TEST Engineering & Management

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

# A Survey on Prediction of Fatty Liver Disease by Using Machine Learning Techniques

[1]L.Ravali, [2]J.Swami Naik,[3]Dr.N.Kasiviswanath

[1]M.Tech(pursuing),CSE Department,G.Pulla Reddy Engineering College(Autonomous),Kurnool,AndhraPradesh,India

ravalisetty666@gmail.com

[2]Associate Professor, CSE Department, G.Pulla Reddy Engineering College(Autonomous),Kurnool,AndhraPradesh,India

swaminaikcse@gmail.com

[3]Professor & Head, CSE Department ,G.Pulla Reddy Engineering College(Autonomous),Kurnool,AndhraPradesh,India

hodcse@gprec.ac.in

***Article Info***
***Volume 82***
***Page Number: 12342 -12350***
***Publication Issue:***
***January-February 2020***

***Article History***
***Article Received****: 18 May 2019*
***Revised****: 14 July 2019*
***Accepted****: 22 December 2019*
***Publication****: 23 February 2020*

***Abstract***
Fatty liver infection (FLD) is an umbrella term for some sorts of liver illnesses. As the name proposes, the fundamental clinical issue, is an excess of fat put away in liver cells. An early determination of patients with FLD will assist doctors with making a fitting technique for counteraction, early conclusion, and treatment. An ML model is planned which will help doctors in ordering theoretical patients, and cause another conclusion, to forestall and oversee FLD.This model represents the comparison of the four classification algorithms on different benchmark dataset to evaluate classification performance and predict fatty liver disease accurately.Likewise, it also presents the performance of the Fatty liver disease (FLD) prediction depends on following scaling factors such as Accuracy, Precision, Sensitivity(Recall), and Specificity. Usage of the ML model in the clinical setting could assist doctors with stratifying greasy liver patients for essential anticipation, in the early hours treatment, and the board.

***Keywords:*** *Fatty Liver Disease, Machine learning, Learningalgorithms.*

## I. INTRODUCTION

Greasy liver malady (FLD) is a usualmedicalproblem; it is related with dismalness and death. FLD likewise prompts non cholestatic cirrhosis and hepato cell carcinoma [1]. Furthermore, FLD has been expanding in corresponding with the pervasiveness of diabetes, metabolic disorder and weight [2]. Higher regularity of FLD has appeared as a progressively imperative fiscal weight. Right now, unmistakable confirmation of individuals at serious risk and early affirmation of FLD could offer immense favorable circumstances for end, preventive or fundamentally authentic treatment. Over the earlier decade, the biopsy has been used. This procedure is incredibly prominent and over the top; it also may trigger side effects and testing bumbles during the utilization of this system. Regardless of the way that, ultrasonography is using

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

as a helpful instrument for FLD end with higher precision, while recognizing exactness is astoundingly director subordinate [3].

AI (ML) is a field of programming building that uses PC estimations to recognize plans in enormous data, and help to predict the various outcomes subject to data[4].ML systems have been developed as a potential apparatus for expectation and basic leadership in a large number of disciples [5].Due to the accessibility of clinical information, ML has been assuming a basic job in restorative basic leadership as well[6,7].Developing an AI model would likewise fill in as a significant guide to recognize infection and settle on a compelling clinical choice. It would likewise take into account enhancement of medical clinic assets by characterizing right patients with a few hazard factors fundamentally at a beginning time.

### A. Why is the liver important?

The liver is that the second biggest organ inside the body and is found beneath the skeletal structure within the higher right aspect of the abdomen. It performs hundreds of functions vital to health and well-being including regulation of metabolism, production of clotting proteins and blood detoxification.

### B. What is FLD?

Greasy liver malady (steatosis) is a typical clinical issue brought about by the gathering of bound fats inside the liver. The liver for the most part contains a limited quantity of fat. In any case, if fat percent in the liver is more than 5 to 15 percent then the patient has fatty liver disease.

### C. Types of FLD

There are two main forms of FLD*:*

- **Non-alcoholic greasy liver:** Development of fat within the liver which isn't identified with drinking liquor.
- **Alcoholic greasy liver:** Fat develop in the liver because of utilization of a lot of liquor.

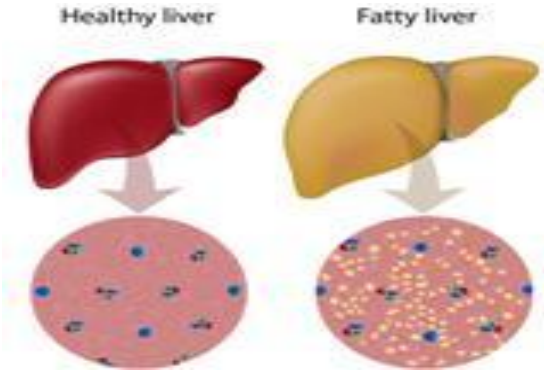



Figure1: Difference between healthy liver and fatty liver

### D. Risk Factors of FLD

The risk factors for FLD include:

- Being hefty or overweight
- Having type 2 diabetes
- Having metabolic disorder
- Some genetic metabolic conditions or prescription medications.

### E. Symptoms of FLD

Individuals with FLD in some cases don't have any side effects. On the off chance that manifestations do appear, they may include:

- A sentiment of totality inside the centre or higher feature of the belly
- Abdominal torment
- Loss of appetency or weight reduction
- Queasiness
- Achilles' heel
- Jaundice (yellowing of the skin and furthermore the whites of the eyes)
- Swelling of the midriff and legs (edema)
- Mental perplexity

TEST Engineering & Management

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

### F. How is fatty liver diagnosed?

Certain blood tests act as screeners, but do not provide a definitive diagnosis. Elevated levels of the enzymesALT, AST and GGT may indicate fatty liver disease.

Imaging concentrates, for example, ultrasound, CT examines what's more, and MRI outputs can push suppliers to outwardly decide whether fat has penetrated the liver. These tests, however, cannot determine how far the disease has progressed. If assessment of severity is needed, a liver biopsy will be necessary.

### G. Is fatty liver disease curable?

There is no medication or procedure which will cure illness. However, reductionof weight often has a positive effect.

### H. What am I able to do to stop the unwellness from progressing?

- Produce associate degree exercise routine. Move for a minimum of a hundred and eighty minutes per week, like by walking with an acquaintance or mate.
- Eat the Mediterranean Diet that consists of fruits and vegetables (5 servings daily), whole grains, beans and buggy. Replace butter with healthier fats, like further virgin vegetable oil.
- Avoid soda and completely different types of sweetening. Attempt to limit intake to fifteen grams of levulose per day. However, analysis shows that restricted amounts of bittersweet chocolate will be useful.
- Specialise in protein-rich foods (2-3 servings daily) like lean poultry, seafood, eggs, and Greek dairy product. Avoid chicken.
- Occasional is also useful. there's proof that 2-3 cups daily will halt progression of NAFLD
- Ponder taking a daily vitamin E supplement of 800 IU. Vitamin E may be a powerful inhibitor and should relieve symptoms of NAFLD.
- Raise your medical care supplier concerning taking a daily Bayer to deal with curdling problems. If your levels of sterol square measure over ancient, jointly raise regarding sterol lowering medication
- Create a meeting with a specialize. World Health Organization will design Associate in Nursing applicable diet, additionally as aid with weight loss if this can be your goal.
- Follow up a minimum of quarterly along with your primary automotive supplier or a GI specialist to watch the health of your liver

## II.METHODOLOGY

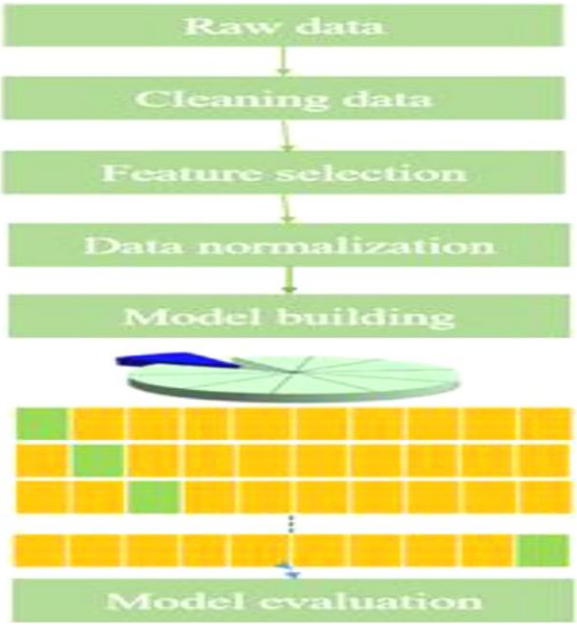



Figure 2: Architecture of FLD recognition

### I. Liver Dataset

Input Attributes of the training dataset contains 10 attributes. The attributes are as follows:

- Age

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

- Gender
- Total_Bilirubin
- Direct_Bilirubin
- Alkaline_Phosphotase
- Alanine_Aminotransferase
- Aspartate_Aminotransferase
- Total_Protiens
- Albumin
- Albumin_and_Globulin_Ratio

## II. Cleaning Data

Cleaning information is a premier measure to decipher each ML difficulties .In request to get exact outcomes from ML calculations, datasets need to purge and change. The most generally utilized pre-preparing procedure is to substitute missing qualities if the characteristics. There are hardly any missing qualities in dataset which are supplanted to prepare ML calculations.

## III. Feature Selection

The different feature selection methods used are as follows:

- **Filter Method**: Filtering dataset and taking just a subset of it containing all the pertinent highlights.
- **Wrapper Method**: We feed a few highlights to our AI model, assess their exhibition and afterward choose to add or evacuate highlight to build exactness.
- **Embedded Method**: It analyzes the diverse preparing emphases of our ML model and afterward positions the significance of each element.

## IV. Feature Optimization

Based on the feature selection, we are going to consider the features only with highest priority and ignore the remaining

## . Model Building

Prescient grouping models were created to distinguish FLD patients precisely. Characterization is a directed learning approach in which the PC gains from the information input given to it and afterward utilizes this to arrange new perception. The various kinds of order calculations utilized are as per the following:

- Support Vector Machine
- Nearest Neighbor
- Decision Trees
- Random Forest

## III. CLASSIFICATION ALGORITHMS

### 1. Support Vector Machine

Support Vector Machine initially saw by methods for Vapnik in 1979 [9]. It turns out to be again suggested by method for Vapnik in 1995 for relapse and arrangement [8]. Support vector is utilized for design class [11] which is formed by multilayer perceptron and spiral premise trademark systems. The SVM is the propelled age with most classification calculations installed in measurable acing theory.SVM procedures are utilized in kind of direct and non-straight records. It changes the bonafide preparing measurements into higher size utilizing non-direct mapping .Within this new size it looks for straight most appropriate detaching hyperplane. Information from classes can be isolated by methods for hyperplane with the exact nonlinear mapping to an adequately high measurement. Utilizing help vectors and edges the SVM uncovers those hyperplane [10]. SVM executes the class task by methods for augmenting the edge orders both class while limiting the sort blunders. Despite the fact that the SVM can be completed to different enhancement inconveniences which incorporate relapse, the great difficulty is that of data class.

TEST Engineering & Management

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

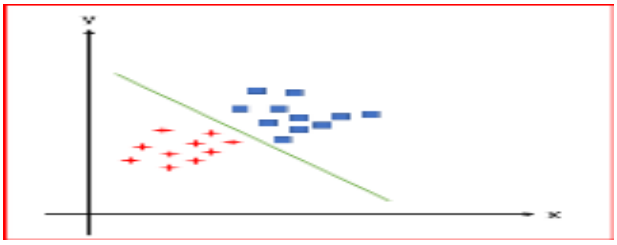


Figure 3: Data Classification

The measurements focuses are analyzed as being successful or negative, and the issue is to discover a hyper plane that isolates the records factors by utilizing a maximal edge. Figure 3: Data Classification the decide 2 propose the 2-dimensional case wherein the data factors are straightly divisible.

$$\min_{\vec{w},b} \frac{1}{2}\|w\|,$$

$$s.t \quad y_i = +1 \Rightarrow \vec{w}\cdot\vec{x}_i + b \geq +1$$
$$y_i = -1 \Rightarrow \vec{w}\cdot\vec{x}_i - b \leq -1 \qquad (1)$$

$$s.t \; y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1, \quad \forall i$$

The identification of the each data point $x_i$ is $y_i$, which can take a value of +1 or -1 (representing positive or negative respectively). The solution hyper-plane is the following:

$$u = \vec{w}\cdot\vec{x} + b \qquad (2)$$

The scalar b is also termed the bias.

A standard method to solve this problem is to apply the theory of Lagrange to convert it to a dual Lagrangian problem. The dual problem is the following:

$$\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j (\vec{x}_i \cdot \vec{x}_j)\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i$$
$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (3)$$
$$\alpha_i \geq 0, \quad \forall i$$

The variables $\alpha_i$ are the Lagrangian multipliers for corresponding data point $x_i$.

## 2. Nearest Neighbor

In design acknowledgment, the k-closest neighbor's equation (k-NN) is additionally a non-parametric philosophy utilized for order and regression.In each case, the information comprises of the k most noteworthy instructing work models inside the element house. The yield relies upon whether or not k-NN is used for characterization or relapse:

In k-NN gathering, the yield is similarly a class enrolment. Accomplice in nursing object is studied by a lion's share vote of its neighbors, with the thing being chosen to the class commonest among its k nearest neighbors (k is in like manner a positive number, consistently little). In case k = 1, by then the thing is simply chosen to the class of that singular nearest neighbor.

- In k-NN backslide, the yield is that the property estimation for the thing. This value is that the customary of the estimations of k nearest neighbors.

K-NN is additionally a strategy for instance based learning, or unconcerned acknowledging, where the work is essentially approximated locally and each one estimation is deferred until portrayal. Both for game plan and backslide, a supportive method is to dispense burdens to the responsibilities of the neighbors, so as that the closer neighbors contribute piles of to the routinely than the stores of far away ones for example, a standard weight subject involves in giving each neighbor a heap of 1/d, where d is that the gap to the neighbor. The neighbors unit taken from a social affair of articles that the class (for k-NN portrayal) or the thing property estimation (for k-NN backslide) is known.This might be thought of because of the utilization set for the recipe, albeit not all out instructing work step is required. An idiosyncrasy of the k-NN recipe is that it's delicate to the local structure of the data.

## 3. Decision Tree

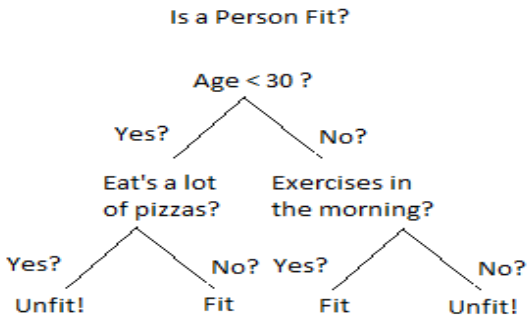
A Decision Tree is a toolthat uses a tree like chart of choices and their outcomes. It is a straightforward

TEST Engineering & Management

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

portrayal for arranging models. It is a Supervised Machine Learning where the information is ceaselessly part as indicated by a specific parameter. Choice Tree comprises of:

- **Nodes:** Represents test for an estimation of a characteristic.
- **Edges/Branch:** Represents the result of a test and it interfaces with the following hub or leaf.
- **Leaf hubs:** Represent class labels.They are additionally called terminal hubs used to foresee the result or class.




**How Decision Trees work: Algorithm**

·Initially fabricate a tree.

- Always Start from information to root hub .
- Select a property and play out a coherent test on that trait.
- Branch on every single result of the test, and move the subset of models which fulfils that result to the comparing youngster hub.
- Recurse on accessible youngster hubs.
- Repeat until all leaves are "unadulterated", i.e., they have model from a solitary class, or "almost unadulterated", i.e., the greater part of the models are from a similar class.
- Prune tree.
- Remove subtrees which don't improve the exactness of order.
- Avoid over-fitting.

**How tobuild tree**

· For all attributes evaluate split-points ·

- Select "best" point and "winning" attribute
- Split data into two parts ·
- Use breadth/depth-first construction ·

**4. Random Forest**

Random forests are a blend of tree indicators such every tree relies upon the estimations of an arbitrary vector inspected severally and with a comparable circulation for all trees in the woods. The speculation blunder for backwoods merges on a farthest point in light of the fact that the assortment of trees inside the woods gets monster. The speculation mistake of woods of tree classifiers relies upon the quality of the individual trees inside the timberland and the relationship between's them. Utilizing an arbitrary selection of choices to separate each hub yields mistake rates that contrast well with Adaboost (Freund and Schapire [1996]), however are increasingly powerful as for commotion. Inward gauges screen mistake, quality, and relationship and thesear wont to demonstrate the reaction to expanding the amount of choices utilized in the parting. Interior assessments are wont to live factor significance. These ideas are applicabletoregression.Critical improvements in classification if cation exactness have come about because of developing partner degree troupe of trees and material belonging them vote in favor of the principal regular class. So as to develop these gatherings, ordinarily arbitrary vectors are created that administer the development of each tree in the troupe. An early model is curio (Breiman [1996]), any place to develop each tree an irregular decision (without substitution) is made from the models inside the instructing set. Another model is irregular part choice (Dietterich [1998]) any place at each hub the split is choosen arbitarily from among the K best parts. Breiman [1999] creates new training sets by randomizing the yields inside the first instructing set. Another methodology is to pick

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

the instructing set from an arbitrary arrangement of loads on the models inside the training set. Ho [1998] has composed various papers on "the arbitrary subspace" system that will an irregular decision of set of alternatives to use to develop each tree.In a significant paper on composed character acknowledgment, Amit and Geman [1997] characterize a larger than usual scope of geometric alternatives and search over an irregular decision of those for the best split at each hub. This last paper has been incredible in my reasoning. The basic part out and out of those methodology is that for the kith tree, an arbitrary vector Θk is created, independent of the past irregular vectors Θ1, Θk−1 yet with consistent circulation; and a tree is full-developed abuse the training set and Θk, bringing about a classifier h(x, Θk) where x is an info vector. For example, in packing the arbitrary vector Θ is produced as the includes in N boxes resulting from N darts tossed self-assertively at the crates, where N is number of models in the preparation set. In irregular split decision Θ comprises of various independent arbitrary numbers among one and K.The nature and spatiality of Θ relies upon its utilization in tree development. After a larger than average assortment of trees is created, they vote in favor of the preeminent in style class. The point of this subsequent from N darts tossed indiscriminately at the cases, where N is number of models in the preparation set. In arbitrary split decision Θ comprises of various independent irregular numbers among one and K. The nature and spatial property of Θ relies upon its utilization in tree development. After a larger than average scope of trees is created, they vote in favor of the most famous class .We call these systems arbitrary

## IV. LITERATURE SURVEY

In 2018 Nazmun Nahar and Ferdousy Ara et. al [18] used different decision tree techniques for Liver Disease Prediction using machine learning. Early forecast of illness is essential to spare human life and find a way to control the sickness. Choice Tree calculations are with progress applied in differed fields. The illness dataset that is decide for this investigation is comprises of qualities like all out hematoidin, direct hematoidin, age, sex, complete proteins, straightforward protein and basic protein extent connection. The point of this work is to ascertain the exhibition of different choice tree methods and think about their presentation. The choice tree systems utilized are J48, LMT, Random Forest, Random tree, REPTree and Decision Stump.Weka is used as a data mining tool which is written in java and developed at Waikato for comparing the performance of various decision trees. These calculations gave different outcomes dependent on Accuracy, Mean Absolute Error, Precision, Recall, Kappa insights and Runtime. These methods were assessed and their exhibitions were analyzed. From the examination, Decision Stump has accomplished a precision of 70.67% and it beat well when contrasted with different calculations.

Later in 2019 Chieh-Chen Wua and Wen-Chun Yehb et.al [19] has portrayed various calculations for expectation of greasy liver sickness utilizing ML. In this paper they have incorporated all patients who had a fundamental oily liver screening at the New Taipei City Hospital some place in the scope of first and 31st December 2009.Classification models, for instance, Random Forest (RF), Naïve Bayes (NB), Artificial neural frameworks (ANN), and calculated relapse (LR) were used to foresee FLD. The gatherer working trademark twist (ROC) was used to evaluate execution among the four models. Among the four counts used the arbitrary backwoods model showed superior to anything the other gathering modelsExecution of arbitrary woodland calculation in the clinical setting could assist doctors with stratifying greasy liver patients for essential counteraction, early treatment, and the board.

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

Again in 2019 another work has been done by Binish Khan, Piyush Kumar Shukla,and Manish Kumar Ahirwar et.al [20] on Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms. The work mainly focused on analyzing the parameters of various classification algorithms and comparing their predictive accuracies so as to find out the best classifier for determining the liver disease. Various attributes that are useful in the prediction of liver disease were examined and the dataset of liver patients were also evaluated. They compared various classification algorithms such as Random Forest, Logistic Regression and Separation Algorithm with an aim to identify the best technique. Thus, Random forest with 100% accuracy has outperformed well.

## V. CONCLUSIONS

The goal of this literature review is to compare four ML techniques on different benchmark FLD data set. The four techniques are(a) Support Vector Machine, (b) Nearest Neighbor, (c) Decision Tree, and (d) Random Forest algorithm for getting accurate results. Finally, analyzing the results with the help of Comparing Models and Confusion Matrix

## VI. REFERENCES

Journal papers

[1] M. Lazo , J.M. Clark , in: The Epidemiology of Nonalcoholic Fatty Liver Disease: A Global Perspective: Seminars in Liver Disease, 28, (c) Thieme Medical Pub- lishers, 2008, pp. 339–350 .

[2] M.H. Le , P. Devaki , N.B. Ha , D.W. Jun , H.S. Te , R.C. Cheung , M.H. Nguyen , Preva- lence of non-alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States, PLoS One 12 (2017) e0173499 [3] Q.M. Anstee , G. Targher , C.P. Day , Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis, Nat. Rev. Gastroenterol. Hepatol. 10 (2013) 330–344.

[4] M. Motwani , D. Dey , D.S. Berman , G. Germano , S.Achenbach , M.H. Al-Mallah , D. Andreini , M.J. Budoff, F. Cademartiri , T.Q. Callister , Machine learning for pre- diction of all-cause mortality in patients with suspected artery dis- ease: a 5-year multicentre prospective registER analysis, Eur. Heart J. 38 (2016) 500–507 .

[5]Sani A. Machine Learning for Decision Making, Université de Lille 1, 2015,

[6]W.Raghupathi, massive knowledge analytics in healthcare: promise and potential, Health Inf.Sci. Syst. 2 (2014). [7]P.Groves, B.Kayyali, D.Knott, S.V.Kuiken, The Big Data Revolution in Health-care: fast price and Innovation, 2016

[8]Grimaldi.M, Cunningham.P, Kokaram.A, associate analysis of other feature selection methods and ensemble techniques for classifying music, in: Work-shop in multimedia system discovery and Mining, ECML/PKDD03, Dubrovnik, Croatia, 2003.

[9] Voltaic Emre Güraksın, Hüseyin Hakli, Harun U˘guz, Support vector machines classification based on particle swarm optimization for bone age determination, Elsevier publications, Science direct, page no 597- 602

[10]Han, J.; Kamber.M. “Data Mining ideas and Techniques”. 2nd Edition, Morgan Kaufamann, San Francisco.

[11]Karthik.S, Priyadarshini.A, Anuradha.J, and Tripathi.B.k, Classification and Rule Extraction victimization Rough set for diagnoising of disease and its varities, Advances in Applied Science Research, 2011, 2(3): page no 334-345.

[12]Kotsiantis.S.B, Increasing the Classification Accuracy of easy Bayesian Classifier, AIMSA, pp.198-207, 2004.

[13] Milano Kumari, Sunila Godara, Comparative Study of information Mining Classification strategies in Disorder Prediction, International Journal of Computer Science and Technology, Vol.2, Iss ue 2, June 2011, page no 304-308

[14]Omar S.Soliman, Eman Abo Elhamd, Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine, International Journal of Scientific & Engineering Research, Volume 5, Issue 3,

*January - February 2020*
*ISSN: 0193 - 4120 Page No. 12342 -12350*

March-2014 122
[15] Patrick Breheny, Kernel density classification, STA 621: Nonparametric Statistics October 25
[16]Pushpalatha.S, Jagdesh Pandya, knowledge model comparison for liver disease diagnosing, International Journal of rising analysis in Management &Technology ISSN: 2278-9359 (Volume-3, Issue-7) 2014, page no 138-141.
[17]Yugal Kumar and G.Sahoo**,** "Prediction of different types of liver diseases using rule based classification model", IOS press,2013,P.p:417-432,DOI:10.3233/THC-13074
[18]Nazmun Nahar and Ferdous Ara, "Liver Disease Prediction By Using Different Decision Tree Techniques", International journal of data Mining and Knowledge Management Process,2018,DOI:10.512/ijdkp.2018.8201
[19]Chieh-Chen Wua, Wen-Chun Yehb**,** "Prediction of fatty liver disease using machine learning algorithms", Computer Methods and Programs in Biomedicine, 2019, P.p:23-29.
[20]Binish Khan, Piyush Kumar Shukla, Manish Kumar Ahirwar, "Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms**",** International Journal of Computer Science and Engineering, Vol: 7, issue: 7, July 2019.