iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XI **Month of publication:** November 2021

DOI: https://doi.org/10.22214/ijraset.2021.38931

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

Anticipation of Living Status of Hepatitis B Patient by using Machine Learning

Mr. G Ragu[1], Bangaru Sri Sai Varun[2], Bondalapati Bhargav[3], Kuppala Siva Chakravarthy[4]
[1]*Asst. Professor,* [2, 3, 4]*Student, Department of Computer Science Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai-89, Tamil Nadu, India*

Abstract: ***Recently the methods of Data mining and machine learning are widely used in medical field. These methods/techniques have given better results in the prediction of respective diseases. Hepatitis B is a Liver inflammation; it can affect people of all age groups. Lakhs of people across the globe are thought to be affected by Hepatitis B. Early prediction of Hepatitis B with accurate results can save many people. Hepatitis B is a tough challenge for public health care system because of limited clinical diagnosis in the early stages of disease.***
This paper presents the decision tree algorithm to diagnose the Hepatitis B. The proposed algorithm includes collection of datasets, pre-processing, EDA (Exploratory Data Analysis), Feature Selection, data visualizing, Interpreting, saving and evaluating the model. After the data visualization process decision tree algorithm is implemented to diagnose the disease along with the patient chances of living.
Keywords: Hepatitis B virus, Machine Learning, Decision Tree, Public Health, EDA

I. INTRODUCTION

Hepatitis B is a serious liver infection caused by the hepatitis B virus (HBV). For some people, hepatitis B infection becomes chronic, means it lasts six months or above. Having chronic hepatitis B leads to liver failure, liver cancer or cirrhosis — a condition that permanently scars of the liver. More than 25 Lakh people are being infected with Hepatitis B annually with around 1 Lakh deaths.
Predicting Hepatitis B is an universal public health problem. More-over being able to help in predicting the need for a diagnosis at an initial stage for decision making about health care, predictions can also provide a brief image of an unpredictable upcoming conditions. Many methods/techniques are being implemented in the prediction of Hepatitis B to give reliable results.[1]
In this paper, decision tree in machine learning used to predict the Hepatitis B by using the dataset obtained from the UCI Machine Learning Repository. Decision Tree is used for better accuracy and good results. The structure of decision tree is used for the terms such as accuracy, sensitivity and confusion matrix. Decision Tree contains attribute nodes with two or more
subtrees and decision nodes. Decision Tree begins with two main divisions, one is training dataset where the data is stored and testing dataset where the accuracy is obtained.

A. What is Hepatitis?

Hepatitis is basically known as a inflammation of a liver.
When the liver is inflamed, then its function can be affected.
Well, In this Hepatitis there are different types are present.
They are

1) Hepatitis A
2) Hepatitis B
3) Hepatitis C
4) Hepatitis D
5) Hepatitis E

But, In this project we are going to anticipate only about Hepatitis B. Naturally, Hepatitis B is known and what we call in medicine and medical terms is a DNA virus.

B. Transmission of Hepatitis B

Hepatitis B is transmitted in a number of different ways.

Some of the transmissions are :

1) Unproctected Sex
2) Infected needles or medical devices
3) Infected blood products
4) Infected toothbrushes

C. Signs of Hepatitis B

1) Flu-like Illness
2) Fatigue
3) Abdominal Pain
4) Jaundice
5) Severe itching

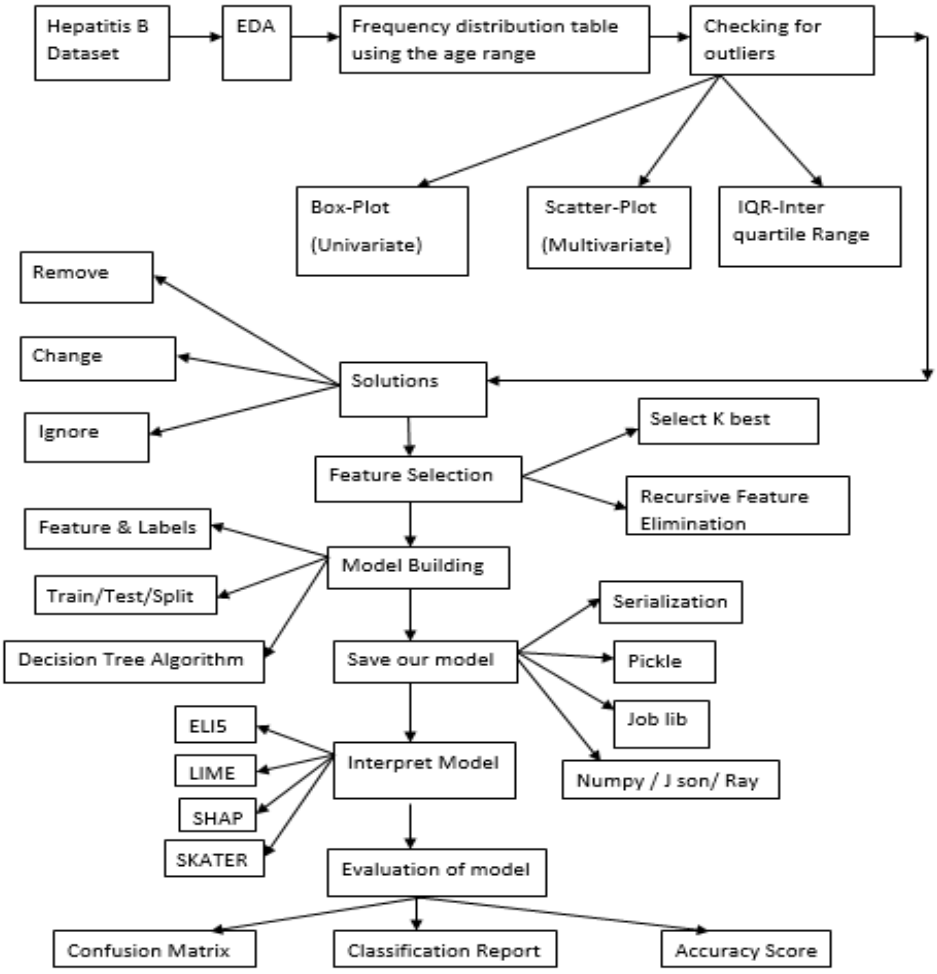
These are some of the signs of Hepatitis B which causes to the inflammation of the liver and some other body parts.

D. Complications of Hepatitis B

As we know Hepatitis B causes for the inflammation of the liver and if suppose, when a patient is unchecked or untreated it will lead to the condition called cirrhosis.

Cirrhosis problem leads to hardening or taking of the liver and the liver stops functioning as it should and it can also heart failure cancer or cancer of the liver.

II. ARCHITECTURE DIAGRAM



III. MODULE DESCRIPTION

A. Hepatitis B Dataset

Start of any project dataset is the crucial thing. In this project, We have taken dataset from UCI Machine Learning Repository. It consists of 19 various parameters to implement and to get better results. In this project we are taking multivariant dataset of Hepatitis B disease.[2]

B. EDA (Exploratory Data Analysis)

Before going to this step, make sure about standard data without any missing or null values.

EDA is a process where we can check the various aspects of our data by performing some methods. EDA is also known as "Descriptive Analysis". Some of the EDA examples are like Describe(): This function gives you all the statistical measures of our dataset counts(): This function gives you about how many values are present in a particular parameter in which you taken etc.

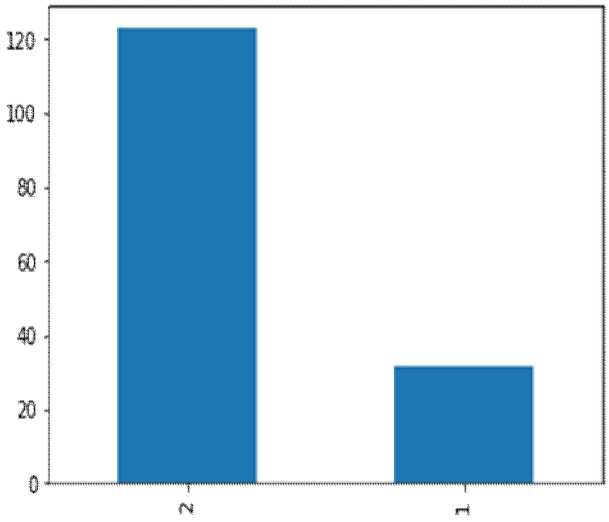


Figure. 1

1) Represents total number of patient's chance of death
2) Represents total number of patient's chances of living

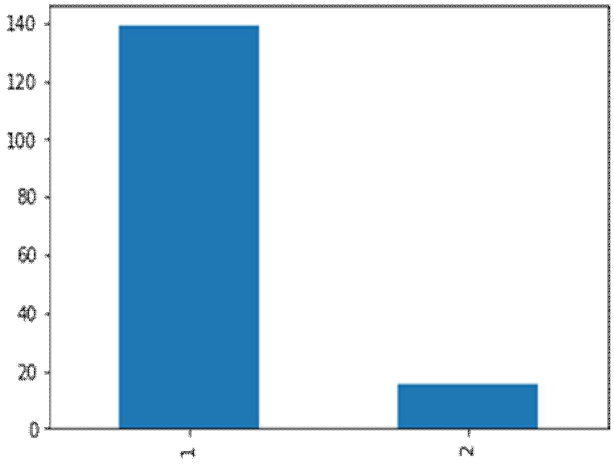


Figure. 2

1) Represents total number of males
2) Represents total number of females

C. Frequency Distribution using Age Range

In this step we are going to create an age range of a person's which are taken in the dataset. We are doing this to find that which age members are getting more infected by the hepatitis b. So, for doing this we are using some data visualization methods like bar plots, line plots, pie charts and etc. we are also using frequency distribution in this, as we know it is like representing in a graphical or tabular format is known frequency distribution.

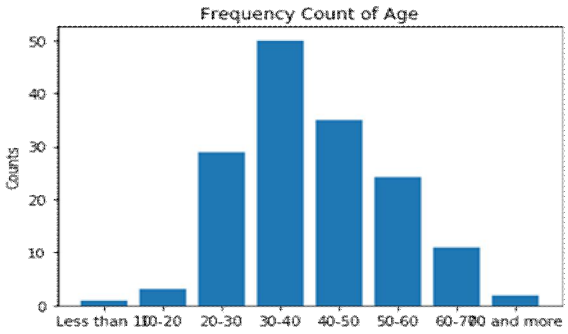


Figure. 3

1) Highest prevalence of Hepatitis is from 30-40 followed by 40-50
2) The least is individual under 10, and elderly above 70

D. Checking for Outliers

Basically, an outlier is a data point that lies so far from the other values or from other data points in a random sample of a population. So, in this we are checking outliers using by two types of analysis

1) Univariate Analysis: [It deals with the single variables]
2) Multivariate Analysis: [It deals with the multiple variables]

For univariate analysis we are using the box plot whereas for multivariate analysis we are using the scatter plot and along with these we are also using the IQR Inter quartile range methods to measure the statistical dispersion using a basic formula.

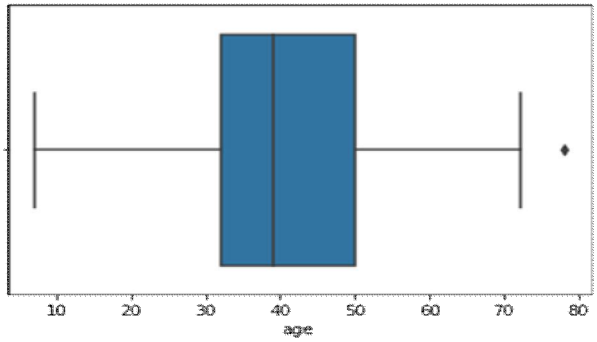


Figure. 4 In the label "Age" we can see one outlier is present in the above figure.

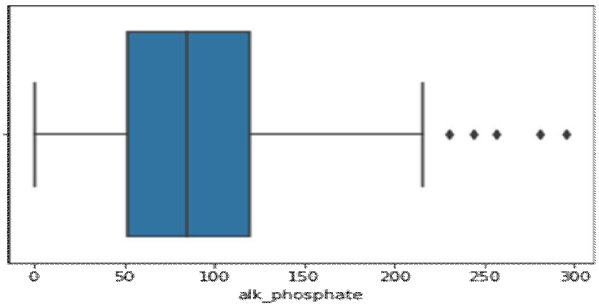


Figure. 5 In the label "Alk-phosphate" we can see that five outliers are present in the above figure Similarly, we can do for all the other labels which we have taken in our dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

E. Solutions

In the previous step we have done the process for checking whether the outliers are present or not. So, after doing some analysis methods we came to know that some of the outliers are present in our dataset. Basically, in any project if you found any outliers, the solution for this is "It's your wish to remove the outliers according to your thinking/ideation".

Basically, to remove outliers we can use three methods

1) Remove
2) Change
3) Ignore

By this, in our project we have removed the outliers using one of these solutions which are mentioned above

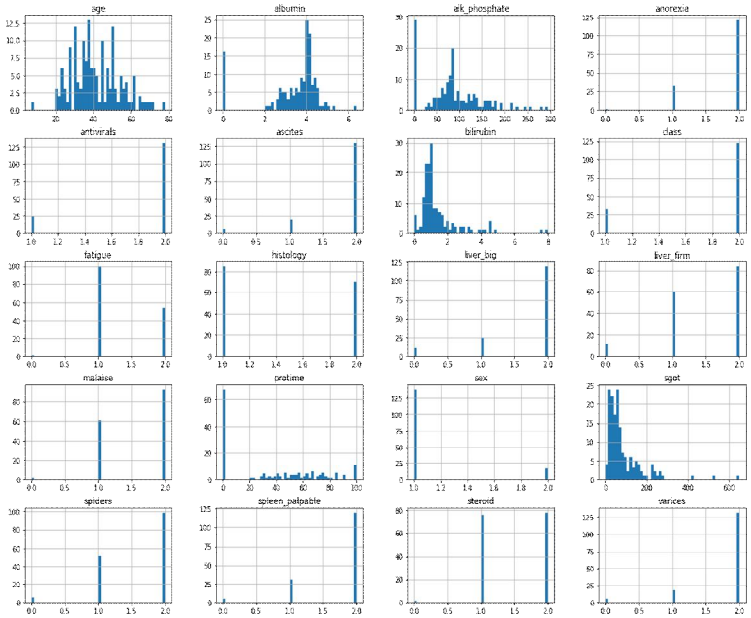


Figure. 6 Lables with outliers

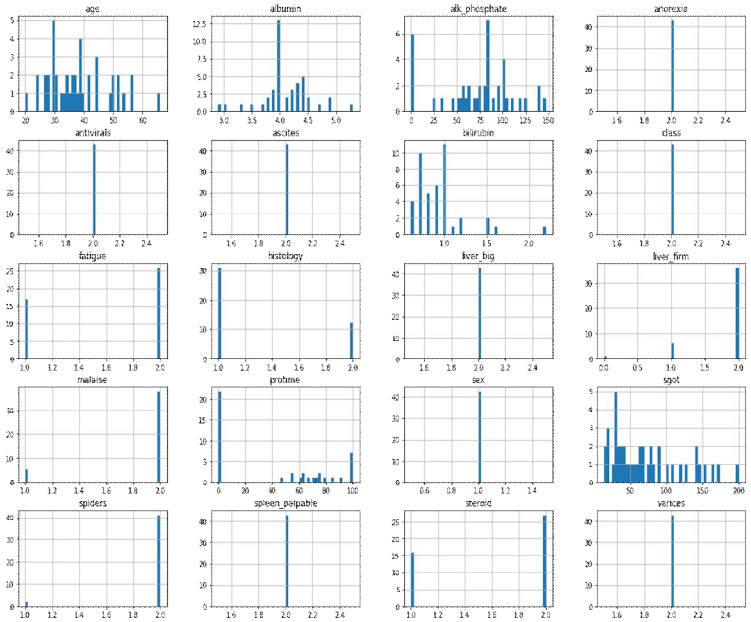


Figure. 7 Labels with no outliers

F. Feature Selection

In this step of our project, we are going to perform the process called "Feature Selection". Basically, feature selection is a process of selecting the important features that are required to get good outcome of our project.

In this process we are going to use two methods

1) *Select K-best Method:* Select K-Best method is one of the feature selection methods and it also gives you a strong relation with the output/target. In simple words we can say that what are all the important features which are required of our project, with that it give a relation. We will do this by declaring some of the sk_learn predefined libraries and also by using some features scores. By this, we can get the method with good outcomes.
2) *Recursive Feature Elimination Method:* Recursive feature elimination method is also a one of the feature selection method and it is a process removing the unworthy feature by declaring some of the sk_learn predefined libraries. So, by applying these libraries finally we will get some lowest ranking feature scores with their Boolean values ["TRUE", "FALSE]. By this we can remove or eliminate the unwanted features.[4] [5]

G. Model Building

In this step we are going to build the model of our project.

For this, we required three main processes.

They are:

1) *Features and Labels:* In feature and labels process we have done some basic methods like checking the columns, labels and after this we have declared some sk_learn predefined libraries to implement train/test/split method. After this we have started building our model with the algorithm called "Decision Tree".
2) *Decision Tree:* A decision tree is a flowchart-like tree structure, helps you in making decisions. We have chosen this algorithm because for predicting. For predicting it gives an effective method of decision making and it also permits us to examine totally the possible results of a decision.

Not only good accuracy it is also easy to explain to people about this algorithm. It is also easy to interpret and prepare. For decision tree algorithm, only a less data is required and we also know that it belongs to the supervised learning algorithms family. After doing this we got the accuracy of 74.4% in our project which says it is a good model. [3]

H. Saving the Model

Implementing the model, we are going to save the model of our project so mainly in these four methods are possible.

They are

- Joblib
- Pickle
- Serialization
- numpy/json/ray

These are the four basic methods to save the model. So in this we are using the joblib library method in our project.

Basically, Joblib is a predefined python library and the main aim of this library is that avoiding

The repetitive functions like which are repeating again and again. We can also say the name like "Recursive function" in computer language. To avoid the recursion we are using this and saving our model, which saves a lot of time and cost too.

1) *Interpret the Model:* We will do this interpretation for our model to understand the decision-making policies better and it also give humans enough trust to use the model in a real-world problem which makes a lot and huge impact on the business and etc.

Basically, there are four packages to interpret our model. They are:

a) Lime
b) Eli5
c) SHAP
d) Skate

In our project we are going to use only two packages and they are "Lime, Eli5".

- *Lime and Eli5 Packages:* These both packages help to debug machine learning classifiers and tells their predictions in simple way to understand an intuitive way. This both packages are very easy and simple machine learning frameworks to get started.

2) Evaluate the Model: This is the final step of our project. we are going to evaluate our model using the confusion matrix.

a) Confusion Matrix: This is the final step of our project. we are going to evaluate our model using the confusion matrix. confusion matrix is used for checking whether this model is best suited or not. In this confusion matrix we are going to use some parameters. The parameters are true positive, true negative, false positive, false negative. these four parameters have their different formulae's. By using those formulae's, we will get our confusion matrix.[3]

Prediction Result	Real Situation	
	Positive Class	Negative Class
Positive Class	2	2
Negative Class	32	11

Table I. Classification Results in the Confusion Matrix

Finally, the accuracy score of our project is 74.44% and we have also defined all this by taking some parameters.

IV. RESULT

	Precision	Recall	F1 Score
Die	0.50	0.15	0.24
Live	0.74	0.94	0.83
Avg	0.62	0.54	0.53

Table II. Shows Other Classification Performance Measures

V. CONCLUSION

The conclusions that can be carry from the results and discussions in this study are:

1) The use of decision tree as an anticipation method is fair enough to give the research objectives by achieving an accuracy rate of 74.4%

2) Factors that trouble patients with hepatitis in addition to the results of laboratory tests, clinical symptoms that need to be examined are diet and nutritional needs of the body and maintaining healthy liver function

We are concluding by saying that we have predicted the living status of the patient who are suffering with Hepatitis B infection by using decision tree algorithm and required parameters. In future, it is anticipated that our proposal will be implemented on the diseases other-than Hepatitis.

VI. ACKNOWLEDGEMENT

The author would like to thank the support provided by our project guide Mr. G Ragu, Asst. Professor, at SRM Institute of Science and Technology

REFERENCES

[1] B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," Proc. 2017 Int.Conf. Intell. Comput. Control Syst. ICICCS 2017, vol. 2018-

[2] E. A. Bayrak, P. Kirci, and T. Ensari, "Performance Analysis of Machine Learning Algorithms and Feature Selection Methods on Hepatitis Disease," Int. J. Multidiscip. Stud. Innov. Technol., vol.3,no.2,pp.135–138,2019.

[3] K. S. Bhargav, "Application of Machine Learning Classification Algorithms on Hepatitis Dataset," Int. J. Appl. Eng. Res., vol. 13, no.16,pp.12732–12737,2018.

[4] M.Dash and H. Liu, " Feature selection for classification", Intelligent data analysis,vol.1,no.3,pp.131-156,1997

[5] M. S. Satu, F. Tasnim, T. Akter, and S. Halder, "Exploring significant heart disease factors based on semi supervised learning algorithmns," in 2018 international conference on computer, communication,chemical,material and Electric Engineering(IC4ME2). IEEE,2018,pp.1-4.

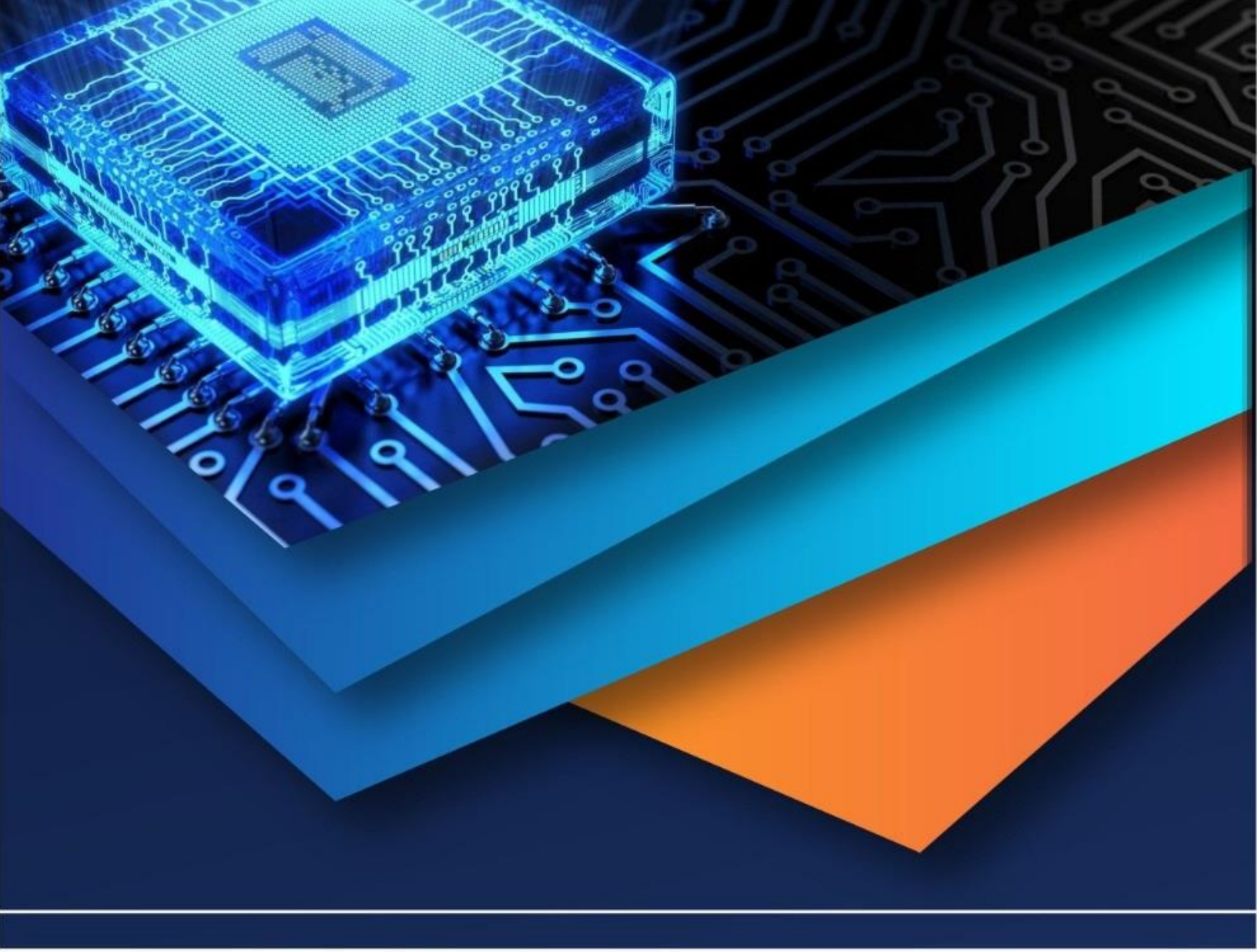
[6] Mun Kyu Lee, Jong Ho Paik, In Seop Na "Outbreak Prediction of Hepatitis A in Korea based on Statistical Analysis and LSTM Network"

[7] Khair Ahmed, Md. Shahriare, Md. Imran Khan "Predicting Infectious state of Hepatits C virus by using various machine learning algorithms.

[8] L. Parisi, N. RaviChandran, and M. L. Manaog, "A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis," Neural Comput. Appl., vol. 0123456789, 2019.

[9] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," J. Intell. Learn. Syst. Appl., vol. 09, no. 01, pp. 1–16, 2017.

[10] M. A. Konerman et al., "Machine learning models to predict disease progression among veterans with hepatitis C virus," PLoS



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY