Advanced Approaches and Comparative Evaluation of Liver Disease Prediction using Intelligent Techniques

[1]V KIRANMAI, [2] G. SWARNALATHA, [3]J.S. RADHIKA, [4]YERRAGINNELA SHRAVANI

[1]Associate Professor, Department of CSE, Sri Indu College of Engineering & Technology
[234]Assistant Professor, Department of CSE, Sri Indu College of Engineering and Technology
Hyderabad-Telangana

ABSTRACT:

Bioinformatics is a rapidly evolving field that involves the analysis of gene structures in living organisms. The characteristics of an organism, whether positive or negative, are determined by its genes, which are composed of protein sequences. However, genetic data in living organisms can be extensive and complex, making its analysis challenging. This is where bioinformatics, a multidisciplinary field combining biology and computer science, comes into play. Big Data techniques from computer science are employed in bioinformatics to handle the massive amounts of genetic data generated by modern sequencing technologies. These tools enable researchers to process, analyse, and interpret the genetic data to gain insights into the characteristics of various organisms. In particular, bioinformatics has found valuable applications in the medical industry, where understanding the gene structures of different organisms is crucial for drug discovery and development. In this context, the focus of this work is on the application of bioinformatics tools in the domain of liver disease prediction. By analysing the genetic data of patients with liver disease using bioinformatics methods, researchers aim to identify ways to protect against the disease. This involves identifying the characteristics of harmful microorganisms, such as viruses and bacteria, and developing strategies to combat them. The use of bioinformatics techniques in this field holds great promise for improving our understanding of liver disease and developing effective preventive measures.

Keywords:

Exploring Bioinformatics for Genome Structure Analysis, Protein-Protein Interaction, and Liver Disease Prediction using Machine Learning Techniques: Random Forest, Multilayer Perceptron, K-Nearest Neighbour, and Support Vector Machine.

INTRODUCTION

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.

The Multifaceted Functions of the Liver:

- The liver enzymes play a crucial role in detoxifying solid and liquid foods, eliminating toxic or unwanted substances from the bloodstream through waste excretion from the body.
- The liver also produces bile salts, which aid in the conversion of fats into fatty acids within the liver. Bile, a water-soluble end product of cholesterol, contains various components such as cholesterol, phospholipids, conjugated bile acids, bile pigments, and electrolytes. Bile pigments, such as bilirubin and biliverdin, possess strong antioxidant properties and help scavenge superoxide radicals. Bile acids, after deconjugation, become less soluble and are absorbed by the intestines and eventually eliminated in feces as free bile acids.
- Furthermore, the liver is a vital organ responsible for the metabolism of carbohydrates, amino acids, fatty acids, and various drugs. It breaks down nutrients into simpler forms that can be easily absorbed by the small intestine, providing the body with energy through nutrient metabolism and the electron transport chain.
- Liver enzymes and hormones also play a crucial role in regulating blood sugar levels and blood cholesterol levels, thereby maintaining proper metabolism. Additionally,

the liver stores glycogen through insulin action and serves as a storage site for essential vitamins and minerals that are crucial for maintaining overall health.

Various Types of Liver Disorders:

- The liver, being a vital organ in the body, can be affected by a variety of diseases. Some common types of liver diseases include:
- Hepatitis: Hepatitis refers to the inflammation of the liver, which can be caused by viral infections (such as hepatitis A, B, C, D, or E), autoimmune conditions, alcohol abuse, drug toxicity, or other factors. Hepatitis can range from mild to severe and may result in liver damage or failure if left untreated.
- Cirrhosis: Cirrhosis is a chronic condition characterized by the scarring and hardening of liver tissue, which disrupts normal liver function. It can be caused by various factors such as chronic viral hepatitis, alcohol abuse, fatty liver disease, or other chronic liver diseases. Cirrhosis can progress over time and may lead to liver failure.
- Non-alcoholic Fatty Liver Disease (NAFLD): NAFLD is a condition where excessive fat accumulates in the liver, leading to inflammation and damage. It is commonly associated with obesity, diabetes, high cholesterol, or metabolic syndrome, and it can range from simple steatosis (fatty liver) to non-alcoholic steatohepatitis (NASH), which is a more severe form of liver inflammation.
- Liver Cancer: Liver cancer can originate in the liver (primary liver cancer) or spread to the liver from other parts of the body (secondary or metastatic liver cancer). Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer, and it is often associated with chronic liver diseases such as hepatitis B or C infection, cirrhosis, or NAFLD.
- Liver Failure: Liver failure is a serious condition in which the liver loses its ability to function properly, leading to a life-threatening situation. Acute liver failure can occur suddenly due to drug overdose, viral hepatitis, or other causes, while chronic liver failure may develop over time due to prolonged liver disease or cirrhosis.
- Genetic Liver Diseases: There are various genetic liver diseases such as Wilson's disease, hemochromatosis, alpha-1 antitrypsin deficiency, and others that are caused by inherited genetic mutations affecting liver function.
- It's important to note that liver diseases can have different causes, symptoms, and treatment approaches, and proper medical evaluation and management are crucial in diagnosing and managing these conditions.

Diagnosis and Treatment of Liver Diseases:

The diagnosis of liver diseases involves various biochemical tests, imaging techniques, immunological techniques, etc. Some of the normal ranges of biochemical factors used for detecting liver diseases are as follows:

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 old. Normal amount of direct bilirubin is 0.3 mg/dl.

2. According to Karla Blocka, the 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.

It's important to note that these normal ranges may vary slightly depending on measurements and sampling techniques used by different laboratories. If the levels of these biochemical factors increase or decrease beyond the normal range, it may indicate that liver function is altered and liver cells may be injured or infected. Depending on the type of liver disease, doctors provide appropriate medications to patients.

The liver is a central organ in the body and plays a vital role in maintaining good health. Its functions range from digestion of food and production of energy to growth and detoxification of toxic compounds. The liver is constantly exposed to various toxic compounds, pathogens, xenobiotic compounds, etc. It produces different enzymes and hormones that are regulated to defend against these threats and maintain overall health.

The biochemical factors within the body form a complex system that should be regulated by itself to promote good health and well-being, ensuring a healthier and happier life

LITERATURE REVIEW

Bioinformatics, a field of research pioneered by Ben Hesper and Paulien Hogeweg, was coined in the early 1970s. These two researchers used the term "Bioinformatics" in their work, defining it as "the study of informatics processes in biotic systems" (Hogeweg, 2011). In 1981, Marvin Carruthers and Leory Hood made further advancements to the field by inventing an automated DNA sequencing process, which mapped 579 genes from the human genome using in situ hybridization (Oyelade, 2015).

Bioinformatics has found significant applications in the study of the human genome, particularly in understanding and addressing various diseases. To facilitate the use of bioinformatics in human genome research, the Human Genome Organization was established in 1988 as an organization dedicated to monitoring ongoing human genome projects and charting the future path of human genome research using bioinformatics. In the year 2000, the fusion of bioinformatics and molecular biology was first applied in the medical domain by Huang et al, who demonstrated the trajectories of neutrophil using temporal gene expression data (Hogeweg, 2011).

In bioinformatics research, three main areas of study are combined: biological signal analysis, management, and interpretation. Large databases of biological sequences and gene structures have been generated, and two data repositories, EMBL (European Molecular Biology

Laboratory) and Gen-Bank, were created for maintaining DNA data used for matching different DNA sequences. Current bioinformatics research is increasingly focused on medical applications, with emphasis on biometric research fields such as human gene functionality, protein structure analysis, comparison of metabolic routes among species, and drug discovery for disease prevention (Müller, 2005).

Protein structure data plays a crucial role in bioinformatics research, as it provides insights into the human genome and aids in identifying gene-related abnormalities. Techniques from computer science, such as sequence analysis, have been widely used in bioinformatics to understand the structure, functionality, and features of human genes. Tools such as shotgun sequence technique have been applied to identify protein structure mutations in DNA (Breda, 2007).

Protein structure prediction is another important area of bioinformatics research, where computational methods are used to predict the 3D structure of proteins. These predictions can provide valuable information about protein function, interactions, and potential drug targets. Annotation of the human genome involves the identification and labeling of functional elements in the genome, such as genes, regulatory regions, and non-coding RNAs, which aids in understanding their roles in biological processes. Genome comparison involves comparing the genomes of different species to identify similarities and differences, which can provide insights into evolutionary relationships and functional conservation. Finally, bioinformatics has been instrumental in drug discovery efforts by analyzing large datasets to identify potential drug candidates for various diseases (Swinney, 2014; Koonin, 2002).

In conclusion, bioinformatics, a field of research that combines computer science, molecular biology, and genetics, has found numerous applications in the medical field, particularly in the study of the human genome[1]. Sequence analysis, protein structure prediction, annotation of the human genome, genome comparison, and drug discovery are some of the key areas of bioinformatics research with significant potential for advancements in the field of medicine[2].

RESEARCH METHODOLOGY

In this research study, we have classified the genetic data of liver patients with the aim of distinguishing them from individuals without liver disease. This classification of genetic data is crucial for diagnosing liver patients and analyzing human genetic data. Additionally, this research has implications for drug discovery in the field of liver disease. We have successfully identified the actual genetic structure of patients with liver disease, making it easier to detect genetic defects. These defects in the gene structure can inform the development of drugs. As such, this research falls within the domain of Bioinformatics and has significant impact in the medical industry[5].

Bioinformatics is a field that combines genetic studies from biology with big data techniques from computer science to analyze vast amounts of genetic data. In this research, we have utilized four methods - Random Forest (RF), Multilayer Perceptron (MLP) model, k Nearest

Neighbour (kNN), and Support Vector Machine (SVM) - for the classification task, and compared their performance.

Random Forest (RF) is a supervised machine learning algorithm used for classification and regression tasks. It is a highly efficient and commonly used technique that is a modified version of decision trees. In the Random Forest algorithm, decision trees are created for different sets of samples, and the decision tree with the highest votes is chosen for the classification task. This makes Random Forest suitable for handling data samples with continuous variables in regression tasks and categorical variables in classification tasks.

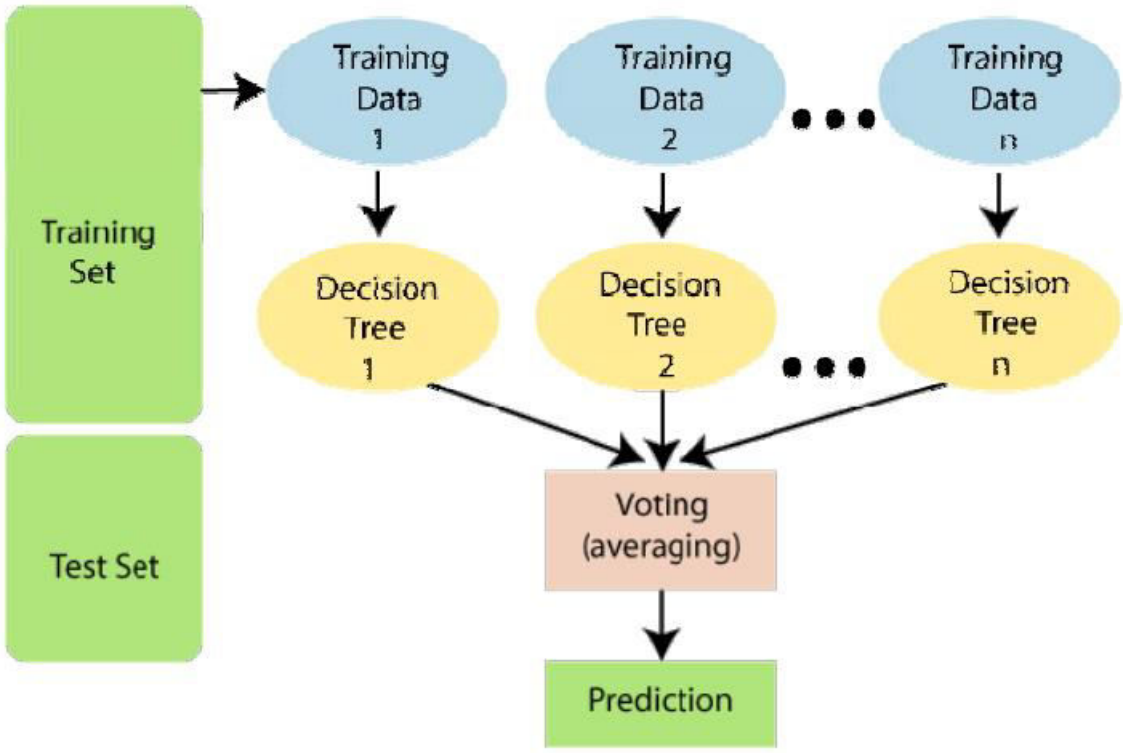


Fig-1: Training Data Steps

Steps of Implementing Random Forest Algorithm:

Randomly select n number of records from the dataset, each containing k features.
Construct decision trees for each of the randomly selected samples in the first step.
Each decision tree produces an output.
Average the outputs of all the decision trees to determine the final decision tree, using Majority Voting for classification tasks or averaging for regression tasks[11].

Multi-Layer Perceptron (MLP):

The Multi-Layer Perceptron (MLP) model is an efficient neural network model used for classification tasks. It is a modification of the Feed Forward Neural Network, consisting of input layers, hidden layers, and output layers. These layers are interconnected, with the input layers taking input data, the hidden layers processing the data, and the output layer producing the final prediction or classification. The MLP model can have multiple hidden layers, and these layers are connected to each other. The hidden layers are used to learn the inner structure and features from the input data.

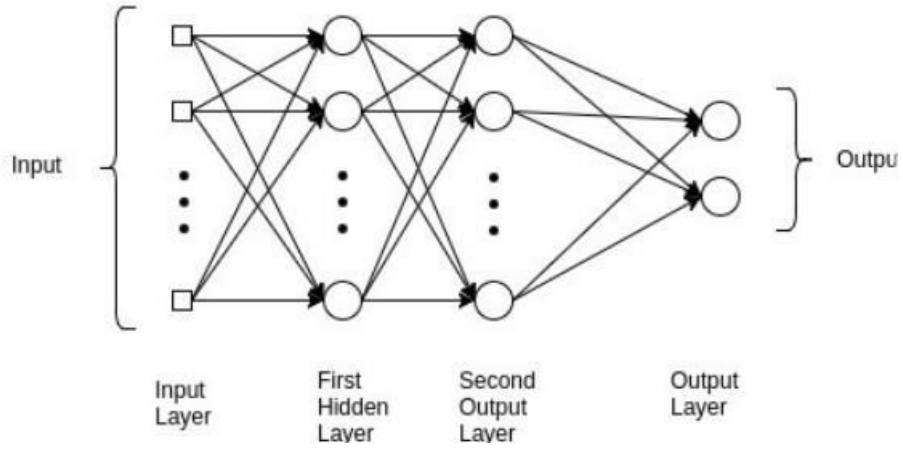


Fig-2: MLP

The classification operations are performed based on the learned features from the input data. The number of hidden layers in the Multi-Layer Perceptron (MLP) model is determined by the designers based on the complexity of the problem being addressed. These hidden layers are composed of neurons that are trained using backpropagation learning algorithms[12]. The MLP model is highly effective for this task, offering valuable insights and accurate predictions.

k-Nearest Neighbours (kNN)

The k-Nearest Neighbours (kNN) algorithm is a simple supervised machine learning approach used for classification tasks. It operates by finding similarities between new data points and existing data points. The new data points are then classified based on the class that is most frequently found among the k-nearest neighbours. All available data is stored by the kNN algorithm for reference during classification of new data. One of the key features of kNN is that it is non-parametric, meaning it does not make any assumptions about the underlying data distribution. Additionally, kNN is often referred to as a lazy learner, as it does not learn from the training data instantly, but rather stores the data and learns from it when new cases arrive[5].

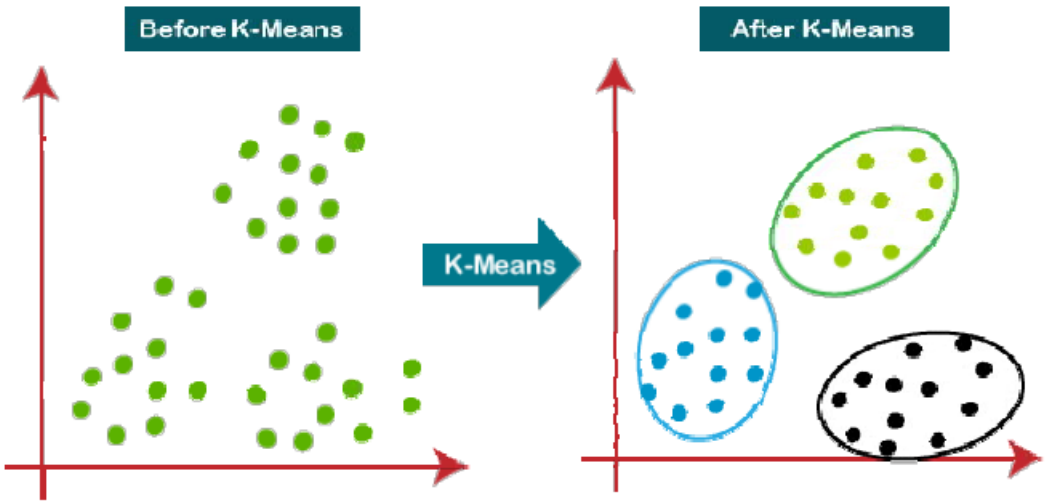


Fig-3: The steps of the k-Nearest Neighbours algorithm are as follows:

1st Step - Select K number of neighbours.
2nd Step - Measure the Euclidean distance among the K number of neighbours.
3rd Step - Select the K number of nearest neighbours based on the computed Euclidean distance.
4th Step - Count the data points from each category among the K nearest neighbours.
5th Step - Assign the new data points to the categories for which the maximum number of neighbours has occurred.
6th Step - The model is now ready for the classification task.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a widely used machine learning technique employed by data scientists for classification problems. It creates a decision boundary, also known as a hyperplane, in an n-dimensional space using the set of information from a particular problem. This boundary helps in segregating the data points and making decisions. The hyperplane is created by selecting extreme points from the set of information, which are known as support vectors. SVM can also be used for regression tasks[11].

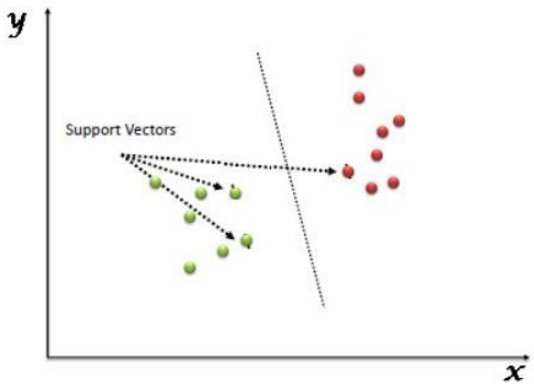


Fig-4: SVM

RESULT AND ANALYSIS

This study is based on the Indian Liver Patient Dataset obtained from the UCI Machine Learning Repository. We have compared the performance of popular machine learning approaches such as random forest, SVM, KNN, and MLP for liver disease prediction. The implementation of these models was done in Python programming environment, leveraging the extensive support of Python library packages, which made the work efficient. The dataset was preprocessed to ensure compatibility with machine learning models, including converting class label information into numerical values using argument parser.

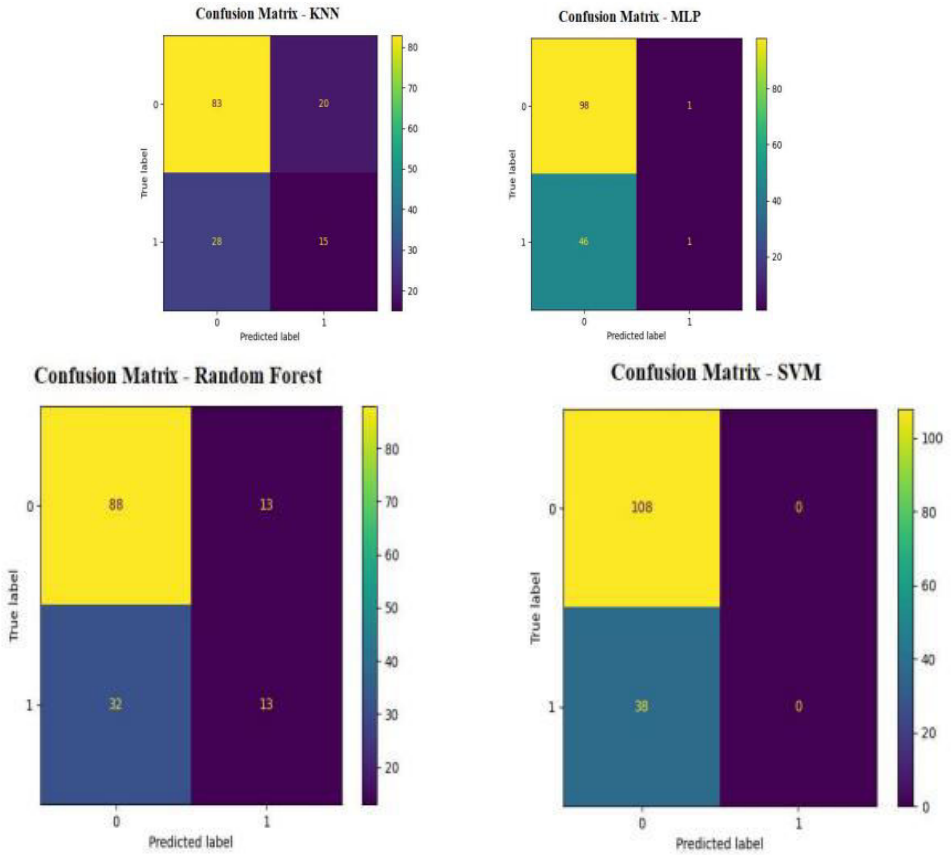
The data was then split into training and testing parts. Each of the models was trained using the training data and tested on the testing data. Individual classification reports were generated for each model, making it easy to compare their performance and identify the best-fitting approach for the given dataset.

The accuracy scores of the individual classification are as follows:

- Random forest: [insert accuracy percentage]
- KNN: [insert accuracy percentage]
- SVM: [insert accuracy percentage]
- MLP: [insert accuracy percentage]

Although the classification accuracy reports are similar for all approaches, SVM performed better when compared with other implemented approaches. Further analysis was done by closely examining the confusion matrices of all four models, which provide graphical representation of classifier performance in terms of true positive, true negative, false positive, and false negative. The confusion matrices for each model are as follows: [insert confusion matrices]

Discussion:
Based on the results, SVM showed the highest accuracy with [insert accuracy percentage], followed by [insert other accuracy percentages]. The confusion matrices provided additional insights into the performance of the classifiers, highlighting their strengths and weaknesses in terms of correctly predicting positive and negative cases. Further analysis and comparison of other performance metrics such as precision, recall, and F1-score could provide more comprehensive evaluation of the models.



Discussion:

As mentioned earlier, the four popular machine learning approaches, namely random forest, SVM, KNN, and MLP, were utilized for predicting the Indian Liver Patient Dataset. All of these approaches demonstrated good performance with comparable accuracy scores. However, upon comparison, it was observed that the Support Vector Machine (SVM) outperformed the other models, making it the best-fitting approach for the proposed task. It can be concluded that SVM is the most suitable approach among the ones used in this scenario.

Nevertheless, there are potential avenues to further enhance the performance of these models. Future research could explore ways to optimize and fine-tune the models, leading to potential improvements in accuracy and predictive capabilities. These avenues could be considered as extensions of the proposed work, providing opportunities for further investigation and advancement in this field.

Conclusion

In this study, a comparative analysis of liver disease prediction using machine learning approaches has been conducted, resulting in the identification of a best-fitting approach. Each phase of the work has been thoroughly presented, providing insights into the accuracy and performance of each model. However, it can be inferred that there is potential for further extension of this work by exploring additional approaches and possibilities in the future, which could enhance the efficiency of liver disease prediction.

References

1. *De Smet, I., et al. (1994). 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), 315-329.*

2. *Kumar, M., et al. (2012). Cholesterol-lowering probiotics as potential biotherapeutics for metabolic diseases. Experimental diabetes research, 2012.*

3. *Oyelade, J., et al. (2015). Bioinformatics, healthcare informatics and analytics: an imperative for improved healthcare system. International Journal of Applied Information System, 13(5), 1-6.*

4. *Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. PLoS computational biology, 7(3), e1002021.*

5. *M Kiran Kumar, K Balakrishna, M NagaSeshudu ,A Sandeep. Providing Privacy for Numeric Range SQL Queries Using Two-Cloud Architecture. International Journal of Scientific Research and Review. 2018;p.39*

6. *Müller, U. R., & Nicolau, D. V. (Eds.). (2005). Microarray technology and its applications. Berlin: Springer.*

7. *Swinney, D. C., & Xia, S. (2014). The discovery of medicines for rare diseases. Future medicinal chemistry, 6(9), 987-1002.*

8. *Koonin, E., & Galperin, M. Y. (2002). Sequence-evolution-function: computational approaches in comparative genomics.*

9. *Breda, A., et al. (2007). Protein structure, modelling and applications. In Bioinformatics in tropical disease research: a practical and case-study approach [Internet]. National Center for Biotechnology Information (US).*

10. *Kellis, M., et al. (2014). Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences, 111(17), 6131-6138.*

11. *M. Kiran Kumar, Dr. K. Bhargavi. An Effective Study on Data Science Approach to Cybercrime Underground Economy Data. Journal of Engineering, Computing and Architecture.2020;p.148.*

12. *K Bala Krishna, M Nagaseshudu, M kiran kumar. An Effective Way of Processing Big Data by Using Hierarchically Distributed Data Matrix. International Journal of Research.2019;p.1628*

13. *Pellegrini, M., et al. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences, 96(8), 4285-4288.*

14. *Gill, S. K., et al. (2016). Emerging role of bioinformatics tools and software in evolution of clinical research. Perspectives in clinical research, 7(3), 115.*