ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

*International Journal of Mechanical Engineering*

# Intelligent Techniques and Comparative Performance Analysis of Liver Disease Prediction

**Sreenivasa Rao Veeranki and Manish Varshney**

Department of Computer Science and Engineering, School of Engg. & Tech., Maharishi university of Information Technology, Lucknow, INDIA

**Abstract:**

Bioinformatics is the process of analysing gene structures living organisms. The complete characteristics of living organisms whether it is bad or good is completely controlled by the human gene. The human genes are made of protein sequences. The genetic data of living organism is very large. The analysis of huge amount data is very critical. The bioinformatics is such an area of research to deal with huge genetic data. These amounts of data can be analysed only by the techniques of Big Data belonging from computer science. Different tools of Big Data used to handle the huge amount of genetic data. The process of Bioinformatics has been applied in the field of medical industry very well. Every organism's gene structure is very much important to identify for knowing the characteristics of the organisms. It is helpful to create drugs to fight against different viruses and bacteria and other microorganisms. The bioinformatics helps to identify the characteristics of the harmful micro biological organisms to create protection against them. In this work, tools of bioinformatics have been used in the domain of liver disease. The genetic data of liver patient has been analysed using the methods belonging from bioinformatics to find the ways to protect the disease.

**Keywords:** Bioinformatics, Genome Structure Analysis, Protein-Protein Interaction, Liver Disease, Machine Learning, Random Forest, Multilayer Perceptron, K-Nearest Neighbour, Support Vector Machine

## 1. Introduction:

Liver is a major organ of the body which is located in the right-side portion and above of the abdominal cavity but just beneath the diaphragm, top of the stomach, under the rib cage. The size of a liver of the human being is large and reddish-brown in colour. Liver is a most important organ of the body of every life form who are vertebrates. Liver should do their proper functions to maintain the body healthy and feet. Liver has major functions like, digestion of foods which we uptake from day to night, secretion of many types enzymes which are required to carry out different biochemical pathways such as, glycolysis, gluconeogenesis, tri-carboxylic acid cycle, protein synthesis, beta-oxidation of fatty acids, detoxification of xenobiotic compounds etc. These biochemical pathways are highly essential to be regulated by the different factors or biochemical components of the body to keep the body feet and energetic. These biochemical pathways are carried out to produce sufficient energy in the form of adenosine tri-phosphate (ATP), so that a vertebrate living body can do their work energetically without feel sick. If these central biochemical pathways are not do their work appropriately as well as enzymes are not regulated properly, then disease will occur in the liver. As the liver is a central organ of the body, so any kind of abnormality in the liver make harm in other organs present in the body, resulting that, cause diseases of other organs. If the diseases are not cured properly, then it will hamper the immune structure of the body as well as losing the ability to combat against the acute diseases. The biological chemistry is responsible for making a living creature in the universe. The detection and treatment of the disease can be done by various methods. Biochemical methods of detection as well as treatment are one of the best methods. There are different biochemical markers are present in the body,

by detecting their amounts present in the blood or serum of the body, we can diagnose that liver is working properly or not.

Liver is a very much essential compartment of the body as it helps in foods digestion resulting in converting them into simple nutrients from complex compounds by the help of various enzymes present within the body. Major as well as important biochemical pathways are accomplished by the enzymes present in the liver cells. Serum glutamic pyruvic transaminase (SGPT) and serum glutamic-oxaloacetic transaminase (SGOT) are enzymes present in liver cells in normal health condition but when these enzymes are released into the bloodstream as well as increase their amount in the blood, indicate that the liver does not function correctly and the liver has been damaged. Other names of SGPT and SGOT are alanine transaminase (ALT) and aspartate aminotransferase (AST). A high amount of these two enzymes in the bloodstream indicate that the liver will be damaged completely in future, if it is not cured properly. Liver can be damaged by various viral attacks, fibrosis, cancer, alcohol consumption, bacterial and parasite attack, appropriate metabolic functions, abnormal gene function etc. These phenomena can alter liver structure and responsible for liver malfunctioning.

**Various Functionalities of The Liver:**

- Liver enzymes help to detoxify the toxic or unwanted part of the solid as well as liquid foods and remove them from blood circulation of the body through excreting them as waste products from the body.
- Liver produce bile salts which help in converting fats into fatty acids in the liver. Bile is a water-soluble end product of cholesterol in the liver. Bile salts are stored and concentrated in the gall-bladder and it release from the gall-bladder into bile duct. Fat soluble vitamins are also digested by bile salts of liver.
- Bile consists of cholesterol, phospholipids, conjugated bile acids, bile pigments and electrolytes (Smet, 1994). Bilirubin and biliverdin are bile pigments which have a strong anti-oxidative property, resulting in scavenging the super-oxide radicals. After deconjugation of bile acids, it becomes less soluble and absorbed by the intestines and this leads to their elimination in the faces as free bile acids (Kumar, 2012).
- Liver is a most critical organ in the body as it is an only compartment where metabolism of carbohydrates, amino acids, fatty acids, various types of drugs, takes place. From the metabolism, every kind of food nutrients are broken down into simple form of nutrients which can be easily absorb by the small intestine. A body can get energy through nutrient metabolism as well as electron transport chain.
- Liver enzymes and hormones help in regulating blood sugar level, blood cholesterol level by maintaining metabolism properly. Most number of drugs is activated by the hormones, enzymes of the liver.
- Liver stores glycogen through insulin action as well as store vitamins and minerals which are very much important to maintain a good health.

**Various Sorts of Liver Diseases:**

Liver diseases can be caused by various reasons such as,

1. Viral infections like, Hepatitis A, Hepatitis B and Hepatitis C
2. Autoimmune diseases like, autoimmune hepatitis
3. Cancers like, liver cancer, bile duct cancer, liver cell adenoma
4. Over consumption of various drugs create effect on the liver function adversely
5. Genetic diseases which are inherited through genes from parents like, hemochromatosis, alfa-1-antitrypsin deficiency etc.
6. Disease causes due to regular alcohol consumption like, liver fibrosis, cirrhosis of liver

7. Due to fat deposition into the liver, causes diseases like, non-alcoholic fatty liver disease (NAFLD)

**Diagnosis And Treatment of The Various Liver Disease:**

Liver diseases can be diagnosed by biochemical tests, imaging techniques, immunological techniques etc. Different Biochemical tests for the detection of the liver diseases are following:

1. Normal amount of total bilirubin is 1.2 milligrams per decilitre (mg/dl) for adults and usually 1 mg/dl for those who are under 18(years). Normal amount of direct bilirubin is 0.3 mg/dl.
2. According to Karla Blocka, Normal range of alanine aminotransferase in blood is 4 – 36 U/L (Units per litre).
3. The normal range of alkaline phosphatase is 44 – 147 international units per litre (IU/L) or 0.73 to 2.45 microkatal per litre.
4. The normal range of aspartate aminotransferase in blood is 8 – 33 U/L.
5. The normal range for gamma-glutamyl transferase (GGT)in blood is 5 – 40 U/L for adults.
6. The normal range of albumin in blood is 3.4 – 5.4 g/dL (34 – 54 g/L).
7. The normal range of sugar level in blood is 99 mg/dL or below.
8. The normal triglyceride level in blood is below 150 mg/dL.

These all-normal ranges of bile salts, enzymes, sugar and cholesterol are slightly may change depending on measurements and sampling techniques by the different laboratories. If the amount of these biochemical factors increases or decrease beyond the normal range then it indicates that liver function is somehow altered as well as the liver cells are injured or infected. Depending on the types of liver diseases, medications are provided to the patients by the doctors.

Liver is a central organ of the body so it plays a vital role to maintain a good health. Digestion of foods to production of energy and growth as well as detoxification of the toxic compounds is main functions of the liver. Liver is exposed to the various toxic compounds, various pathogens, xenobiotic compounds etc. Liver produces different enzymes, hormones which are regulated properly and regularly to defend those kinds of threats and makes a sound health.

The biological chemistry is a system of the body which should be regulated by the biochemical factors itself and build up a sound health to live the life healthier and happier.

## 2. Literature Review:

The “Bioinformatics” is the area of research, which has been invented by two great researchers Ben Hesper and Paulien Hogeweg. The invention of the term “Bioinformatics” had been created in the early stage of the 1970s. The two researchers Ben Hesper and Paulien Hogeweg used the term “Bioinformatics” in their research work. In this research work, the term bio informatics has been defined as “the study of informatics processes in biotic systems” (Hogeweg, 2011). Further modification of the Bioinformatics had been done in the year 1981. In the year, two other researchers had modified the Bioinformatics domain. The two researchers ware Marvin Carruthers and Leory Hood. The two researchers had found five hundred seventy-nine genes from human being, which had been mapped using situ hybridization. The researchers Marvin Carruthers and Leory Hood invented a DNA sequencing process that is an automated process (Oyelade, 2015). Bioinformatics has a very prominent application in human gene. Bioinformatics has been applied quite successfully in the research of human genome and this research helps to deal with different diseases. For better usability of the bioinformatics in the field of human genome, human genome projects had been taken and for successful maintenance of the projects related with human genome using Bioinformatics an organization was established. The main concern of the organization is to monitor the ongoing projects on human genome and find the future path of the human genome research using Bioinformatics. The organization is named as Human Genome Organization. The human genome organization was established in the year 1988. After long time after the creation of human genome organization, the concept of Bioinformatics with molecular

biology was combined. Huang et al mixed the two concepts Bioinformatics and molecular biology for the first time in the year 2000. The fusion of two concepts Bioinformatics and molecular biology was applied in the medical domain. In this work, the trajectories of neutrophil ware demonstrated in the alternative ways. This research work has been done using expression of temporal gene data (Hogeweg, 2011). In the research domain of Bioinformatics, three things have been combined. These three things are 'Biological signal analysis', 'management', and 'interpretation'. Huge databases on biological sequences and gene structure have been generated and for maintaining the databases, two data repositories ware created. The two repositories were – 'EMBL (European Molecular Biology Laboratory)' and 'Gen-Bank'. These data repositories contain DNA data that can be used for matching different DNA sequences. The current research on Bioinformatics has aiming towards the medical industries. For the medical industry applications, several fields of biometric research have been encouraged. These fields are like – functionality of the human gene, protein structure of the gene, analysis if the protein structure, and metabolic rout comparison for various species comparison (Müller, 2005). The research of the bioinformatics is mainly the analysis of the big protein structure data. The protein structure data of human genome is being generated by technology related to the molecular biology. The data basically is the array of DNA data in large scale and there are several tools available for building positions of protein coding areas. The possibility of various quality verbalization assessment is used to uncover the DNA chip involves huge number of nucleotide areas in thousands. It is far off from every other person used to perceive the quality coding DNA structure the model and track down the subsets of characteristics. In genomics and proteomics astonishing achievements are more important in data mining systems. The accomplishment rate in ID of characteristics related wrecks is depending upon the exposure of novel quiets and its sufficiency. Various around the world, public, autonomous and great goal enabling together to store the genetic information in kind of electronic prosperity records (Swinney, 2014). The computation and understanding issues related with the nuclear level can be tended to by using math and estimations. The computer programming and information advancement is used to arrangement suitable computational instruments (Koonin, 2002). The DNA structure is the main input of the bioinformatics-based research. The DNA structures have been analysed to know the protein structure of the human genome. The protein structures are the key to identify the disabilities in human being. Therefore, for the applications in the medical sector, the bioinformatics tries to analyse the protein structure of human genome.

According to the recent state of the art researches, bioinformatics can be applied in the medical field in the following sections. These sections are – 'Sequence Analysis', 'Protein Structure Prediction', 'Annotation of Human genome', 'genome comparison', and 'Discovery of drug to protect diseases. Here, one research work of recent years from each section, have been discussed.

The genetic structure is the sequence of human genes. The genes are the building block of any living organisms. Sequence Analysis is the process of understanding the structure, functionalities, and features of the gene structure. Computer Science is very much helpful in the process of Sequence analysis. Computer Science offers several tools, which are very much powerful to analyse the sequence of genomes. Each of the such tools has some advantages and disadvantages. These tools have been applied on the merit of problems addressed. These tools are capable of identify the mutation of protein structure in the DNA of any organisms. Shotgun sequence technique is the best example of tools used for sequence analysis (Breda, 2007).

Protein structure prediction is one of the finest sections of research work in the field of bioinformatics. The human genes are generated by the proteins. Proteins are the chain of amino acids. The protein structures are two types – primary protein structures and secondary protein structure. The primary protein structures are easy to be predicted because the primary protein structure is three dimensional. However, the secondary protein structures are higher than the three dimensional. Therefore, it is very complicated to determine protein structure of the secondary section. Mainly, these proteins are quaternary in structure. To predict the secondary protein structure, biometric tool can be used with the help of crystallography (Kellis, 2014).

The annotation of Genome is another very important task to identify the gene structure in the organisms. The regulatory sequence has been done by the Genome Annotation. In addition, with that protein coding has also been done using the annotation of Genomes. According to the process of annotation of

genomes, location of all genes has been identified. The process of annotation of Genome helps to find the region of coding and genome structure as well (Pellegrini, 1999).

The comparison of different parts of genes helps to determine the structure of genomics. The general Genomic features are like gene, mRNA, tRNA, rRNA, and CDS. These features have been compared to find the functional similarities among the biological organisms. The complete way of evolution occurring in the genomes of various species can be traced by the researchers using the maps of inter genomics. These traced maps can be used to find the information about the mutation point and the segments of the large chromosome.

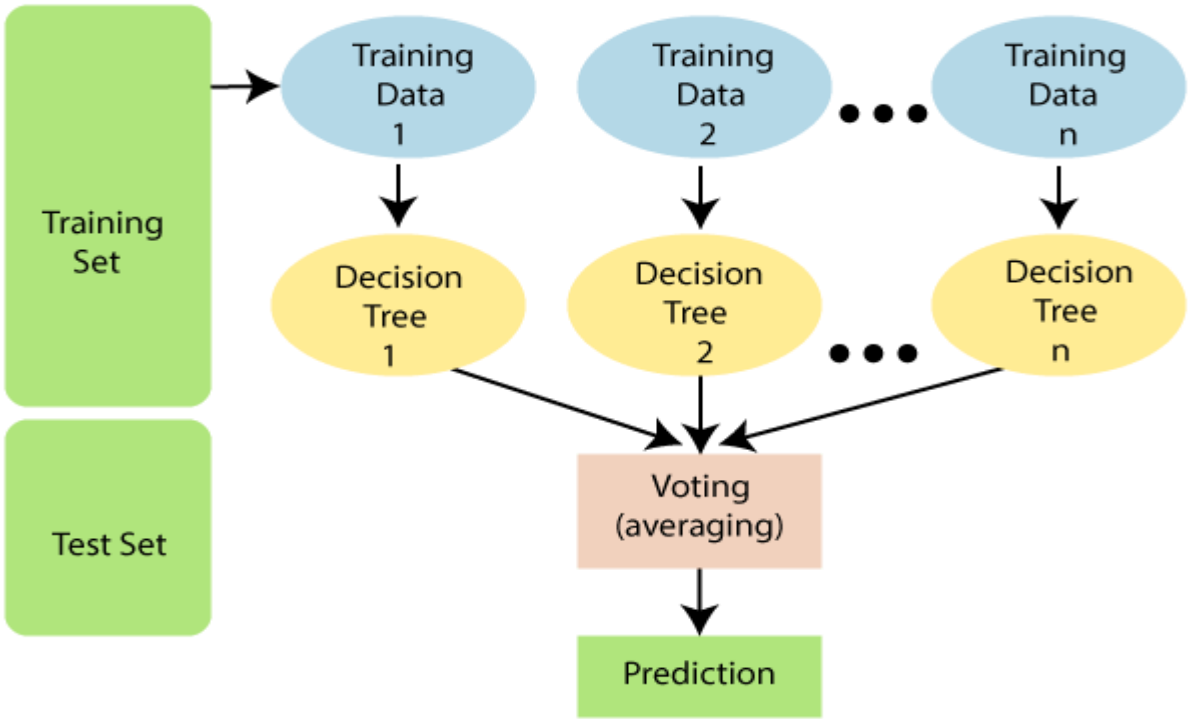
The bioinformatics tools have been used very efficiently in the work of drug discovery. The drug discovery using the bioinformatics has been done on the basis of molecular disease. The researchers can develop medicines and drugs suitable for more than 500 genes according to the 'disease and diagnosis management'. The development of drugs has been done using different powerful bioinformatics tools (Gill, 2016).

## 3. Research Methodology:

In this research-based work, the data of liver patient has been classified. The main goal of the research work is to classify the genetic data of liver patients from the genetic data of persons, who are not suffering any liver disease. The classification of genetic data is very much needed to diagnose the liver patient, by analysing the genetic data of a human being. This research work is also applicable in the research of drug discovery. The actual genetic structure of the patients suffering from liver disease has been identified properly. Therefore, it is easy to identify the genetic structure of patients suffering from liver disease. This genetic structure has been analysed to find out the defects in the gene structure. According to the defect in the gene structure, drugs can be invented. Therefore, this research work is belonging from the research domain of Bioinformatics and it has a very prominent impact in the domain of medical industry. The research field of Bioinformatics is the fusion of genetic study from biology and big data from computer science. The huge genetic data can be analysed by the techniques of the Big Data. In this research work, we have used four methods namely – Random Forest (RF), Multilayer Perceptron (MLP) model, k Nearest Neighbour (kNN), and Support Vector Machine (SVM) for the classification task and the performance of the methods have been compared.

### Random Forest (RF):

Random Forest is a classification technique belonging from the class of Supervised Machine Learning algorithms. This is a very efficient and very frequently used machine learning technique used for classification and regression task. The random forest is the modified version of decision tree. According to the process of Random Forest, decision trees have been created for different set of samples and the best voted decision tree have been chosen for the classification task. Therefore, the Random Forest algorithm is very much suitable to handle data samples containing continuous variables for the regression task and categorical variable for the classification task.




**Steps of Implementing Random Forest Algorithm:**

**1st step:** At the first step of implementing Random Forest algorithm, n number of random records have been taken initially from the data set, which have records of k numbers.

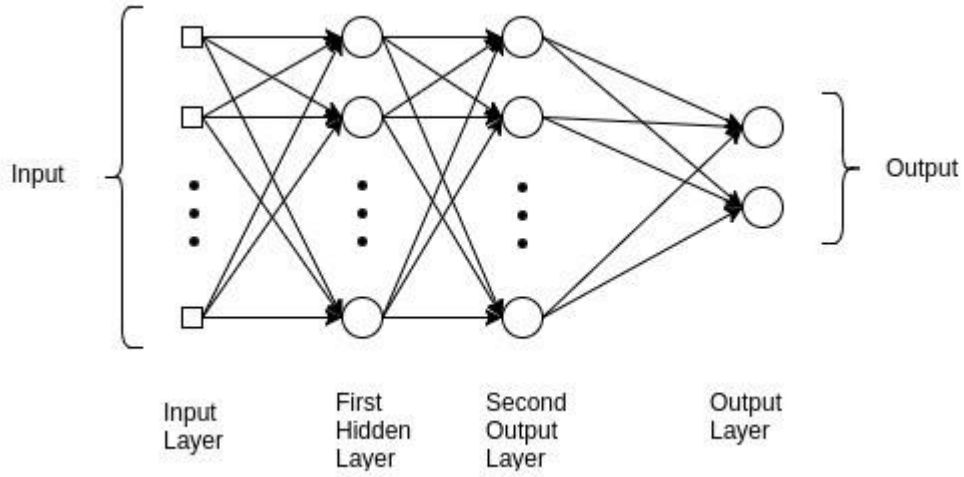
**2nd Step:** After the first step, for each of the samples considered in the first step, decision trees have been constructed for individual samples.

**3rd Step:** Each of the created decision tree has produced some output.

**4th Step:** The created decision trees have been averaged to find out the best decision tree using the process of Majority Voting for classification task or the process of averaging for the regression task.

**Multi-Layer Perceptron (MLP):**

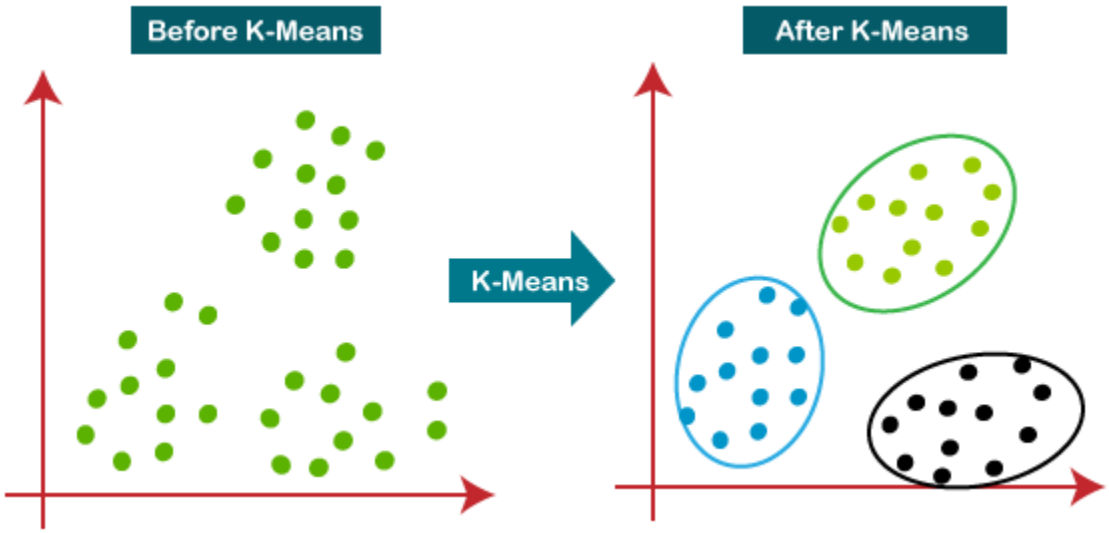
Multi-Layer Perceptron (MLP) model is a very efficient network model used for classification task. The Multi-Layer Perceptron (MLP) model is the modification of the Feed Forward Neural Network. The Multi-Layer Perceptron (MLP) model has basically three layers namely input layers, hidden layers, and output layers. These three layers are connected with each other's. The input layers taking input. The hidden layers process the input data, and the output layer produces the output. The output layer performs the actual work of prediction and classification. Between the input layer and the output layer, more than one layers can be resides. All of the layers are known as hidden layers. All of the hidden layers are connected to each other's. Therefore, these hidden layers are called connected layers. The hidden layers have been used to learn the inner structure and features from the input data.




According to the learned features the classification operations have been done. The number of hidden layers has been chosen by the designers of the model according to the complexity of the problem addressed. The hidden layers consist of neurons. The neurons have been trained using the back propagation learning algorithms. The Multi-Layer Perceptron (MLP) model is very useful for this task.

**k Nearest Neighbour (kNN)**

The k Nearest Neighbour (kNN) is the simplest supervised machine learning algorithm used for the classification task. This algorithm finds the similarity among the new cases and the previous cases. The new cases have been fallen into the class, which is the mostly closed. All the data, which are available, has been stored by the k Nearest Neighbour (kNN) algorithm. The new data has been classified by comparing with stored data. The k Nearest Neighbour (kNN) is mainly an algorithm, which is non parametric. Therefore, the algorithm has not made any assumption on the basis of the underlying data. The k Nearest Neighbour (kNN) algorithm is also known for lazy learner. The main reason is that the algorithm cannot learn from the training data instantly. It stores the data and learns from it when the new cases arrived.




The steps of the k Nearest Neighbour algorithm are as follows –

**1st Step** – At first K number of neighbours have been selected

**2nd Step** – Measure the Euclidean distance among the K number of neighbours.

**3rd Step** – The most K number of nearest neighbours have been selected on the basis of the computed Euclidean distance.

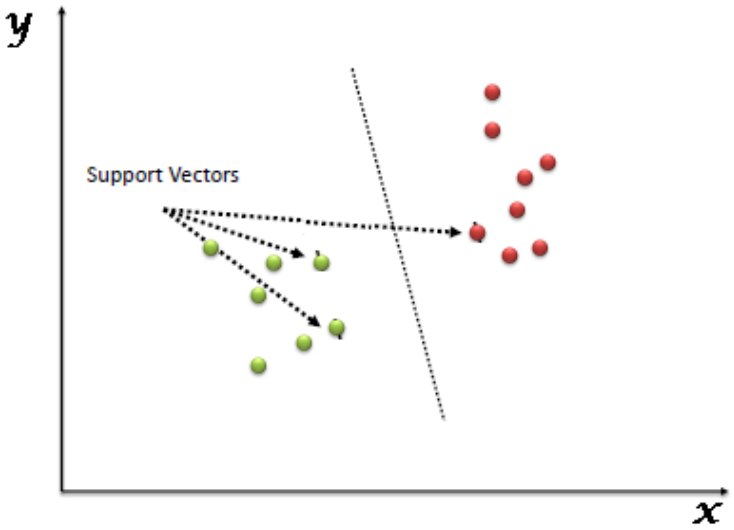
**4th Step** – Count the data points from each category from the k nearest neighbours.

**5th Step** – New data points have been assigned to the categories for which maximum number of neighbours has occurred

**6th Step** – The model is ready for the classification task.

**Support Vector Machine (SVM):**

There are several machine learning techniques are used by the data scientists now a day. Among them one of the mostly used machine learning technique is Support Vector Machine or SVM. This approach is used for the solving the problems related to classification as well as regression. In the n-dimensional space, the set of information of a particular problem is used in classification and this renowned approach is taken to create a line in n-dimensional space. This line is known as best line or boundary line for taking a decision that helps in segregation of the used set of information and thus it creates new points from the used dataset. This boundary or boundary for taking decisions is known as hyperplane that is created by choosing extreme points from the set of information. These points that are chosen in an extreme choice is known as support vectors.



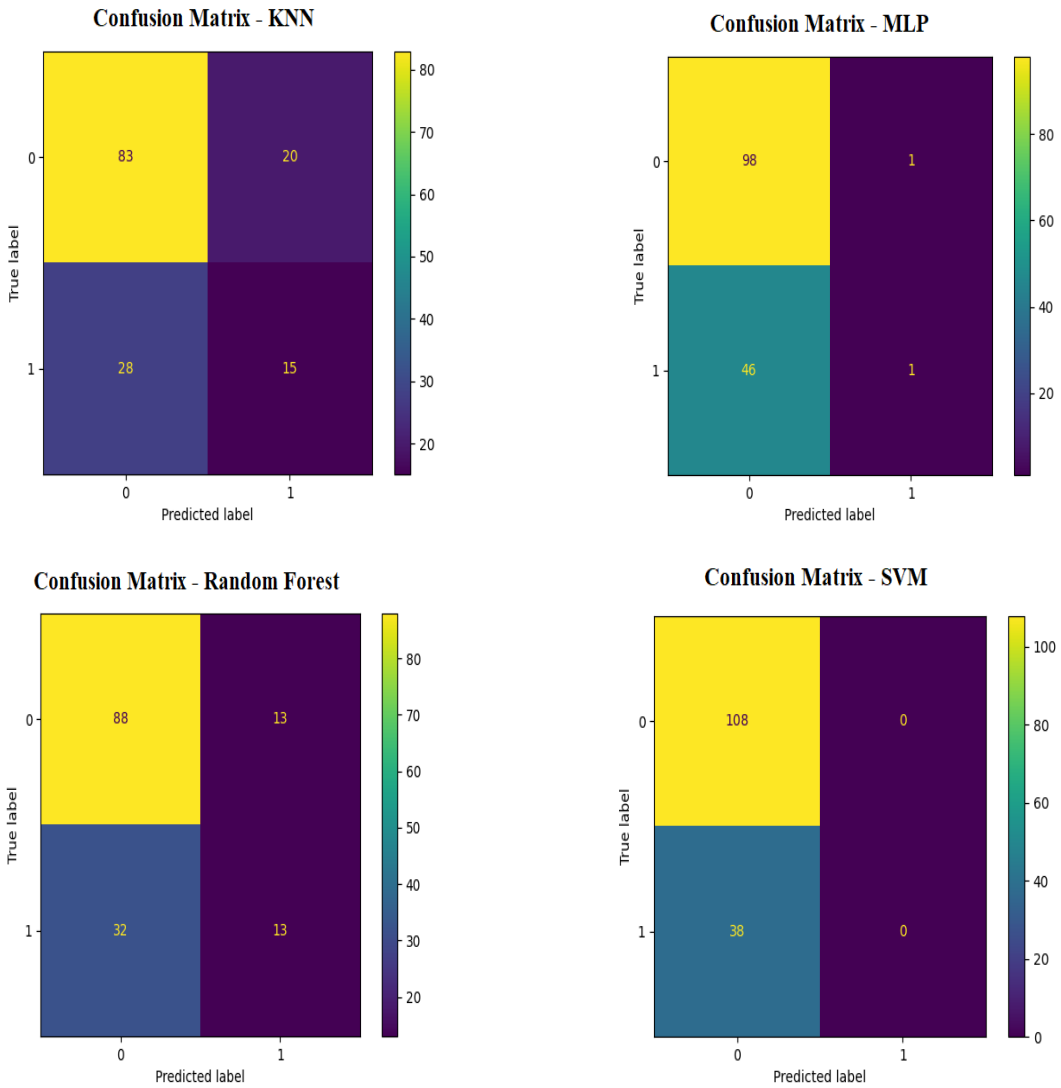

**Result and Analysis:**

This proposed work is created with the dataset of Indian Liver Patient Dataset that is collected from UCI Machine Learning Repository. For creating a successful comparative study on the proposed work, we have implemented the mostly used machine learning approaches like random forest, SVM, KNN and MLP. All of them are used to predict liver disease and thus a comparative study on their performances are created. This whole work is created by using python programming environment thus there is a huge library support comes from the python library packages that makes the proposed work efficient. With the import of library packages, the models are also called and implemented. The used dataset is needed to be converted so that the machine learning models can use them properly. The argument parser is used to convert the class label information into numbers. Once the data is ready for the processing, it is divided into two parts where the first one is the training part and the other one is the testing part. Each of the mentioned models are trained with the newly created training data and then the trained models are tested over the test data. Thus, we got individual classification report the makes the proposed work easier to find out the best fitted approach for the proposed dataset among other implemented approaches.

In this proposed work it has been found that each of the models created their own accuracy to predict the lever disease in this scenario. The accuracy scores of the individual classification are following:

- The random forest approach secures 69% of accuracy.
- The KNN approach secures 67% of accuracy.
- The SVM approach secures 74% of accuracy.
- The MLP approach secures 68% of accuracy.

The classification accuracy reports are almost same for the all approaches but the SVM is performing better when it is compared with other implemented approach. To understand the classification reports closely, here we have also pretended the confusion matrix of all of the four implemented model. We can say that confusion matrix is the graphical representation of classifiers performance that comes up with the four matrices – true positive, true negative, false positive and false negative. Each of the confusion matrices are:




**Discussion:**

It has been mentioned earlier that four most popular machine learning approaches like random forest, SVM, KNN and MLP has been used to predict the dataset of Indian Liver Patient Dataset. All of them are performed well enough and retunes with the nearly accuracy score. But it has been found that the Support Vector Machine or SVM is performed well when it has been compared with other models that are used here for the successful completion of the proposed task so it can be concluded that the SVM

approach is the best fitted approach among other used approaches in this scenario. But there is scope to increase their performance in various ways. This can be chosen in future as an extension of this proposed work.

**Conclusion:**

A comparative performance analysis of Liver Disease prediction has been performed here by using machine learning approach and hence a best fitted approach is established here. Each and every phase of the entire work has been presented that is helpful to understand the prediction and each model performance in terms of accuracy here. But finally, we can conclude that this work can be extended with other approaches and possibilities in future that makes more efficient way for the prediction of liver diseases.

**References**


1. Smet, I. De, et al. "In vitro study of bile salt hydrolase (BSH) activity of BSH isogenic Lactobacillus plantarum 80 strains and estimation of cholesterol lowering through enhanced BSH activity." *Microbial ecology in health and disease* 7.6 (1994): 315-329.
2. Kumar, Manoj, et al. "Cholesterol-lowering probiotics as potential biotherapeutics for metabolic diseases." *Experimental diabetes research* 2012 (2012).
3. Oyelade, Jelili, et al. "Bioinformatics, healthcare informatics and analytics: an imperative for improved healthcare system." *International Journal of Applied Information System* 13.5 (2015): 1-6.
4. Hogeweg, Paulien. "The roots of bioinformatics in theoretical biology." *PLoS computational biology* 7.3 (2011): e1002021.
5. Müller, Uwe R., and Dan V. Nicolau, eds. *Microarray technology and its applications*. Berlin: Springer, 2005.
6. Swinney, David C., and Shuangluo Xia. "The discovery of medicines for rare diseases." *Future medicinal chemistry* 6.9 (2014): 987-1002.
7. Koonin, Eugene, and Michael Y. Galperin. "Sequence—evolution—function: computational approaches in comparative genomics." (2002).
8. Breda, Ardala, et al. "Protein structure, modelling and applications." *Bioinformatics in tropical disease research: a practical and case-study approach [Internet]*. National Center for Biotechnology Information (US), 2007.
9. Kellis, Manolis, et al. "Defining functional DNA elements in the human genome." *Proceedings of the National Academy of Sciences* 111.17 (2014): 6131-6138.
10. Pellegrini, Matteo, et al. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proceedings of the National Academy of Sciences* 96.8 (1999): 4285-4288.
11. Gill, Supreet Kaur, et al. "Emerging role of bioinformatics tools and software in evolution of clinical research." *Perspectives in clinical research* 7.3 (2016): 115.