International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

An Approach Combining Feature Selection with Machine Learning Techniques for Prediction Reliability and Accuracy in Hepatitis Diagnosis

Prasenjit Maity[1], Arup Kumar Dey[2], Krishna Prasad Singha[3], Dr. Avijit Kumar Chaudhuri[4], Mrs. Sulekha Das[5]

[1]UG-Computer Science & Engineering, Techno Engineering College Banipur
[2]UG-Information Technology, Techno Engineering College Banipur
[3]UG-Computer Science & Engineering, Techno Engineering College Banipur
[4]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur
[5]Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur
Orcid ID: 0009-0003-4369-628X[1], 0009-0007-4718-6593[2], 0009-0004-0615-9888[3], 0000-0002-5310-3180[4], 0000-0002-6641-3268[5]

Abstract
Viral hepatitis is the inflammation and damage of liver cells due to infection. There are different causes of hepatitis, but the symptoms can be similar. As of 30 June 2022, 473 cases of acute hepatitis of unknown aetiology among children aged 16 years and under have been reported from the World Health Organization European Region. Simply more than (56.7%) of these cases have been reported from the UK. The majority (76.1%) of reported cases are five years old or younger. In this review, some machine learning order procedures are applied to Hepatitis informational index procured from UCI AI Storehouse. Naïve Bayes Classifier, Logistic Regression, and J48 Decision Tree are utilized as grouping algorithms and they have been contrasted agreeing with the filter-based selection method. For channel-based feature determination, Cfs Subset Eval, Data Gain Attribute Eval, and Principal Components have been utilized and their presentation of them is assessed as far as accuracy, review, F-Measure, and ROC Region. Among the pre-owned classification algorithms, Naïve Bayes Classifier has higher classification accuracy on the data set than the others with applied and non-applied filter-based feature choices. Additionally, we proclaim that the best filter-based feature choice is Principal Components because of the highest classification precision for hepatitis patients.
Keywords: Hepatitis, Feature Selection, Genetic Algorithm (GA), Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF).

Introduction:

Hepatitis is the term that is utilized to suggest aggravation of the liver. A very rare example of disease most often causes hepatitis and it is as a rule called viral hepatitis. The most notable kinds of viral hepatitis can't avoid being Hepatitis A, Hepatitis B, and Hepatitis C. The most known 5 head hepatitis diseases are called types hepatitis A, B, C, D, and E. Considering the causing episode and happening lots of passing, these 5 sorts of hepatitis stress all people. Remarkably, types B and C motivate steady sickness in 100 million people, and they are the sweeping clarification of liver cirrhosis and the dangerous development of cells. In this paper, three of the most renowned artificial intelligence systems are applied to Hepatitis educational assortment and the outcome of applied artificial intelligence methodologies is taking a gander at using some show estimation. The rest of the paper is facilitated as follows: Section II portrays the investigations about man-made intelligence computations on prosperity instructive assortment. Fragment III gets a handle on used material and the urgent spot of used simulated intelligence procedures. The exploratory results are gotten a handle on and frame in Region IV. The paper shut in Fragment V with the end remarks.

Hepatitis A is an irritation of the liver brought about by hepatitis A infection. The infection is spread when an uninfected (and unvaccinated) individual ingests food or water that is polluted with the dung of a tainted individual. The sickness is firmly connected with perilous water or food, insufficient

DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

disinfection, unfortunate individual cleanliness, and oral-butt-centric sex. Not at all like hepatitis B and C, hepatitis A doesn't cause persistent liver infection yet it can cause incapacitating side effects and seldom fulminant hepatitis (intense liver disappointment), which is frequently deadly. WHO gauges that in 2016, 7134 people kicked the bucket from hepatitis the around world (representing 0.5% of the mortality because of viral hepatitis). Hepatitis A happens irregularly and in scourges around the world, with a propensity for cyclic repeats. Scourges connected with sullied food or water can eject violently, like the plague in Shanghai in 1988 that impacted around 300 000 individuals (1). They can likewise be delayed, influencing networks for quite a long time from one individual to the next transmission. Hepatitis An infections endure in the climate and can endure food creation processes regularly used to inactivate or control bacterial microbes.

Hepatitis B is a serious liver contamination brought about by hepatitis B infection (HBV). For the vast majority, hepatitis B is the present moment, likewise called intense, and endures under a half year. In any case, for other people, the disease becomes constant, meaning it endures over a half year. Having persistent hepatitis B expands your gamble of creating liver disappointment, liver malignant growth, or cirrhosis — a condition that for all time scars the liver. Most grown-ups with hepatitis B recuperate completely, regardless of whether their side effects are extreme. New born children and kids are bound to foster an enduring hepatitis B disease. This is known as constant contamination. An immunization can forestall hepatitis B, yet there's no fix on the off chance that you have the condition. If you're contaminated, playing it safe can assist with forestalling and spreading the infection to other people.

Hepatitis C is a fervid disease causing liver irritation and prompting serious liver effects. Hepatitis C spreads through defiled blood. Hepatitis C cure required week-by-week infusions and oral prescriptions that numerous HCV-polluted individuals couldn't take as a result of other medical issues. That is evolving. Today, constant HCV is generally recoverable with medication required consistently for two to a half year. In any case, about a portion of individuals with HCV don’t realize They donated, principally because they have no side effects, which can require a long time to show up. Thus, the U.S. Preventive Administrations Team prescribes that all grown-ups ages 18 to 70 years be evaluated for hepatitis C, even those without side effects or known liver illness.

Hepatitis D is called "delta hepatitis," liver contamination brought about by HDV. Hepatitis D just happens in individuals who are additionally tainted with HBV infection. Hepatitis D can be an enormous, epidemic or become a hard-labour, tenacious disease. Hepatitis D can cause extreme side effects and difficult ailments that can prompt long-lasting liver harm and even ailments. Individuals can belong-lasting with both HBD and HDD infections simultaneously (known as "coinfection"). There is no immunization to forestall hepatitis D. Be that as it may, the anticipation of hepatitis B with hepatitis B antibodies additionally safeguards against future hepatitis D contamination.

Hepatitis E is a liver brought about by hepatitis E infection (HEV). HEV is tracked down in a tainted individual. It is spread through ingesting the infection - even in minute sums. In agricultural nations, individuals most frequently get HED from water tainted by dung from individuals who are contaminated with the infection. In the US and other created nations where hepatitis E isn't normal, individuals have become ill with hepatitis E in the wake of eating crude or half-cooked pork, venison, wild pig meat, or shellfish. Previously, most cases in created nations affected individuals who have as of late headed out to nations where HED is normal. Side effects of hepatitis E can incorporate weakness, unfortunate cravings, stomach torment, sickness, and jaundice. Be that as it may, many individuals with hepatitis E, particularly small kids, have no side effects. Except the uncommon event of persistent hepatitis E in individuals with compromised resistant frameworks, a great many people recuperate completely from the illness with no confusion. No antibody for hepatitis E is right now accessible in the US.

In this review, some machine learning algorithms are prompted to abbreviate the demonstrative time frame and smooth out the evaluation process for doctors. Does this paper resolve the basic inquiries - (1) What are the critical highlights that make sense of the circumstances and logical results of Hepatitis? furthermore (2) Which information mining apparatus yields higher accuracy.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

The creators propose a mixture of highlight determination and stacked speculation model (HFSSGM) way to deal with consecutively decide the significant highlights, foresee utilizing different demonstrated data mining strategies, and reproduce with different train-test parts as portrayed in Fig. 1. The elements are resolved to utilize a Genetic algorithm (GA) and Logistic Regression (LR) and the dataset is changed keeping the significant highlights. Genetic algorithm and Logistic Regression are applied to the modified dataset and further refined. The emphasis goes on the precision of the forecast utilizing LR increments. The last dataset with fewer elements is then parted to make the train-test sets and likely to very much demonstrated data mining strategies (DMTs), specifically, Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF). The step is rehashed with a different train-test proportion and forecast utilizing the DMTs. This cycle goes on till the accuracy, specificity, sensitivity, AUC-ROC, and test result move along.

Moreover, the proposed classifier has been contrasted with RF, NB, and SVM with outspread premise capability piece, DT, also, MLP, which are all cutting-edge machine-learning classifiers. Different execution measurements were utilized to test the classifiers, including accuracy, sensitivity, specificity AUC-ROC trademark, the Area under the curve, also, factual tests like kappa measurements. To test the validity of the arrangement models in taking care of uneven information, this review utilized different parts of training and testing information, including 50%–50%, 66%-34%, 80%-20%, and 10-fold cross-validation. The architecture for the proposed classifier for cervical disease expectation has been portrayed in Fig. 2 below.

Literature Study:

Hepatitis implies irritation of the liver. It is a vital organ that processes important elements, channels the blood, and battles contaminations. If the liver is excited or harmed, its capability can be impacted. Alcohol, poisons, a few drugs, and certain medical condition can cause hepatitis. The disease is normal in low and centre pay nations with poor clean circumstances and hygienic practices, and most children (90%) have been contaminated with hepatitis an infection before the age of 10 years, most frequently without side effects. Contamination rates are low in big-league salary nations with great clean and hygienic circumstances. The disease might happen among teenagers and grown-ups in high-risk gatherings, for example, people who infuse drugs (PWID), men who have intercourse with men (MSM), individuals venturing out to areas of high endemicity and in disconnected populaces, like shut strict gatherings. In the US of America, enormous flare-ups have been accounted for among people encountering vagrancy. In center-pay nations and locales where sterile circumstances are variable, kids frequently get away from the disease in youth and arrive at adulthood without resistance. The most elevated weight of sickness is in the Eastern Mediterranean Area and European District, with 12 million individuals persistently contaminated in every locale. In the South-East Asia District and the Western Pacific Locale, an expected 10 million individuals in every area are constantly tainted. 9,000,000 individuals are constantly tainted in the African District and 5 million the Locale of the Americas.

In a review distributed in the Journal of f Hepatology in 2020 ¬, led by WHO with What its identity was, assessed that hepatitis D infection (HDV) influences almost 5% of individuals worldwide who have an ongoing disease with hepatitis B infection (HBV) and that HDV co-contamination could make sense of around 1 of every 5 instances of liver sickness and liver malignant growth in individuals with HBV contamination. The review has recognized a few geological focal points of HDV contamination's great commonness, including Mongolia, the Republic of Moldova, and nations in western and focal Africa.

Populations at highest risk for HAV infection include travelers from high-income developed countries who visit endemic areas of Africa, Asia, and parts of Central and South America, men who have sex with men, close contacts (household or sexual) with infected persons, persons exposed to daycare centers, as well as the homeless, the incarcerated, and illicit drug users [1][2][3]. In the 2016–17 Michigan and San Diego outbreaks in the U.S., half to three-quarters of infected individuals were

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

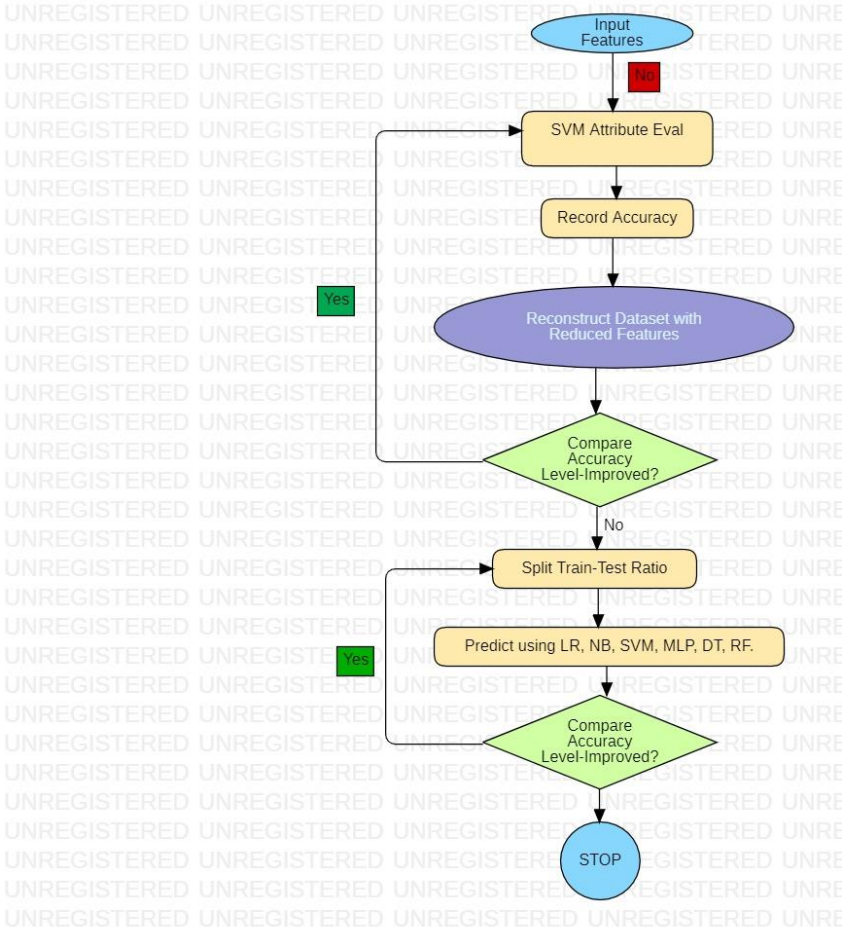
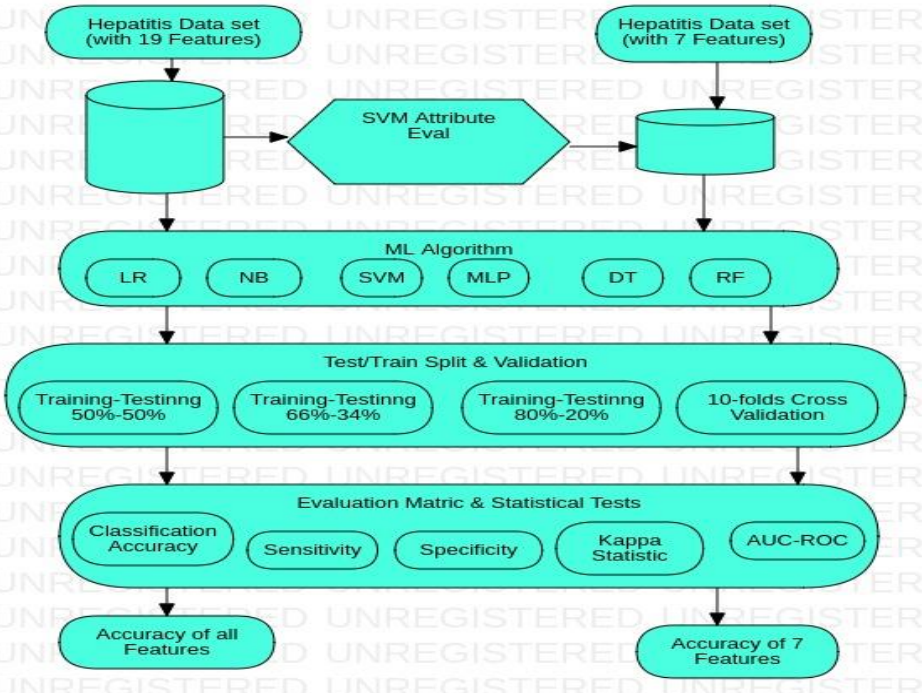


Figure 1. Hybrid Feature Selection and Stacked Generalization Model (HFSSGM) Algorithm



homeless, recently incarcerated, or illicit drug users [4][5]. In the Hawaii outbreak from scallops and the multistate outbreak from frozen strawberries, these populations were not at higher risk. Between December 2016 and June 2017, there has been an ongoing HAV outbreak in 20 European countries

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

and Tel Aviv, Israel. As of September 27, 2017, 2,873 cases of HAV infection have been identified. 980 of these cases involved male patients. Of cases among male patients, 738 (76%) occurred among men having sex with men [6]. Additionally, 17 cases of HAV in Tel Aviv have been linked to men having sex with men [7]. Between January and August of 2017, there has also been an increase in HAV infection in men who have sex with men in New York City, with 46 identified patients as of September 22, 2017 [8] .

Dataset Description:

The dataset of Hepatitis risk factors obtained from the UCI Machine Learning Repository is utilized for this research and is shown in Table 2 underneath. The dataset is made out of demographics, custom, and clinical records of 155 patients, with 19 attributes/features. Table 2 presents every one of the features existing in the Hepatitis dataset and the data type of those features. Some patients chose not to respond to a portion of the subjects in view of their protection concerns. The traits addressed by whole numbers, floating numbers, and Boolean (1 or, 2) are viewed as data types. The missing values for the whole number sort have been loaded up with the Boolean example means values.

Table 1: Description of the Cervical Cancer Dataset

Serial No.	Attributes/Features	Available Data	Missing Data	Data Type
1	Age	155 (100%)	0 (0%)	Integer
2	Sex	155 (100%)	0 (0%)	Boolean
3	Steroid	154 (99.36%)	1 (0.64%)	Boolean
4	Antivirals	155 (100%)	0 (0%)	Boolean
5	Fatigue	154 (99.36%)	1(0.64%)	Boolean
6	Malaise	154 (99.36%)	1 (0.64%)	Boolean
7	Anorexia	154 (99.36%)	1 (0.64%)	Boolean
8	Liver Big	145(93.54%)	10 (6.45%)	Boolean
9	Liver Firm	144(92.9%)	11 (7.1%)	Boolean
10	Spleen Palpable	150 (96.77%)	5 (3.23%)	Boolean
11	Spiders	150 (96.77%)	5 (3.23%)	Boolean
12	Ascites	150 (96.77%)	5 (3.23%)	Boolean
13	Varices	150 (96.77%)	5 (3.23%)	Boolean
14	Bilirubin	149 (96.13%)	6 (3.87%)	Float
15	ALK Phosphate	126 (81.29%)	29 (18.71%)	Integer
16	SGOT	151 (97.42%)	4 (2.58%)	Integer
17	Albumin	149 (89.68%)	16 (10.32%)	Float
18	Protein	89 (56.77%)	66 (43.23%)	Integer
19	Histology	155 (100%)	0 (0%)	Boolean
20	Outcome	155 (100%)	0 (0%)	Boolean

Methodology:

This paper compares data-mining models, namely, LR, NB, SVM, DT, MLP and RF for analysis of the reason for hepatitis and accurate prediction of the disease.

Step 1: All records from the hepatitis dataset are collected and read.

DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

Step 2:SVM Attribute Eval is applied to get the relevant features.
Step 3: The classifiers are trained using the validation set. A train-test partitioning and 10-fold cross-validation techniques are used for testing purposes.
Step 4: Classification of optimal feature subset using algorithms such as LR, NB, SVM, DT, MLP and RF.

SVM Attribute Eval:
Evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class seperately using a one-vs-all method and then "dealing" from the top of each pile to give a final ranking.

Logistic Regression (LR):
LR is named after the logistic equation the capability applied at the centre of the framework. The logistic method, too perceived as the sigmoid capability, was made by analysts to make sense of the environmental features of populace development increment dramatically and upgrade environmental execution. Mainly an S-formed curve can get any number with a genuine value and guide it to a value somewhere in the range of 0 and 1, yet under no circumstances at those cut off points. A predictor or, assortment of predictors of a dichotomous dependent variable, like the presence of hepatitis in patients, the patient who survived or, passed on, or, the patient who responded or, didn't make a recovery from treatment, is evaluated by Logistic Regression. Independent variables are stretch, proportion level (or nonstop), nominal or, ordinal (ranked) type data [9]. This can then be generalized to forecast the utilization of a few predictors on a collection of dependent variables [10]. Each of the activities of taking care of oneself is a dichotomous variable; with either 'yes' or, 'no' reactions (no response uncovers 'missing' information). Models are built utilizing logistic regression from the statistics that better depict the connections.

Naïve Bayes (NB):
Naïve Bayes Classifier is a straightforward kind of Bayesian network classifier in view of the utilization of the Bayes Theorem, with a solid assumption of the nonalignment of the attributes. Because of its conventionality, dependability, and great data set execution, the Naïve Bayes (NB) is treated as the most well-known classification type. NB is struggling with data sets where confounded property conditions are available. For large data sets, the NB classification doesn't produce accurate outcomes [11].

Support Vector Machine (SVM):
SVMs are one sort of efficient ML technique with a high generalization limit in practice. They are a gathering of edge classification models proposed by Vapnik and his gathering at AT&T Laboratories during the 1990s [13]. In differentiation to the techniques for statistical learning in view of practical risk minimization, the target of SVM is to minimalize structural risk, which explains a strong ability to avoid over-fitting [12]. In the SVM model, a choice hyperplane is utilized for a separation gap that splits the maximum limit between two classes. Contrasted with traditional ML approaches, SVMs have been used in many fields for their boundless speculation capacities. In specific, as a data-driven prediction technique, lately, SVM models certainly stand out in the determination of infections.

Multi-layer Perceptron (MLP):
It trains a multilayer perceptron with one hidden layer by minimizing the squared error and a quadratic penalty with the BFGS method. The edge parameter is utilized to decide the penalty on the size of the weights. The number of hidden units can likewise be determined. At last, utilizing conjugate is a conceivable inclination plunge instead of BFGS updates, which might be quicker for cases with many parameters. An estimated variant of the logistic function is utilized as the activation function to make it quick. Additionally, assuming that delta values in the backpropagation step are inside the client-determined tolerance, the inclination isn't refreshed for that specific occurrence, which saves some extra time. Parallel computation of squared error and inclination is conceivable at the point when

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

numerous CPU cores are available. Data is split into groups furthermore, handled in discrete strings for this situation. Note that this just improves runtime for bigger datasets. Nominal attributes are handled utilizing the unsupervised Nominal to Binary channel, what's more, missing qualities are supplanted worldwide by utilizing Replace Missing Values.

Decision Tree (DT):

Classification with decision trees is fundamentally founded on the formation of a decision tree, the use of each occurrence in the data set to that tree, and the grouping of the case as per the outcome. Contrasted and other groupings strategies, decision trees are more straightforward to produce and comprehend. There are many algorithms created in view of decision trees, what's more, these algorithms contrast concerning the selection of root, node, and branching criteria [14].

Random Forest (RF):

RF is a well-known supervised classification technique utilized in different classification fields. It is an ensemble learning method [15] that works on the idea of utilizing a collection of feeble students to set up areas of strength for a. RF utilizes the Classification and Regression Tree (CART) techniques [16] to make a combination of numerous choice trees in light of the bootstrap aggregation (bagging) technique [17]. The CART methodology correctly classified the dependent and independent variables and makes a connection between them. In RF, each tree randomly picks a subset of the dataset to construct an independent decision tree. RF parts chose an irregular subset from the root node to the child node over and again until each tree arrives at the leaf node without pruning. Each tree independently classifies the features and the objective variable and decisions in favour of the final tree class. Depending upon the edge of the votes cast, RF decides the final overall classification.

Performance Matrix:

Execution appraisal of the proposed work is achieved utilizing the accompanying measures. A confusion Matrix is used to survey the exhibition of a learning model. Four terms, connected with the confusion Matrix are applied to lay out the performance matrices. The quantity of hepatitis disease patients classified as hepatitis patients is a True Positive (TP). False Positive (FP) is the quantity of non-hepatitis patients delegated to hepatitis disease patients. True Negative (TN) is the quantity of patients that are named non-hepatitis patients without hepatitis. False negative (FN) is the number of patients with delegated disease patients without hepatitis [18].

Accuracy: It is the ratio between numbers of correctly predicted instances to the total number of instances.

$$\text{ACCURACY} = \frac{TP + TN}{TP + FP + FN + TN}$$

Recall/Sensitivity: It measures the proportion of individuals with hepatitis and individuals predicted by an algorithm to be hepatitis patients.

$$\text{RECALL} = \frac{TP}{TP + FN}$$

Specificity: It evaluates the proportion of individuals without hepatitis and individuals predicted by an algorithm to be non-hepatitis patients.

$$\text{SPECIFICITY} = \frac{TN}{TN + FP}$$

Kappa Statistics: A chance-corrected method for evaluating agreement between raters is Cohen's kappa () statistic [19]. Kappa is defined as follows:

$$K_{STAT} = \frac{A_{OBS} - A_{EXP}}{N - A_{EXP}}$$

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

AUC-ROC Curve: AUC-ROC or, AUROC (Area Under the Receiver Operating Characteristics) is a probability curve denoting the ability of a model to differentiate between classes in the case of binary classification. ROC indicates an exchange between True Positive Rate (TPR) and False Positive Rate (FPR). AUC signifies degree or, a measure of separability where the value closer to 1 means an algorithm effectively classifies patients with and without Hepatitis.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Results and Discussion:

The following research questions are tended to in this research article: Which Data Mining Procedure (DMT) is best for anticipating sicknesses like hepatitis disease? what's more, Which DMT structure can help with meeting the three rules of consistency, sensitivity, and specificity? To accomplish the most elevated levels of consistency, sensitivity, and specificity, the author thinks about the most famous methodologies and researches their group. Previous authors have focused exclusively on reducing variables to improve prediction. However, this method results in a loss of data. In this manner, the author lays out a system in this paper that proposes the utilization of data mining approaches, the estimation of consistency utilizing kappa statistics, and the improvement of specificity and sensitivity parameters utilizing a group learning approach. As a result, the system introduced in this paper adds to mankind's prosperity by considering better sickness prediction.

Table 2: Hepatitis Dataset with 19 Features and 1 Target Variable:

Attributes
Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, ALK Phosphate, SGOT, Albumin, Protein, Histology, Outcome.

Table 3: Hepatitis Dataset with 7 Features and 1 Target Variable:

Attributes
Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big.

Table 4.1: Comparison of Accuracies with 19 Features

Method	LR	NB	SVM	DT	MLP	RF
50-50	75.3247	75.3247	74.026	76.6234	79.2208	74.026
66-34	72.5499	75.4902	75.4902	64.7059	79.4118	80.3922
80-20	60.4839	73.3871	66.9355	75.8065	63.7097	75.8065
10F	76.129	78.0645	76.7742	79.3548	74.8387	79.3548

Table 4.2: Comparison of Accuracies with 7 Features

Method	LR	NB	SVM	DT	MLP	RF
50-50	89.6104	87.013	92.2078	92.2078	92.2078	88.3117
66-34	90.1961	86.2745	93.1373	93.1373	94.1176	95.098
80-20	72.5806	76.6129	74.1935	74.1935	78.2258	79.8387
10F	83.871	85.1613	94.8387	100	97.4194	96.129

As displayed in Table 4.1 and Table 4.2, different machine learning classifiers have yielded fluctuating levels of accuracy. 6 classifiers, specially LR, NB, SVM, DT, MLP and RF performed

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

exceptionally with a accuracy of more than 60% for a particular number of features i.e. 19 and 7. LR classifier quantified accuracy in ascending order with lowest using 19 features and highest using 7 features. This leads to highly optimistic outcome and does not reflect the actual predictive performance of the model. The exclusion of excess variables guaranteed improvement in the classification accuracy of hepatitis patients yet in general anticipated accuracy could show a shrinking effect. In this unique circumstance accuracy is not the best measurement for assessing predictive performance.

Table 5.1: Comparison of Sensitivity for with 19 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.893	0.827	0.865	0.799	0.865	0.827
66-34	0.905	0.844	0.828	0.719	0.813	0.891
80-20	0.714	0.821	0.744	0.859	0.679	0.821
10F	0.891	0.885	0.813	0.823	0.813	0.865

Table 5.2: Comparison of Sensitivity for 7 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.962	0.893	0.893	0.893	0.893	0.964
66-34	0.957	0.892	0.905	0.905	0.946	0.946
80-20	0.89	0.681	0.648	0.648	0.703	0.736
10F	0.883	0.918	0.927	1.000	0.964	1.000

As displayed in Table 5.1 and Table 5.2, different machine learning classifiers have yielded fluctuating levels of sensitivity. 6 classifiers, specially LR, NB, SVM, DT, MLP and RF performed exceptionally with a sensitivity of more than 65% for a particular number of features i.e. 19 and 7. LR classifier quantified sensitivity in ascending order with lowest using19 features and highest using 7 features.

Table 6.1: Comparison of Specificity for with 19 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.35	0.375	0.368	0.3793	0.3043	0.3684
66-34	0.352	0.3030	0.314	0.4736	0.2926	0.2424
80-20	0.5306	0.3414	0.444	0.2894	0.490	0.3548
10F	0.296	0.2340	0.305	0.2786	0.3214	0.2439

Table 6.2: Comparison of Specificity for 7 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.24	0.2608	0.222	0.222	0.2222	0.0869
66-34	0.2187	0.2666	0.2	0.2	0.1333	0.1481
80-20	0.5098	0.4677	0.492	0.4923	0.45	0.41176
10F	0.2727	0.225	0.151	0	0.0816	00

As displayed in Table 6.1 and Table 6.2, different machine learning classifiers have yielded fluctuating levels of specificity. 6 classifiers, specially LR, NB, SVM, DT, MLP and RF performed exceptionally with a specificity of more than 0.2340 for a particular number of features i.e. 19 and 7. LR classifier quantified specificity in ascending order with lowest using7 features and highest using 19 features.

Table 7.1: Comparison of Kappa statistic for with 19 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.4065	0.4314	0.3683	0.4882	0.5161	0.3683
66-34	0.4	0.4613	0.4672	0.2451	0.5666	0.5475
80-20	0.1645	0.4168	0.2884	0.4624	0.2394	0.3148
10F	0.4853	0.5158	0.5074	0.565	0.4611	0.4476

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

Table 7.2: Comparison of Kappa statistic for 7 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.7528	0.6821	0.8197	0.8197	0.8197	0.9364
66-34	0.7643	0.6629	0.8401	0.8401	0.8555	0.776
80-20	0.402	0.5323	0.4953	0.4953	0.5578	0.5779
10F	0.606	0.6277	0.881	1	0.939	1

As displayed in table 7.1 and table 7.2, different machine learning classifiers have yielded fluctuating levels of kappa statistic. 6 classifiers, specially LR, NB, SVM, DT, MLP and RF performed exceptionally with a kappa statistic of more than 0.1645 for a particular number of features i.e. 19 and 7. MLP classifier quantified kappa value in ascending order with lowest using 19 features and highest in DT(J48), RF classifiers using 7 features.

Table 8.1: Comparison of AUC for with 19 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.79	0.818	0.673	0.862	0.796	0.832
66-34	0.791	0.831	0.73	0.661	0.847	0.874
80-20	0.651	0.836	0.644	0.757	0.763	0.836
10F	0.852	0.857	0.754	0.815	0.837	0.89

Table 8.2: Comparison of AUC for 7 Features

Method	LR	NB	SVM	DT(J48)	MLP	RF
50-50	0.912	0.898	0.946	0.967	0.906	1.000
66-34	0.931	0.895	0.953	0.968	0.99	0.957
80-20	0.821	0.916	0.824	0.824	0.912	0.908
10F	0.935	0.929	0.964	1.000	0.988	1.000

The performance metric, ROC-AUC score given in Table 8.1 and Table 8.2, is used for comparing the performance of several classifiers and has provided clarity than accuracy, sensitivity, and specificity. Kappa statistic gives the agreement rate between the expected and predicted outcome where values ranging from (1.0), (0.81-0.99), (0.61-0.80), (0.41-0.60), (0.21-0.40), (0.1-0.20) to (0) represent perfect, near-perfect, significant, moderate, fair, slight and close to near possibility agreements respectively. All classifiers with seven features and 10-fold cross validation confirmed the great agreement in terms of kappa statistic value as displayed in Table 7.1 and Table 7.2. In general, RF is provided the best outcomes for accuracy, sensitivity with 7 features, followed by LR, MLP, SVM, DT, NB and RF. The reduction in feature subspaces from 19 to 7, through feature determination, worked on the presentation of all classifiers as exhibited in Table 9.

Table 9: Effect of Feature Reduction

Performance Metrics	19 Features			7 Features		
	ML Techniques	Maximum Score	Train – Test Split	ML Techniques	Maximum Score	Train – Test Split
Accuracy (In Percentage)	RF	80.39	66-34	DT(J48)	100	10-Folds
Sensitivity	LR	0.905	66-34	DT(J48), RF	1	10-Folds
Specificity	LR	0.5306	80-20	LR	0.5098	80-20
Kappa Statistics	MLP	0.5666	66-34	DT(J48), RF	1	10-Folds

DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

AUC	RF	0.874	66-34	DT(J48), RF	1	10-Folds, 50-50

Conclusion:

Hepatitis is brought about by an assortment of irresistible infections and non-infectious specialists prompting a scope of medical conditions. There are five fundamental kinds of hepatitis infection (A, B, C, D, and E). They all cause liver sickness, they contrast in significant ways including methods of transmission, the severity of the illness, geographical distribution, and prevention methods. Specifically, HBV and HCV lead to ongoing illness in a huge number of individuals and together are causing liver cirrhosis, and liver malignant growth. An expected 354 million individuals overall live with hepatitis B or C, and in general, testing and treatment stay distant.

A few sorts of hepatitis are preventable through vaccination. A WHO investigation discovered that an expected 4.5 million unexpected losses could be prevented in low-and centre pay nations by 2030 through immunization, diagnostic tests, prescriptions, and training efforts. WHO's worldwide hepatitis procedure, supported by all WHO Part States, plans to decrease new hepatitis contaminations by 90% and deaths by 65% somewhere in the range between 2016 and 2030.

Reference:

1. Daniels, D. (2009). Centers for Disease Control and Prevention (CDC). Surveillance for acute viral hepatitis-United States, 2007. *MMWR Surveil Summ*, *58*, 1-27.

2. Klevens, R. M., Miller, J. T., Iqbal, K., Thomas, A., Rizzo, E. M., Hanson, H., ... & Spradling, P. (2010). The evolving epidemiology of hepatitis a in the United States: incidence and molecular epidemiology from population-based surveillance, 2005-2007. *Archives of internal medicine*, *170*(20), 1811-1818.

3. Bohm, S. R., Berger, K. W., Hackert, P. B., Renas, R., Brunette, S., Parker, N., ... & Teshale, E. H. (2015). Hepatitis A outbreak among adults with developmental disabilities in group homes—Michigan, 2013. *Morbidity and mortality weekly report*, *64*(6), 148.

4. Koenig, K. L., Shastry, S., & Burns, M. J. (2017). Hepatitis A virus: essential knowledge and a novel identify-isolate-inform tool for frontline healthcare providers. *Western Journal of Emergency Medicine*, *18*(6), 1000.

5. Peak, C. M., Stous, S. S., Healy, J. M., Hofmeister, M. G., Lin, Y., Ramachandran, S., ... & McDonald, E. C. (2020). Homelessness and hepatitis A—san diego county, 2016–2018. *Clinical infectious diseases*, *71*(1), 14-21.

6. Update, E. (2017). hepatitis A outbreak in the EU/EEA mostly affecting men who have sex with men.

7. Gozlan, Y., Bar-Or, I., Rakovsky, A., Savion, M., Amitai, Z., Sheffer, R., ... & Mor, O. (2017). Ongoing hepatitis A among men who have sex with men (MSM) linked to outbreaks in Europe in Tel Aviv area, Israel, December 2016–June 2017. *Eurosurveillance*, *22*(29), 30575.

8. Latash, J., Dorsinville, M., Del Rosso, P., Antwi, M., Reddy, V., Waechter, H., ... & Balter, S. (2017). Notes from the field: increase in reported hepatitis A infections among men who have sex with men—New York City, January–August 2017. *Morbidity and Mortality Weekly Report*, *66*(37), 999.

9. Connelly, L. (2020). Logistic regression. *Medsurg Nursing*, *29*(5), 353-354.

10. Tzeng, E., Devin, C., Hoffman, J., Finn, C., Abbeel, P., Levine, S., ... & Darrell, T. (2020). Adapting deep visuomotor representations with weak pairwise constraints. In *Algorithmic Foundations of Robotics XII: Proceedings of the Twelfth Workshop on the Algorithmic Foundations of Robotics* (pp. 688-703). Springer International Publishing.

11. Jabbar, M. A., & Samreen, S. (2016, October). Heart disease prediction system based on hidden naïve bayes classifier. In *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)* (pp. 1-5). IEEE.

12. Zhang, L., Zhou, W., & Jiao, L. (2004). Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *34*(1), 34-39..
13. Ayat, N. E., Cheriet, M., & Suen, C. Y. (2005). Automatic model selection for the optimization of SVM kernels. *Pattern Recognition*, *38*(10), 1733-1745.
14. Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. *Int J Intell Syst Appl*, *13*(5), 46-63.
15. Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23. [16] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5-32, 2001.
16. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.
17. Ray, A., & Chaudhuri, A. K. (2021). Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. *Machine Learning with Applications*, *3*, 100011.
18. Chalak, L. F., Pavageau, L., Huet, B., & Hynan, L. (2020). Statistical rigor and kappa considerations: which, when and clinical context matters. *Pediatric research*, *88*(1), 5-5.
19. Chalak, L. F., Pavageau, L., Huet, B., & Hynan, L. (2020). Statistical rigor and kappa considerations: which, when and clinical context matters. *Pediatric research*, *88*(1), 5-5.
20. Nusrath Unnisa A; Manjula Yerva; Kurian M Z. "Review on Intrusion Detection System (IDS) for Network Security using Machine Learning Algorithms". *International Research Journal on Advanced Science Hub*, 4, 03, 2022, 67-74. doi: 10.47392/irjash.2022.014
21. Mohammadibrahim Korti; Basavaraj S. Malapur; Smita Gour; Rajesh M. Biradar. "Shuchi 1.0: Robotic System For Automatic Segregation of Waste & Floor Cleaning". *International Research Journal on Advanced Science Hub*, 4, 02, 2022, 31-37. doi: 10.47392/irjash.2022.007
22. Bhuneshwari Nayak; Rachana Choudhary; Roymon M. G.. "Isolation, Screening and Morphological characterization of Laccase producing fungi". *International Research Journal on Advanced Science Hub*, 4, 02, 2022, 38-43. doi: 10.47392/irjash.2022.008
23. Khaled Salem Ahmad Amayreh; Ahmad Taufik Hidayah Bin Abdullah. "Conjunction in Expository Essay Writing by Jordanian Undergraduate Students Studying English as a Foreign Language (EFL)". *International Research Journal on Advanced Science Hub*, 4, 02, 2022, 24-30. doi: 10.47392/irjash.2022.006
24. Pravin T, M. Subramanian, R. Ranjith,Clarifying the phenomenon of Ultrasonic Assisted Electric discharge machining, "Journal of the Indian Chemical Society", Volume 99, Issue 10, 2022, 100705, ISSN 0019-4522, Doi: 10.1016/j.jics.2022.100705
25. R. Sudhakaran, P.S. Sivasakthivel, M. Subramanian, Investigations on the effect of surface coatings on the weldment properties on chromium manganese stainless steel gas tungsten arc welded plates, Materials Today: Proceedings,Volume 46, Part 17,2021, Pages 8554-8560, ISSN 2214-7853, Doi: 10.1016/j.matpr.2021.03.541
26. Sivalakshmi B; Mahisha sri K.B; Swetha P. "Survey on Mammogram, Ultrasound, MRI, Spectroscopy, Biopsy for Detecting Tumor in Breast". *International Research Journal on Advanced Science Hub*, 3, 2, 2021, 30-37. doi: 10.47392/irjash.2021.027
27. Hari Prasada Raju Kunadharaju; Sandhya N.; Raghav Mehra. "Detection of Brain Tumor Using Unsupervised Enhanced K-Means, PCA and Supervised SVM Machine Learning Algorithms". *International Research Journal on Advanced Science Hub*, 2, Special Issue ICSTM 12S, 2020, 62-67. doi: 10.47392/irjash.2020.262
28. Suresh P; Justin Jayaraj K; Aravintha Prasad VC; Abishek Velavan; Mr Gokulnath. "Deep Learning for Covid-19 Identification: A Comparative Analysis". *International Research Journal on Advanced Science Hub*, 4, 11, 2022, 272-280. doi: 10.47392/irjash.2022.068

29. Rajesh P.; Vetrivel Govindarasu. "Analyzing and Predicting Covid-19 Dataset in India using Data Mining with Regression Analysis". *International Research Journal on Advanced Science Hub*, 3, Special Issue 7S, 2021, 91-95. doi: 10.47392/irjash.2021.216
30. Menonjyoti Kalita; Golam Imran Hussain. "Determining the Influencing Factors of COVID 19 on Mental Health Using Neural Network". *International Research Journal on Advanced Science Hub*, 3, Special Issue 6S, 2021, 126-129. doi: 10.47392/irjash.2021.177
31. Ajitha K; Samuel Joseph C; Mahila Vasanthi Thangam D. "Online marketing of agricultural products during COVID pandemic: Farmers and customers perspectives". *International Research Journal on Advanced Science Hub*, 3, Special Issue 6S, 2021, 94-101. doi: 10.47392/irjash.2021.173
32. Yeshi Ngima; Dorjee Tsering. "Impact of COVID-19 on Education". *International Research Journal on Advanced Science Hub*, 2, Special Issue ICAMET 10S, 2020, 34-39. doi: 10.47392/irjash.2020.196
33. Siddavatam rammohan reddy; Balaji krushna potnuru. "3D Printing Innovation during Covid-19 Pandemic". *International Research Journal on Advanced Science Hub*, 2, 8, 2020, 62-67. doi: 10.47392/irjash.2020.95
34. Pooja Dahiya; Roopsi Kaushik; Anil Sindhu. "Corona virus: an Overview Along with Its Alternative Diagnostic Measures". *International Research Journal on Advanced Science Hub*, 2, Special Issue ICARD 2020, 2020, 163-169. doi: 10.47392/irjash.2020.113
35. Remya S. "Covid19 and Environment-A Theoretical Review from Higher Education Students Perspective". *International Research Journal on Advanced Science Hub*, 2, Special Issue ICARD 2020, 2020, 227-230. doi: 10.47392/irjash.2020.124
36. Mohd. Akbar; Prasadu Peddi; Balachandrudu K E. "Inauguration in Development for Data Deduplication Under Neural Network Circumstances". *International Research Journal on Advanced Science Hub*, 2, 6, 2020, 154-156. doi: 10.47392/irjash.2020.55
37. Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty; Sriranga H.A. "Machine Learning: An Intuitive Approach In Healthcare". *International Research Journal on Advanced Science Hub*, 2, 7, 2020, 67-74. doi: 10.47392/irjash.2020.67
38. Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. "Machine learning amalgamation of Mathematics, Statistics and Electronics". *International Research Journal on Advanced Science Hub*, 2, 7, 2020, 100-108. doi: 10.47392/irjash.2020.72
39. Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty. "Latent Approach in Entertainment Industry Using Machine Learning". *International Research Journal on Advanced Science Hub*, 2, Special Issue ICARD 2020, 2020, 304-307. doi: 10.47392/irjash.2020.106
40. Suneetha V; Salini Suresh; Niharika Sinha; Sabyasachi Prusty; Syed Jamal J. "Enhancement in the World of Artificial Intelligence". International Research Journal on Advanced Science Hub, 2, Special Issue ICARD 2020, 2020, 276-280. doi: 10.47392/irjash.2020.132
41. Logeswari T.. "Performance Analysis of Ml Techniques for Spam Filtering". International Research Journal on Advanced Science Hub, 2, Special Issue ICIES 9S, 2020, 64-69. doi: 10.47392/irjash.2020.161
42. Hari Prasada Raju Kunadharaju; Sandhya N.; Raghav Mehra. "Detection of Brain Tumor Using Unsupervised Enhanced K-Means, PCA and Supervised SVM Machine Learning Algorithms". International Research Journal on Advanced Science Hub, 2, Special Issue ICSTM 12S, 2020, 62-67. doi: 10.47392/irjash.2020.262Export Citation
43. Maneesha M; Savitha V; Jeevika S; Nithiskumar G; Sangeetha K. "Deep Learning Approach For Intelligent Intrusion Detection System". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 45-48. doi: 10.47392/irjash.2021.061
44. Kalki N; Karthick M; Mr Kavin; Keerthana S; Sangeetha K. "Advanced Face Mask Detection System". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 112-115. doi: 10.47392/irjash.2021.076

DOI:10.46647/ijetms.2023.v07i02.023 ISSN: 2581-4621

45. Naveenkumar S; Kirubhakaran R; Jeeva G; Shobana M; Sangeetha K. "Smart Health Prediction Using Machine Learning". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 124-128. doi: 10.47392/irjash.2021.079
46. Kavya Shakthi R.P; Kavin Raja A.S; Janani S.R; Sangeetha K. "Industrial Machine Identification Using Augmented Reality". *International Research Journal on Advanced Science Hub*, 3, Special Issue ICARD-2021 3S, 2021, 68-71. doi: 10.47392/irjash.2021.066
47. Salma Begum; Sampurna P.. "A Study on growth in Technology and Innovation across the globe in the Field of Education and Business". *International Research Journal on Advanced Science Hub*, 3, Special Issue 6S, 2021, 148-156. doi: 10.47392/irjash.2021.181
48. Dhanya S Karanth; Kumaraswamy H.V; Rajesh Kumar. "Workaround prediction of cloud alarms using machine learning". *International Research Journal on Advanced Science Hub*, 3, 8, 2021, 164-168. doi: 10.47392/irjash.2021.231
49. Gyanendra Kumar Pal; Sanjeev Gangwar. "Discovery of Approaches by Various Machine learning Ensemble Model and Features Selection Method in Critical Heart Disease Diagnosis". *International Research Journal on Advanced Science Hub*, 5, 01, 2022, 15-21. doi: 10.47392/irjash.2023.003