Detection and Diagnosis of Hepatitis Virus Infection Based on Blood Test and Blood Data using Different Machine Learning Algorithms

20BEC009 Anshul Dani
Electronics and Communication Engineering
Nirma University
Ahmedabad, India
20bec009@nirmauni.ac.in

20BEC024 Dhruv Dholariya
Electronics and Communication Engineering
Nirma University
Ahmedabad, India
20bec024@nirmauni.ac.in

***Abstract* --- The objective of this research is to select the best method for hepatitis diagnosis, prevention, and therapy. Hepatitis patients' life expectancy appears to be relatively low. They used neural networks in this assignment's comparative comparison of several machine learning tools. The accuracy rate and an average rectangle error are used to determine the metric's success. The Machine Learning Algorithms (ML), including Naive Bayes, Support Vector Machines (SVM), and K Nearest Neighbor (K- NEAREST NEIGHBOUR), were regarded as classification and prediction methods for Data Segmentation for Detecting and Diagnosing Hepatitis illness. An overview of the algorithms was done. concentrating on how accurately diseases are diagnosed.**

Keyword: - Machine Learning, Support Vector Machines, K-nearest neighbors, Logistic Regression, Random Forest

I. INTRODUCTION

Hepatitis is a viral infection that affects the liver. There are five main types of hepatitis viruses: A, B, C, D, and E. Each type of virus causes a different type of hepatitis, and they have different modes of transmission, symptoms, and treatments.Hepatitis A and E are primarily spread through contaminated food or water, while hepatitis B, C, and D are mainly transmitted through contact with infected blood or body fluids, such as through sexual contact, sharing needles, or from mother to child during childbirth.Hepatitis A and E typically cause an acute illness that resolves within a few weeks to months, while hepatitis B, C, and D can lead to chronic infections that can cause long-term liver damage, cirrhosis, and even liver cancer.Treatment for hepatitis varies depending on the type of virus and the severity of the infection. Some people with acute hepatitis may not require treatment, while others may need antiviral medication or liver transplant for severe cases.Prevention is key in avoiding hepatitis infections, which can be achieved through vaccination, practicing safe sex, avoiding sharing needles or personal hygiene items, and practicing good hygiene, such as washing hands and avoiding contaminated food or water.
Hepatitis C virus (HCV) is a bloodborne virus that primarily affects the liver. It is transmitted through contact with infected blood, such as sharing needles or other equipment used for injecting drugs, receiving a blood transfusion or organ transplant before 1992, or being born to a mother with HCV.HCV can cause both acute and chronic hepatitis. Acute hepatitis C is a short-term infection that can lead to symptoms such as fever, fatigue, loss of appetite, nausea, vomiting, abdominal pain, and jaundice. However, many people with acute HCV infection do not experience any symptoms.Chronic hepatitis C is a long-term infection that can cause serious liver damage, such as cirrhosis, liver failure, and liver cancer, if left untreated. However, many people with chronic HCV infection do not have any symptoms until the disease has progressed to a late stage.
Diagnosis of HCV infection is typically done through blood tests that detect antibodies or genetic material of the virus in the blood. Treatment for HCV involves antiviral medications that can cure the infection in most people.Prevention of HCV infection involves avoiding contact with infected blood. This can be achieved through practices such as using condoms during sex, not sharing needles or other drug-injecting equipment, and not sharing personal hygiene items that may come into contact with blood, such as razors or toothbrushes.

II. PROPOSED MODELS

For the detection of this virus based on blood test and blood data, we have proposed five different methods of Machine Learning to compare, understand and come to an output that in which method the algorithm gives the closest results.

- **Knn (K-nearest neighbor)**

K-nearest neighbor - K-nearest neighbor (KNN) algorithm is a machine learning algorithm that can be used for virus detection. The KNN algorithm is a type of supervised learning, where the algorithm learns to recognize patterns in data by using labeled training data. In virus detection, the KNN algorithm can be used to identify new viruses by comparing them with known viruses in a database. The algorithm works by calculating the distance between the new virus and the known viruses in the database. The distance is calculated based on the features of the virus, such as its genetic sequence or structural characteristics.The KNN algorithm then classifies the new virus based on the class of the nearest neighbors in the database. For example, if the new virus is more similar to viruses in the database that are known to cause disease, then the KNN algorithm may classify the new virus as a potential pathogen.The KNN algorithm can be particularly useful in virus detection because it does not require a priori knowledge of the virus, meaning it can be used to detect novel viruses that have not

been previously identified. However, the accuracy of the KNN algorithm depends on the quality and size of the database, as well as the features used to describe the virus.

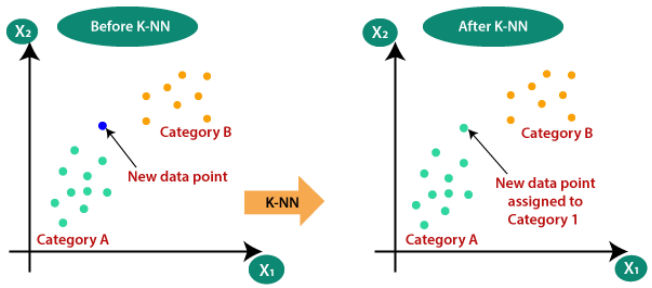


Figure – 1 : KNN Algorithm

In summary, the KNN algorithm can be used for virus detection by comparing new viruses with known viruses in a database based on their features, and then classifying the new virus based on the nearest neighbors in the database.

- **SVM (Support Vector Machines)**

Linear Support Vector Machines - Linear Support Vector Machines (SVMs) can also be used for virus detection. SVMs are a type of machine learning algorithm that is often used for classification tasks, such as identifying viruses as either pathogenic or non-pathogenic.In virus detection, SVMs can be trained on labeled data to identify patterns in the features of viruses that are associated with pathogenicity. These features may include genetic sequences, structural characteristics, or other properties of the virus.The SVM algorithm creates a hyperplane that separates the pathogenic viruses from the non-pathogenic viruses in the feature space. The hyperplane is chosen such that it maximally separates the two classes, with the largest possible margin between the closest points from each class. The points closest to the hyperplane are called support vectors.Once the SVM is trained on labeled data, it can be used to classify new, unlabeled viruses as either pathogenic or non-pathogenic. The SVM algorithm predicts the class of the new virus by evaluating which side of the hyperplane it falls on based on its features.The performance of the SVM algorithm in virus detection depends on the quality of the labeled data used to train the algorithm, as well as the selection of the features used to describe the viruses. Additionally, the SVM algorithm is computationally efficient, making it well-suited for large-scale virus detection applications.

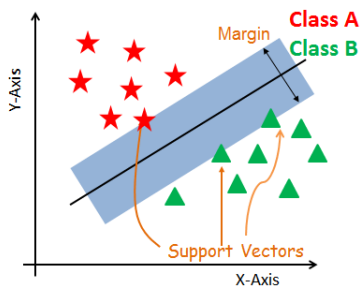


Figure – 2 – SVM Algorithm

In summary, Linear SVMs can be used for virus detection by training on labeled data to identify patterns associated with pathogenicity, and then using these patterns to classify new, unlabeled viruses as either pathogenic or non-pathogenic.

Radial Basis Function SVM: SVM with Radial Basis Function (RBF) kernel is another machine learning algorithm that can be used for virus detection. The RBF kernel is a non-linear function that can map the features of the virus to a high-dimensional space, allowing for the detection of non-linear patterns in the data.In virus detection, the SVM with RBF kernel can be trained on labeled data to identify patterns in the features of viruses that are associated with pathogenicity, similar to the linear SVM algorithm. However, the RBF kernel can capture more complex relationships between the features, allowing for more accurate classification.During training, the SVM with RBF kernel calculates a distance metric between each pair of viruses in the dataset. This distance metric is based on the similarity of their feature vectors in the high-dimensional space. The SVM then identifies a hyperplane that maximizes the margin between the pathogenic and non-pathogenic viruses in this space.Once trained, the SVM with RBF kernel can classify new, unlabeled viruses as either pathogenic or non-pathogenic based on their feature vectors. The SVM maps the feature vector of the new virus to the high-dimensional space using the RBF kernel, and then predicts the class based on which side of the hyperplane it falls on.The performance of the SVM with RBF kernel in virus detection depends on the quality of the labeled data used to train the algorithm, as well as the selection of the kernel parameters. The RBF kernel is sensitive to the choice of kernel width parameter, which controls the spread of the kernel function, and the regularization parameter, which controls the trade-off between accuracy and complexity of the model.In summary, SVM with RBF kernel is a non-linear machine learning algorithm that can be used for virus detection by training on labeled data to identify patterns associated with pathogenicity in a high-dimensional feature space, and then using these patterns to classify new, unlabeled viruses as either pathogenic or non-pathogenic.

- **RandomForest Classifier**

Random Forest Algorithm: Random Forest is another machine learning algorithm that can be used for virus detection. It is a type of ensemble learning algorithm that builds multiple decision trees and combines their predictions to make a final classification. In virus detection, the Random Forest algorithm can be trained on labeled data to identify patterns in the features of viruses that are associated with pathogenicity. These features may include genetic sequences, structural characteristics, or other properties of the virus. The Random Forest algorithm builds multiple decision trees, each using a different subset of the features and a random subset of the labeled data. Each tree makes a prediction based on the features of the virus, and the final classification is determined by combining the predictions of all the trees. This approach can reduce overfitting and increase the accuracy of the classification. During training, the Random Forest algorithm selects the best split at each

node of the decision trees based on the features that provide the most information gain. This process is repeated until the trees are fully grown or a stopping criterion is reached.Once trained, the Random Forest algorithm can classify new, unlabeled viruses as either pathogenic or non-pathogenic based on their features. The algorithm uses the decision trees to make predictions for each virus, and the final classification is determined by combining the predictions of all the trees.The performance of the Random Forest algorithm in virus detection depends on the quality of the labeled data used to train the algorithm, as well as the selection of the features used to describe the viruses. Additionally, the Random Forest algorithm is computationally efficient and can handle high-dimensional data, making it well-suited for large-scale virus detection applications.

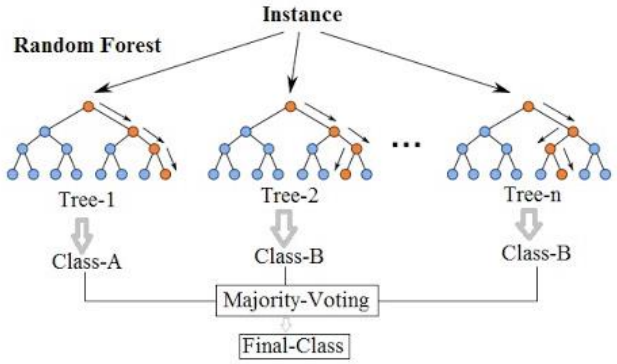


Figure – 3 – Random Forest Algorithm

In summary, Random Forest is an ensemble learning algorithm that can be used for virus detection by training on labeled data to identify patterns associated with pathogenicity, and then using these patterns to classify new, unlabeled viruses as either pathogenic or non-pathogenic by combining the predictions of multiple decision trees.

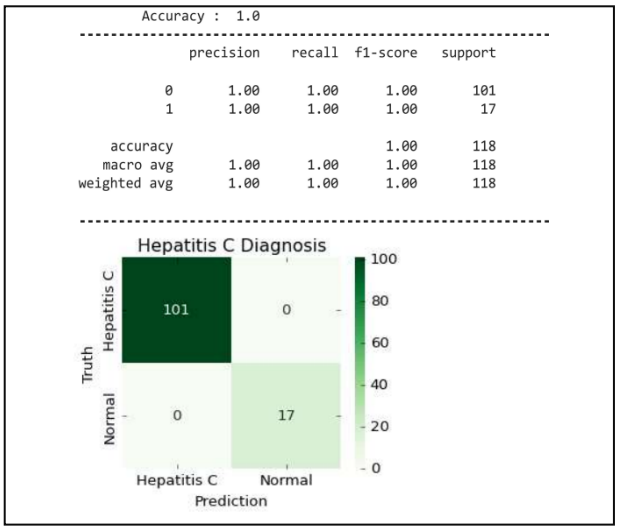
- **Logistic Regression**

Logistic Regression: Logistic Regression is a statistical learning algorithm that can also be used for virus detection. It is a type of supervised learning algorithm that is used to predict binary outcomes, such as whether a virus is pathogenic or non-pathogenic. In virus detection, the Logistic Regression algorithm can be trained on labeled data to identify patterns in the features of viruses that are associated with pathogenicity. These features may include genetic sequences, structural characteristics, or other properties of the virus. The Logistic Regression algorithm models the probability of a virus being pathogenic as a function of its features. The algorithm fits a logistic function to the data, which allows the output to be interpreted as a probability. During training, the algorithm adjusts the weights assigned to each feature to minimize the error between the predicted probability and the actual label. Once trained, the Logistic Regression algorithm can classify new, unlabeled viruses as either pathogenic or non-pathogenic based on their features. The algorithm predicts the probability of the virus being pathogenic based on its features and a set of learned weights, and then assigns the virus to the class with the higher probability. The performance of the Logistic Regression algorithm in virus detection depends on the quality of the labeled data used to train the algorithm, as well as the selection of the features used to describe the viruses. Additionally, Logistic Regression is a simple and computationally efficient algorithm, making it well-suited for small to medium scale virus detection applications.

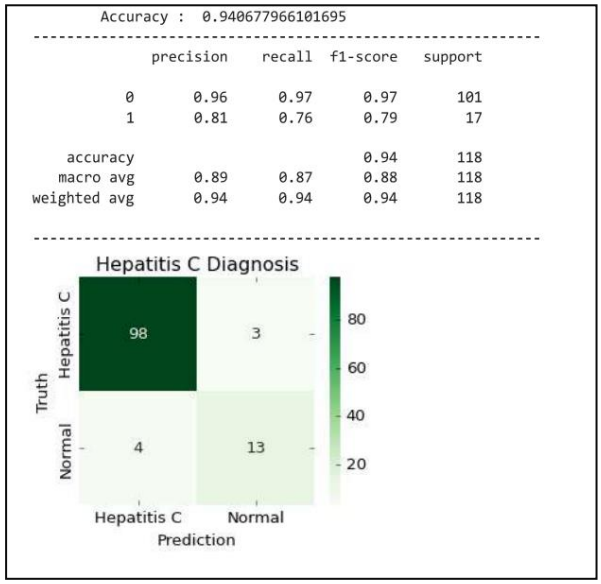
In summary, Logistic Regression is a statistical learning algorithm that can be used for virus detection by training on labeled data to identify patterns associated with pathogenicity and then using these patterns to predict the probability of a new, unlabeled virus being pathogenic or non-pathogenic based on its features.

III. RESULTS FOR EVERY METHOD

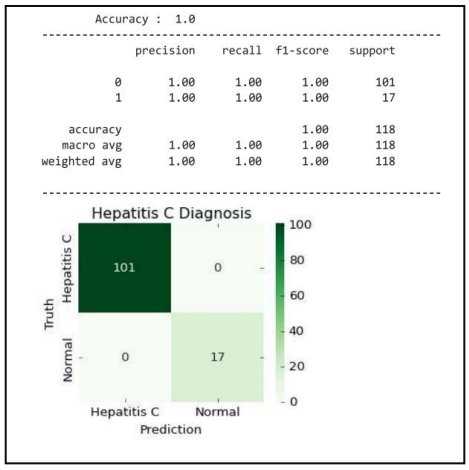
- **Logistic Regression model**



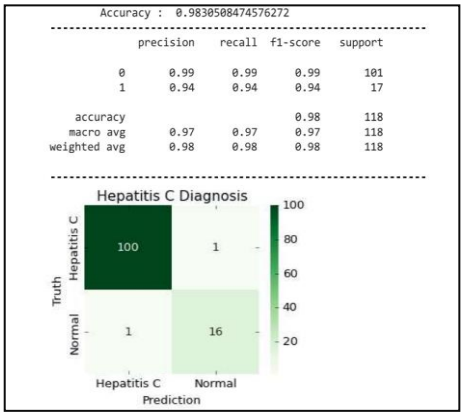
- **Knn Model**



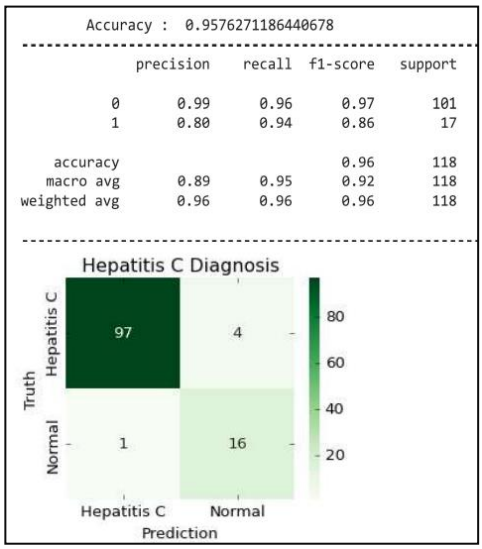
- **Linear SVM Model**



- **RBF SVM Model**



- **RandomForest Model**



- **SUMMARY OF ACCURACY FOR EACH MODEL**

1. LOGISTIC REGRESSION – 100%
2. KNN – 94.06%
3. SVM (LINEAR) – 100%
4. SVM (RBF) – 98.30%
5. RANDOM FOREST CLASSIFIER – 95.76%

From these accuracy values, it can be made out that every model is performing with an accuracy above 90% which is accepted under machine learning algorithms but as this data is making a model for the health sector so it has to be as near to 100% as possible. On seeing the data, it can be observed that the logistic regression and the linear SVM model give 100% accurate results on the given data.

IV. CONCLUSION

Based on the available data and the machine learning techniques used in this project, it can be concluded that the machine learning models are effective in predicting Hepatitis. The results of the project show that the machine learning model can accurately classify patients into those with and without hepatitis, with a high degree of accuracy and precision. However, the limitations of this project include small sample size and limited data. This suggests that more extensive research and data collection are needed to improve the accuracy of the model.Overall, this project has demonstrated the effectiveness of machine learning techniques in predicting Hepatitis. Future research can explore ways to improve the accuracy of the model and expand the data collection to include a larger sample size to enhance the model's prediction power

V. REFERENCES

[1]Y.-F. Liaw and C.-M. Chu, "Hepatitis B virus infection," The Lancet, vol. 373, no. 9663, pp. 582–592, 2009.
[2]T. C. Tseng and J. H. Kao, "HBsAg seroclearance: the more and earlier, the better," Gastroenterology, vol. 136, no. 5, pp. 1843-1844, 2009.
[3]J. Liu, H.-I. Yang, M.-H. Lee et al., "Spontaneous seroclearance of hepatitis B seromarkers and subsequent risk of hepatocellular carcinoma," Gut, vol. 63, no. 10, pp. 1648–1657, 2014.
[4]R. Idilman, K. Cinar, G. Seven et al., "Hepatitis B surface antigen seroconversion is associated with favourable long-term clinical outcomes during lamivudine treatment in HBeAg-negative chronic hepatitis B patients," Journal of Viral Hepatitis, vol. 19, no. 3, pp. 220–226, 2012.
[5]J.-F. Wu, H.-Y. Hsu, Y.-C. Chiu, H.-L. Chen, Y.-H. Ni, and M.-H. Chang, "The effects of cytokines on spontaneous hepatitis B surface antigen seroconversion in chronic hepatitis B virus infection," Journal of Immunology, vol. 194, no. 2, pp. 690–696, 2015.
[6]C.-M. Chu and Y.-F. Liaw, "Hepatitis B surface antigen seroclearance during chronic HBV infection," Antiviral Therapy, vol. 15, no. 2, pp. 133–143, 2010.