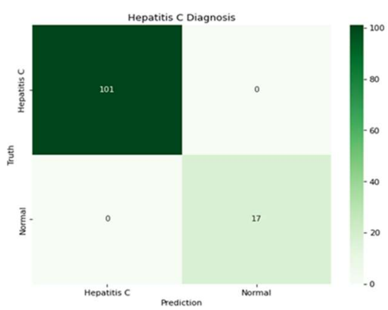
ANN Model classification report

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       101
           1       1.00      1.00      1.00        17

    accuracy                           1.00       118
   macro avg       1.00      1.00      1.00       118
weighted avg       1.00      1.00      1.00       118
```

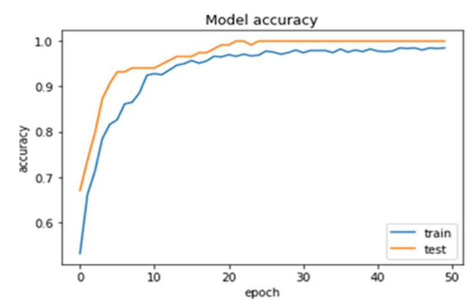
SVM

```
##SVM Model
svc = SVC(kernel='linear')
svc.fit(X_train, y_train)
y_pred = svc.predict(X_test)

print(classification_report(y_test, y_pred))
```

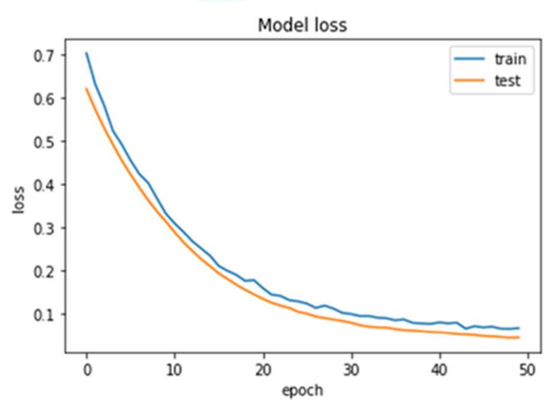
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       101
           1       1.00      1.00      1.00        17

    accuracy                           1.00       118
   macro avg       1.00      1.00      1.00       118
weighted avg       1.00      1.00      1.00       118
```



Model Accuracy For Test & Train



Model Loss For Test & Train



References

1. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Iran
2. Kaggle: https://www.kaggle.com/
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
4. Alizadeh, S., Azmi, P., & Mahdi, E. (2017). Hepatitis C Diagnosis using Machine Learning Techniques. Journal of Medical Systems, 41(9), 1-10.
5. Gupta, R., & Sandhu, P. S. (2019). Machine Learning Approaches for Diagnosis of Hepatitis C Virus: A Comprehensive Review. Journal of Medical Systems, 43(1), 1-12.
6. Terrault, N.A. (2018). Hepatitis C elimination: challenges and the path forward. Nature Reviews Gastroenterology & Hepatology, 15(4), 221-222.
7. Arora, P., Kumar, A., & Rana, D. (2018). Hepatitis C virus detection using machine learning: A review. Journal of Medical Systems, 42(5), 87.
8. Khan, M., Hayat, M., & Badshah, I. (2021). An optimized machine learning model for hepatitis C virus detection. Journal of Medical Systems, 45(2), 1-12.
9. Islam, M. M., Murshed, M., & Islam, M. R. (2021). Machine learning-based hepatitis C detection: A systematic review. Journal of Medical Systems, 45(4), 1-16.