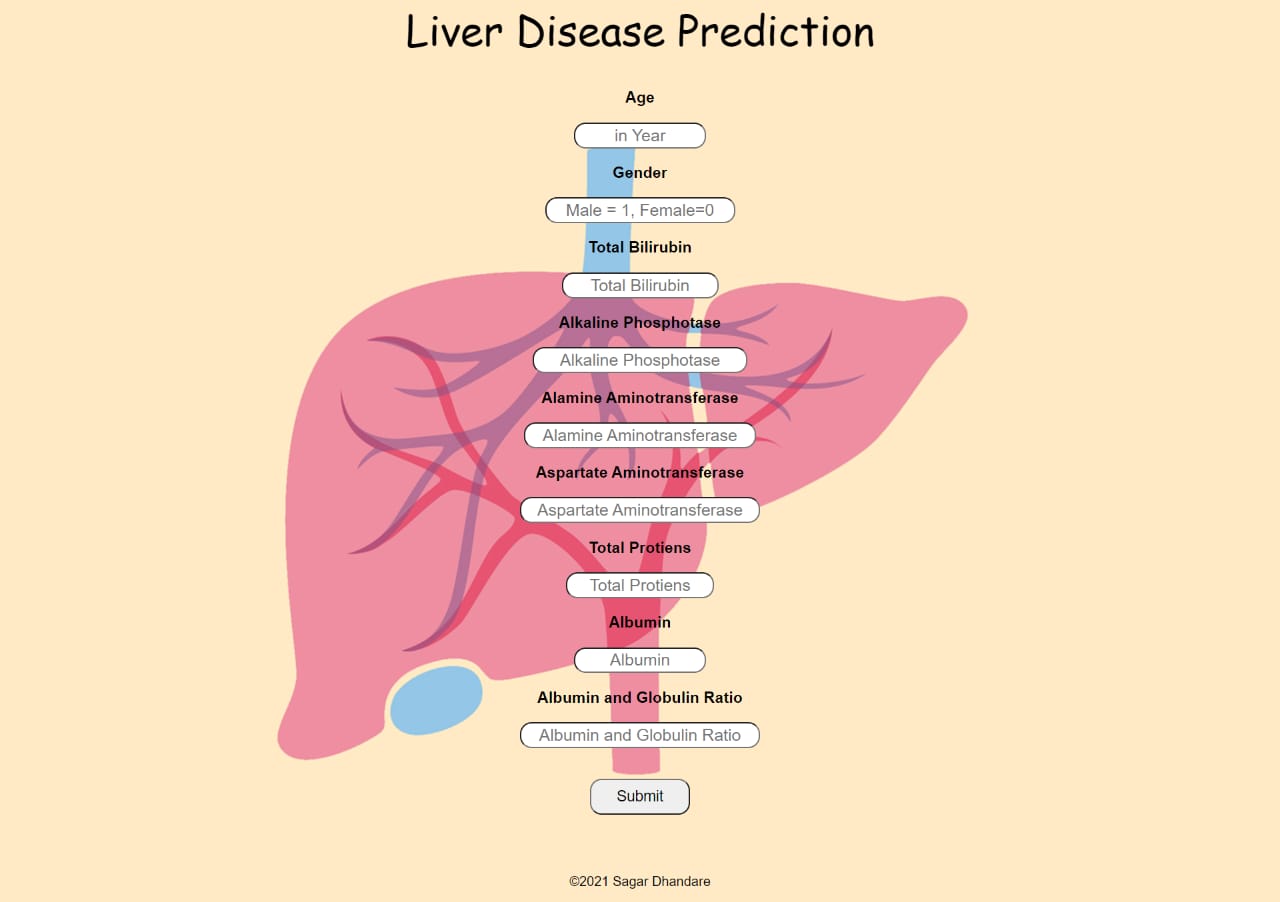
Liver Disease Prediction using Machine Learning Algorithms

Abstract :

Early diagnosis is essential to improving patient outcomes and decreasing the costs of healthcare. Liver disease is a big worldwide health problem, and early detection plays a critical role in both of these areas. A data-driven technique that makes use of supervised learning algorithms is presented in this research report as a method for estimating the likelihood of developing liver disease. For the objectives of research, the study makes use of a dataset collected from www.kaggle.com that contains information on the demographics, lifestyle, and medical history of 416 patients who were treated at a hospital in India. In order to construct accurate prediction models for the risk of liver illness, we use three different supervised learning techniques. These are decision trees, random forests, and logistic regressions. Accuracy, specificity, sensitivity, and the area under the receiver operating characteristic (ROC) curve are the metrics that are used in order to assess the performance of these models. According to the findings, the hybrid method surpasses the other two algorithms by obtaining higher levels of accuracy (75.7 percent), sensitivity (71.4 percent), and specificity (77.2 percent), as well as a higher area under the ROC curve (0.80). This work demonstrates the potential of supervised learning algorithms in forecasting the risk of liver disease using patient data, especially in areas where there is a limited availability of resources



INTRODUCTION:

It is becoming clear that the digital technological revolution has the potential to be a really disruptive breakthrough [1]. The rise of cutting-edge medical technologies is marked by developments in nanotechnology and genetics [2] Prognosis, therapy, and healthcare monitoring in the digital age have great promise, and there seems to be no end to the possible applications of this potential. Every time a medical procedure is carried out, a huge quantity of data is processed and dealt with on a frequent basis [3]. These data sets may be inferential, referential, or raw enough to draw conclusions about further valuable medical data sets. There is a wide variety of origins and purposes for this data. They have applications in illness prediction, diagnosis, and therapy [4]. The same research might be studied to speed up similar initiatives. It could be useful for making statistical conclusions about future trends that may aid the process as a whole. Classification methods are widely used in data mining for illness prediction and medical diagnosis

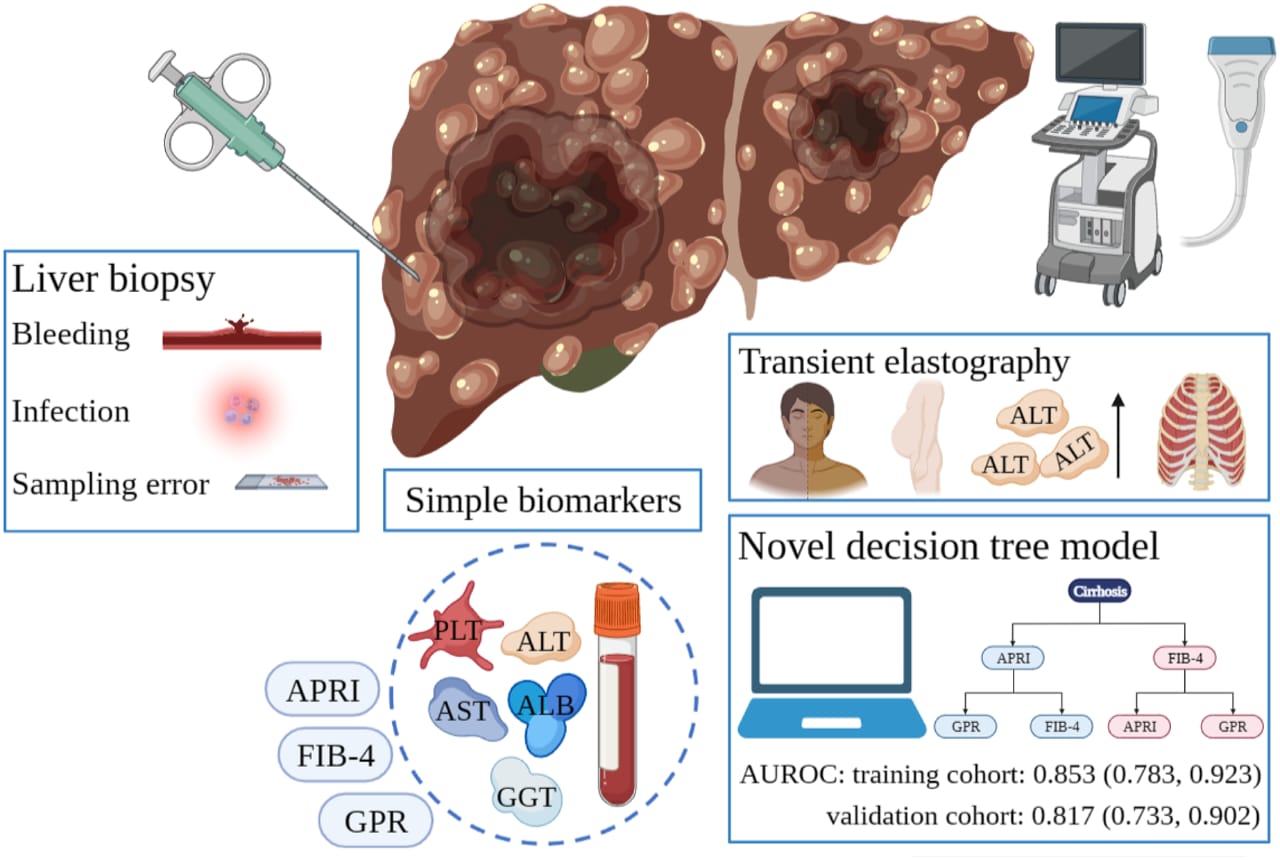
The liver is a crucial organ that sits at the very top of the digestive system. It's important for things like breaking down food, making energy, eliminating toxins, producing antibodies, and storing nutrients [6]. The liver is the body's second largest internal organ, and it's responsible for a wide variety of functions, including digestion, clotting blood, and producing bile. About three pounds is its estimated weight [7]. The liver is responsible for a wide variety of tasks, including digestion, metabolism, immunity, and nutrition storage. For these reasons, the liver is a vital organ; without it, cell death from a lack of oxygen and nutrients would occur rapidly throughout the body

● Analyze the efficacy of the random forest, decision tree, and the logistic regression supervised learning algorithms in determining the likelihood of liver disease.

● Evaluate the prediction models created by the three algorithms with respect to accuracy, sensitivity, specificity, and area under the ROC curve.

● Based on the results of the assessment, choose the algorithm that produces the best accurate predictions of risk for liver disease.

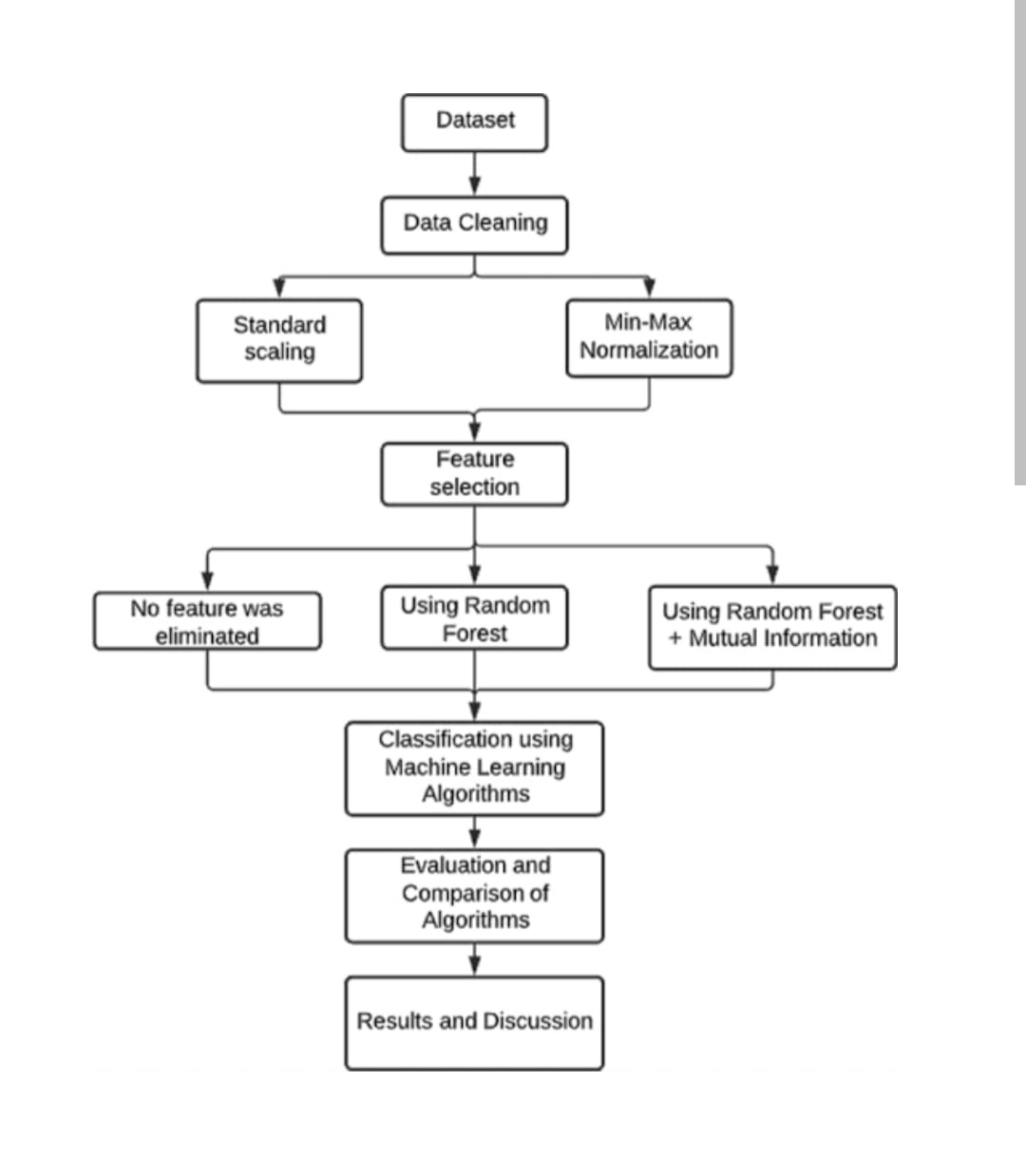
● Evaluate supervised learning algorithms for their potential use in forecasting the risk of liver disease in low-resource areas



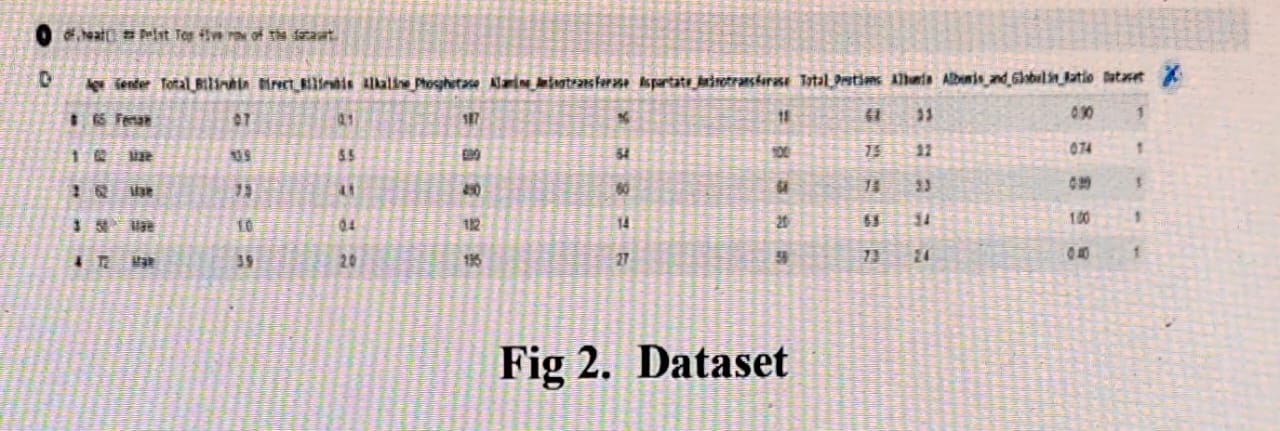
2. LITERATURE REVIEW:

Rahman, Shamrat, Tasnim, Roy, and Hossain, [13]“A comparative study on liver disease prediction using supervised machine learning algorithms” Many people throughout the globe suffer from chronic liver disease, making it a major international mortality. It's caused by things like being overweight, having an undetected hepatitis infection, and drinking excessively. Serious complications from this syndrome include abnormal nerve function, hepatic encephalopathy hemoptysis or vomiting blood, liver failure, renal failure, jaundice. Chronic liver disease is difficult and costly to diagnose. The primary goal of this study is to apply six different supervised machine learning classifiers to develop a reliable method of identifying people who suffer from chronic liver disease. With the hope of lowering the astronomical expenses involved with diagnosing chronic liver disease, this research compares the efficacy of many machine learning algorithms for making such a prediction. In this study, researchers used six different machine learning algorithms.

3. MATERIALS AND METHODS:

Including class balancing and ranking features in the balanced data, we will describe the dataset we utilized and the primary stages of the selected strategy for forecasting the risk of liver sickness. Finally, we detail the ML models that were used in order to make sense of the experimental results. 3. MATERIALS AND METHODS Including class balancing and ranking features in the balanced data, we will describe the dataset we utilized and the primary stages of the selected strategy for forecasting the risk of liver sickness. Finally, we detail the ML models that were used in order to make sense of the experimental results. This dataset seems to have liver detection as its major emphasis, suggesting that its goal is to identify and label occurrences as having or not having features connected to the liver. There are a total of 11 columns in the dataset, all of which indicate different characteristics that may be useful for liver identification. Age, gender, weight, height, body mass index (BMI), liver enzyme levels, and other medical or physiological parameters are all examples of characteristics that are associated with liver health or problems. These features will be used to build a machine learning model for liver detection, and each column gives unique information or measurements about the samples. Researchers and data scientists might benefit greatly from this dataset by examining the connections between the various characteristics and the existence or absence of liver-related disorders

This dataset seems to have liver detection as its major emphasis, suggesting that its goal is to identify and label occurrences as having or not having features connected to the liver. There are a total of 11 columns in the dataset, all of which indicate different characteristics that may be useful for liver identification. Age, gender, weight, height, body mass index (BMI), liver enzyme levels, and other medical or physiological parameters are all examples of characteristics that are associated with liver health or problems. These features will be used to build a machine learning model for liver detection, and each column gives unique information or measurements about the samples. Researchers and data scientists might benefit greatly from this dataset by examining the connections between the various characteristics and the absence of liver-related disorders.



Researchers and data scientists might benefit greatly from this dataset by examining the connections between the various characteristics and the existence or absence of liver-related disorders. In addition, researchers may use this data to train and test machine learning algorithms and prediction models for effectively categorizing new occurrences based on their liver features [17]. As a result, the dataset is a useful tool for training and assessing such algorithms, which might eventually lead to better patient outcomes via more refined liver detection methods

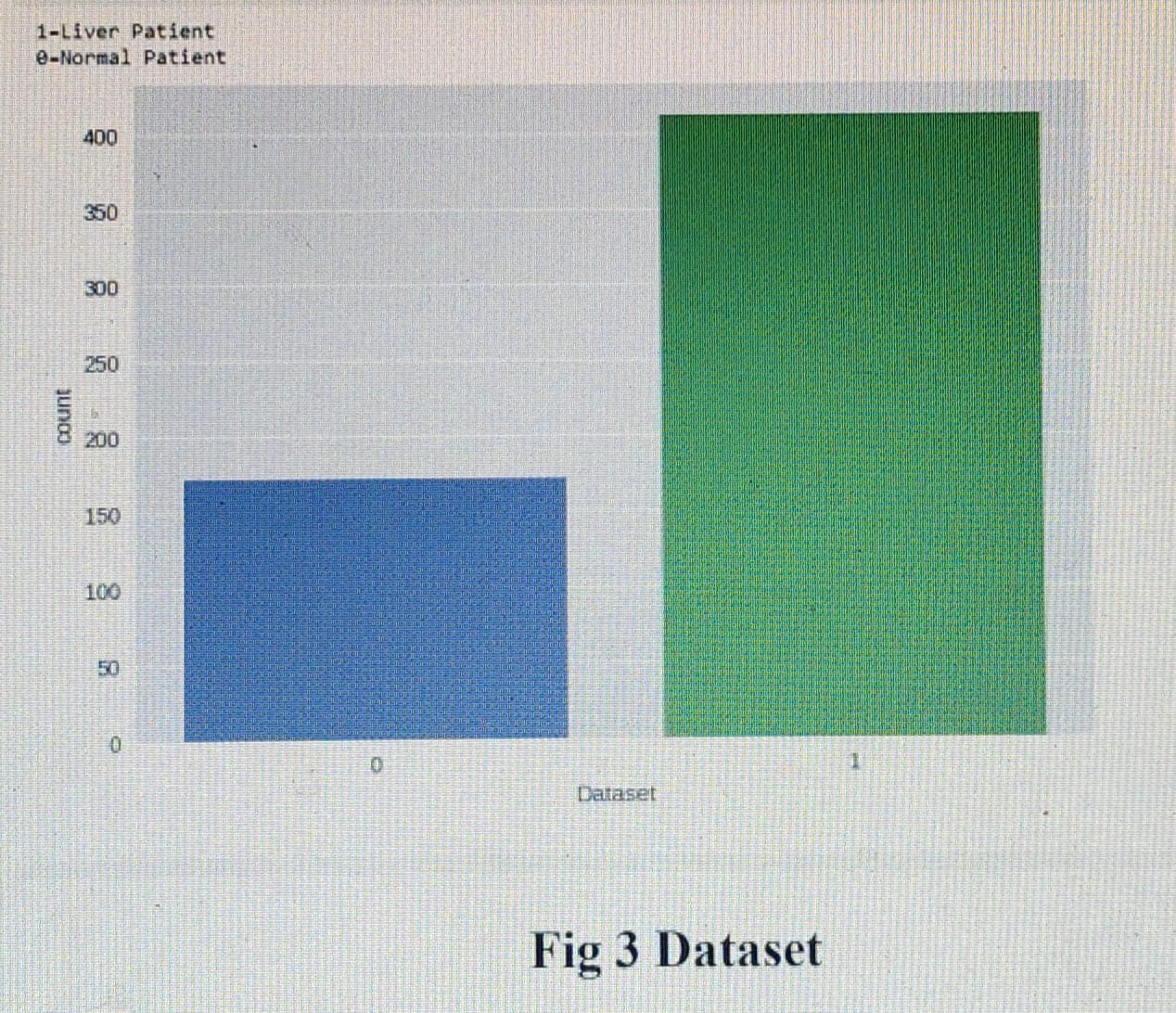


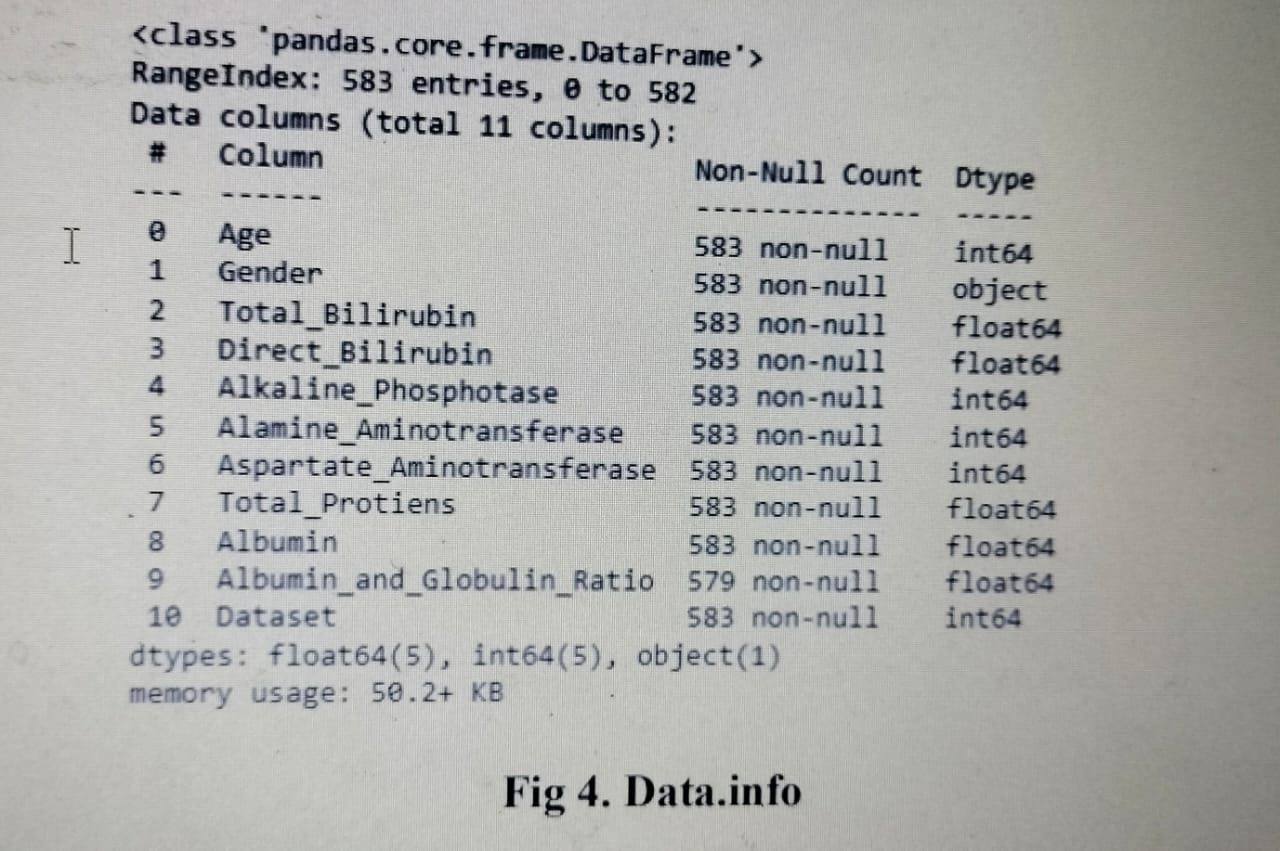
Figure 3 shows that the liver detection dataset has 583 rows, or instances. Each instance is a unique record. There are a total of 583 patient records, 171 of which show that the patient did not have liver disease and 412 of which had. This pattern indicates that there is an inequity between the percentage of patients with and without liver disease in the dataset. In machine learning and statistical analysis, imbalanced datasets may be problematic because models may overfit to the dominant class and underfit to the minority class. Understanding the class distribution of the dataset is crucial for developing efficient modeling approaches and evaluation metrics [18]. Oversampling the minority class, under sampling the majority class, or using techniques designed for imbalanced data are all effective ways to deal with class imbalance and ensure equitable representation and improved model performance.

Liver Disease Risk Prediction

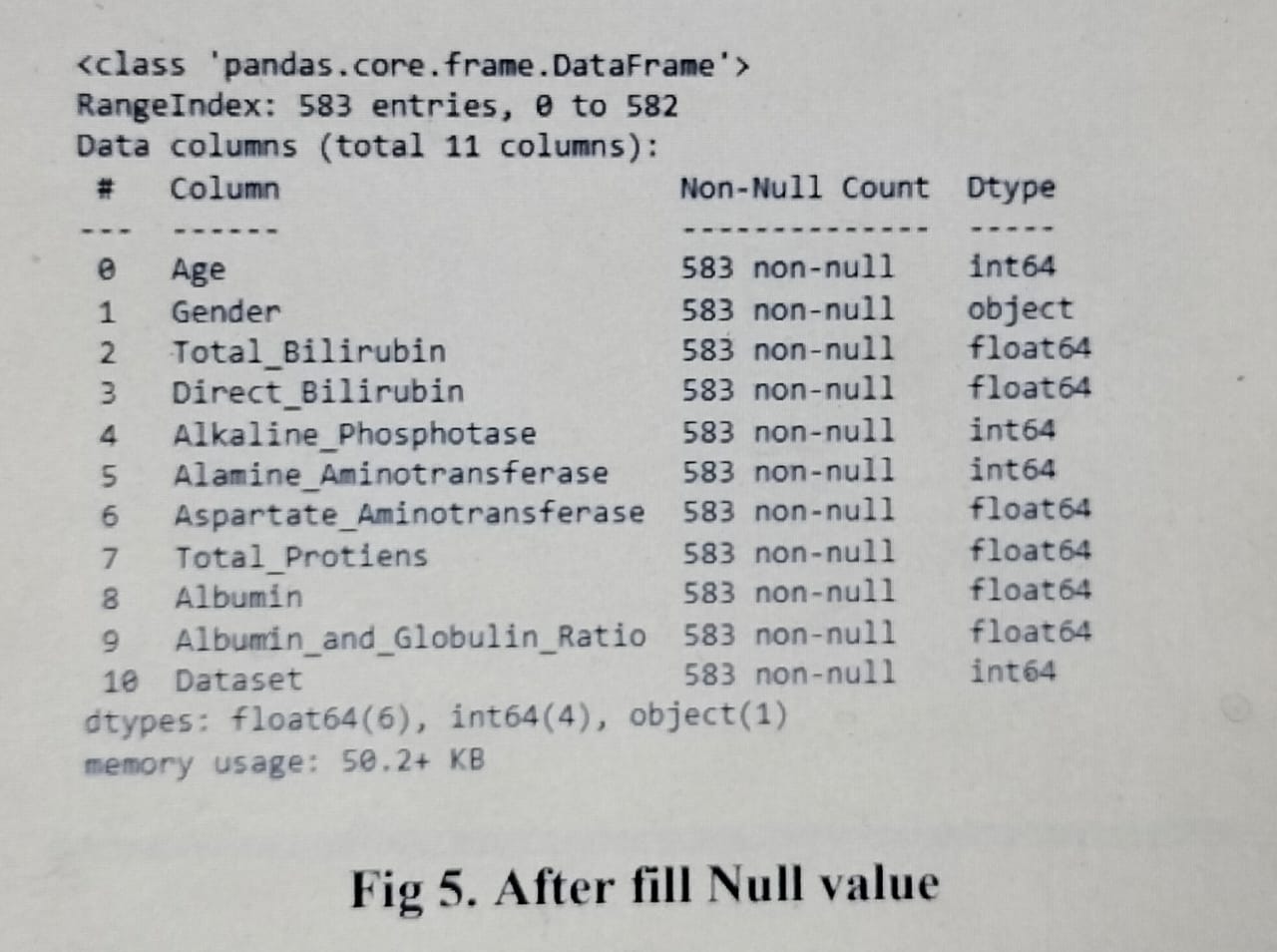
Liver Disease Risk Prediction These days, physicians and health care providers use machine learning algorithms to create accurate tools for determining the likelihood of illness incidence given a set of risk variables. Here, we frame the issue of predicting the long-term risk of liver disease as a classification problem with two classes: c = "Liver-Disease" (LD) and c = "Non-Liver-Disease" (Non-LD). The probability of liver disease is predicted by the trained ML models, which can classify a new unclassified instance as LD or Non-LD depending on the values of the input characteristics..

Data Preprocessing

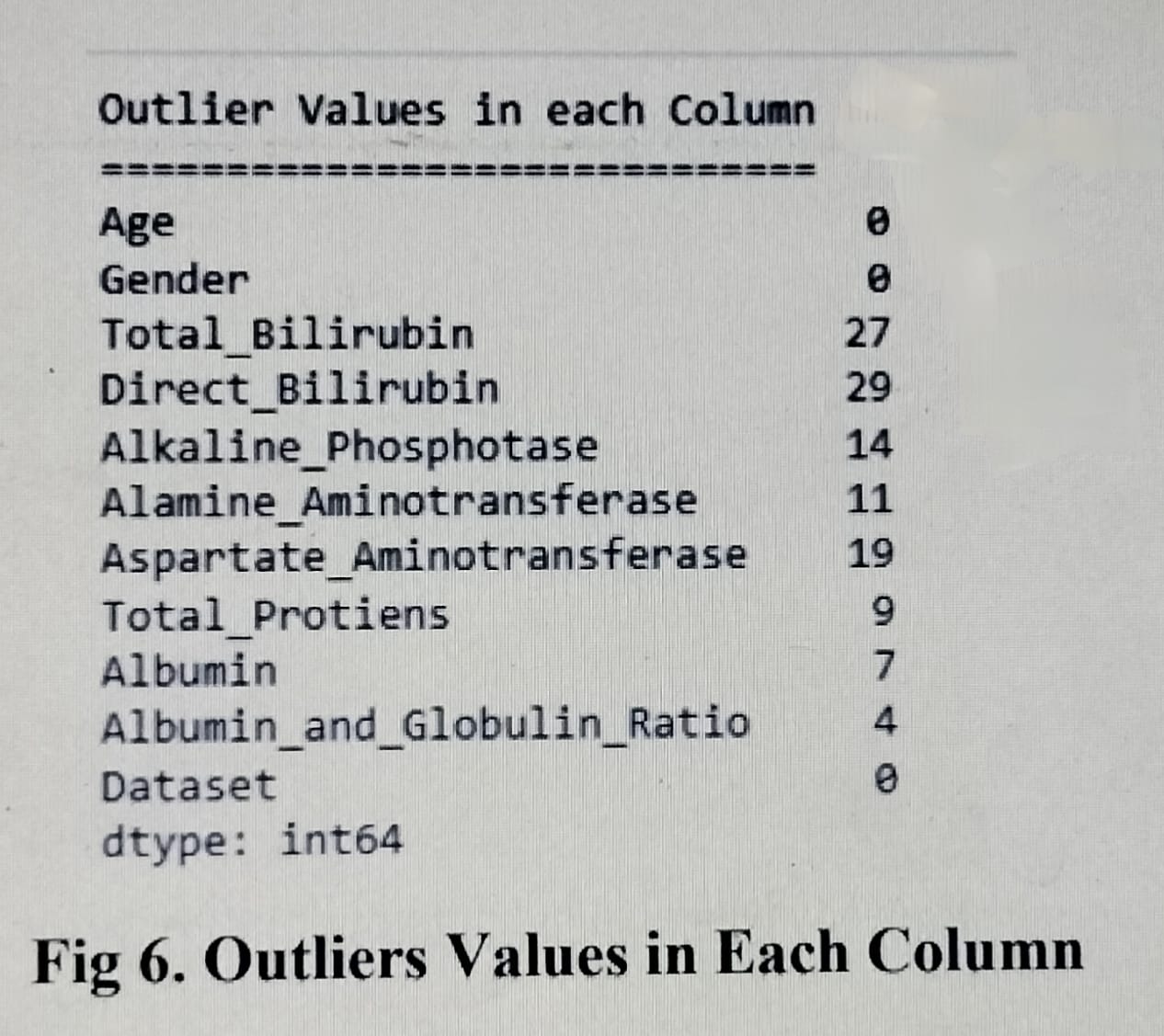
The dataset is then subjected to preprocessing, which involves cleaning the data to remove inconsistencies and handling missing values appropriately. Uneven distribution of LD and Non-LD occurrences in the dataset may interfere with their proper identification. Here, we use an oversampling technique, which generates synthetic data on the underrepresented group by means of a 5-NN classifier. Non-LD cases are oversampled to ensure that the two populations are evenly represented. The dataset is now evenly split between LD and NonLD cases, totaling 416. Finally, we give the information of the dataset in fig.1

n

The column "Albumin and Globulin Ratio" in the liver detection dataset is missing four values as shown in figure 4. One popular method for dealing with such gaps is to substitute the column's mean for the missing value. The mean is found by adding together all the non-missing values and dividing by the total number of such values. By using the column's mean to fill in missing values, we can preserve the column's statistical features and reduce the effect of missing data on our analyses and models. We may maintain the general features of the data by utilizing the mean value to estimate what the missing values could be. This method relies on the assumption that missing values occur at random and that the current data faithfully represents the missing ones. While the median imputation approach is simple and extensively used, there are several methods that may be used instead. To forecast missing values based on other pertinent factors, for instance, regression imputation or more complex machine learning methods might be used. To keep the dataset whole and usable for analysis and modeling purposes with regards to liver detection, we imputed the missing values in the "Albumin and Globulin Ratio" column with the mean. But it's still smart to check how missing data imputation affects the final analysis findings and switch to a different approach if it works better for your data and goals



After missing data was filled in, the final results were shown in Figure 5.

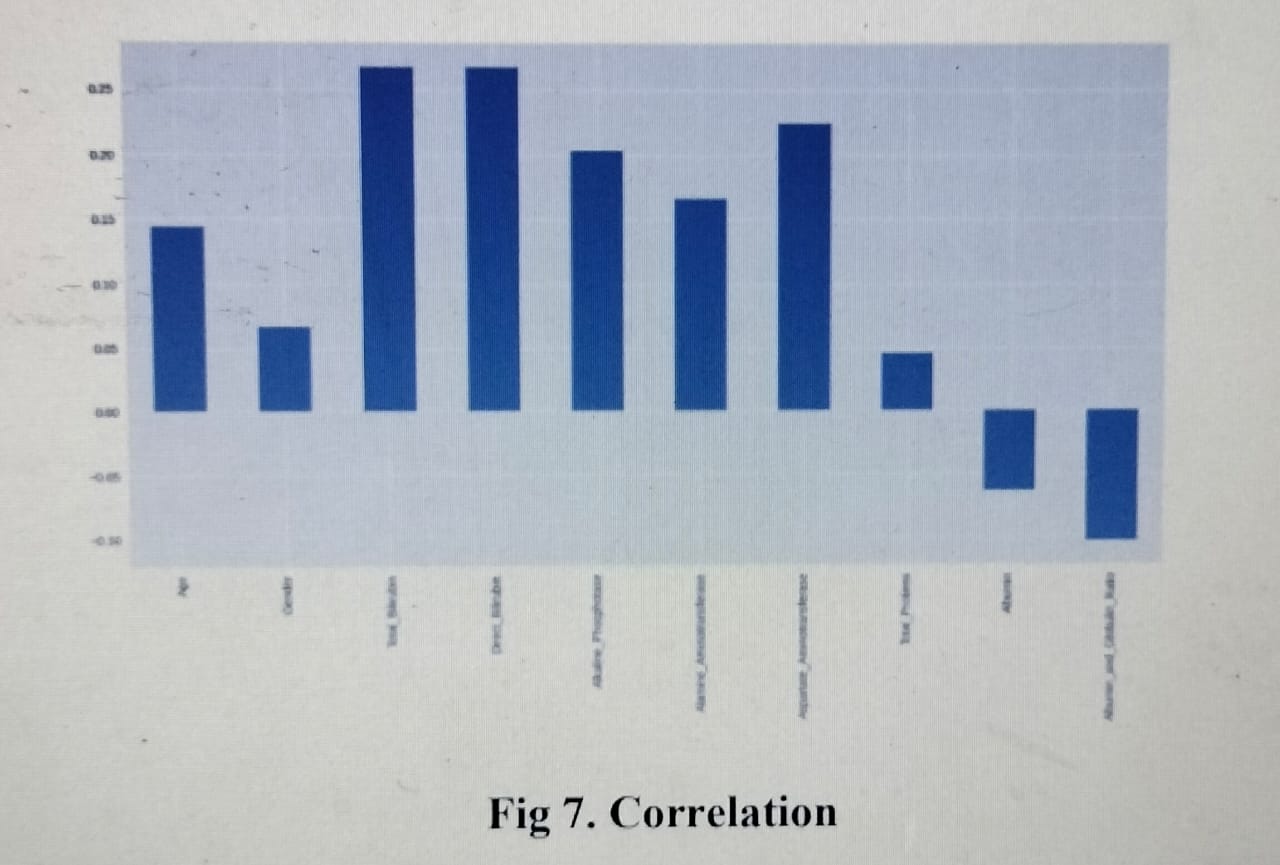


There are 11 columns in the liver detection dataset; eight of them contain outliers. The results of statistical studies or models may be skewed by outliers as shown in figure 6, or data points that differ dramatically from the norm. Capping is a frequent method used to deal with these extreme cases. Capping, also known as winsorization, is the process of substituting highly improbable numbers with more moderate ones that fall within an established range. The 5th and 95th percentiles are common ways to describe this range. Outlying values are substituted with their respective percentile values when they are found to be outside this range. However, care must be used while dealing with extreme cases. However, before deciding to cap outliers, it is important to think about the data and the situation in which the analysis will be used. Blindly eliminating or manipulating outliers might lead to potentially biased or misleading conclusions since they can occasionally hold significant information or indicate true oddities in the data. It is best practice to review the data, get familiarity with the domain, and if feasible, contact domain experts before implementing any outlier management strategy, including capping. In order to make an informed choice on whether or not to utilize capping or another outlier management approach, one must first consider the dataset, the analytic goals, and the possible influence on the final results or conclusions

Feature selection

Then we apply feature selection techniques to identify the most relevant features for predicting liver disease. We select the features that have high predictive power and are meaningful from a medical perspective. To gain insights into the selected features, data visualization techniques are employed, allowing for the visualization of distributions, correlations, and trends. The dataset is then split into training and testing sets to facilitate model development and evaluation as seen in figure.

As can be seen in Fig. 7, three columns—age, gender, and total protein—were removed from the liver detection dataset after correlation analysis was used to determine which ones were most closely connected to the target variable. This choice was made since these columns were shown to provide no substantial benefits in terms of accuracy and had weaker correlations with the target variable.



1. Age: Age is often seen as a crucial factor in a number of studies. However, it was shown that age had no role in the reliability of the liver detection predictions in this dataset. Based on the results of the correlation study, it seems that age is only a moderate predictor of the outcome variable (liver disease). As a result, it was disregarded in favour of other, more important characteristics in the dataset.

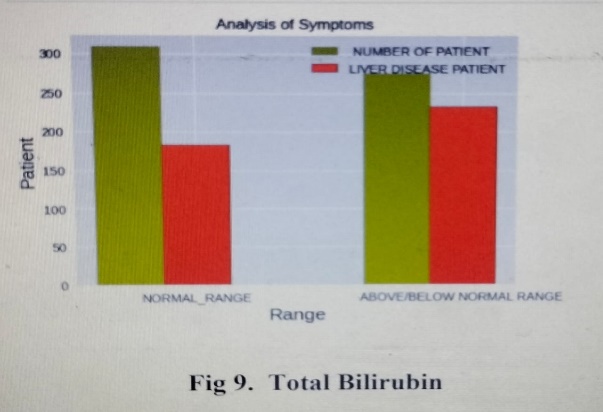
2. Gender: Gender was also removed since it did not provide any significant improvement in accuracy when predicting the outcome variable. The study of correlations showed (Figure 7) that the link between gender and liver illness was minimal at best. It's possible that gender alone doesn't give enough information to reliably distinguish people with liver disease from those without, reducing its utility for the prediction.

3. Total protein: The decision to drop the total\_protein column was based on its relatively lower correlation with the target variable. A weak correlation suggests that the variation in total\_protein values does not strongly align with the presence or absence of liver disease. Although total\_protein may still provide some information, it was considered less influential compared to other attributes in accurately predicting the target variable.

The columns that show better relationships with the target variable are highlighted after removing demographic information such as age, gender, and total protein from the dataset. The efficacy and precision of liver detection models are more likely to be affected by these columns. Researchers and data scientists may use correlation analysis to determine which aspects are most useful and then prioritize those features to increase the accuracy and efficiency of future modeling and prediction work.

Applying hybrid model:

Researchers construct hybrid models, which combine diverse machine learning algorithms and approaches, to enhance the precision of their predictions. The quality of the models is evaluated using performance metrics like accuracy and F1 score. Statistical analysis is used to establish the significance and trustworthiness of the obtained results when comparing models or assessing the effect of particular factors on prediction. The developed models are then used to make predictions on fresh data, and the results are compared to the testing dataset to see how well they performed. The efficacy of the models is also evaluated by contrasting their findings with those of the primary study or with those of previous studies.

Using the liver disease detection confusion matrix. The performance of a classification model may be shown using the confusion matrix. It is really useful for checking how well the model predicts and how well it performs overall using measures like precision, recall, and F1-score. Applying the constructed models to bigger datasets or actual situations is the last stage in assessing their scalability. As our datasets grow, this statistic will help us determine whether or not our models are still computationally efficient. Researchers may use this strategy to create reliable models for predicting liver disease, and the models' efficacy can be evaluated including accuracy, F1 score, statics 

As can be seen in the green area, the total number of patients whose Direct Bilirubin levels are within the normal range is 30. Figure 10 shows 175 people have liver disease (indicated in red)

Compassion with base paper:

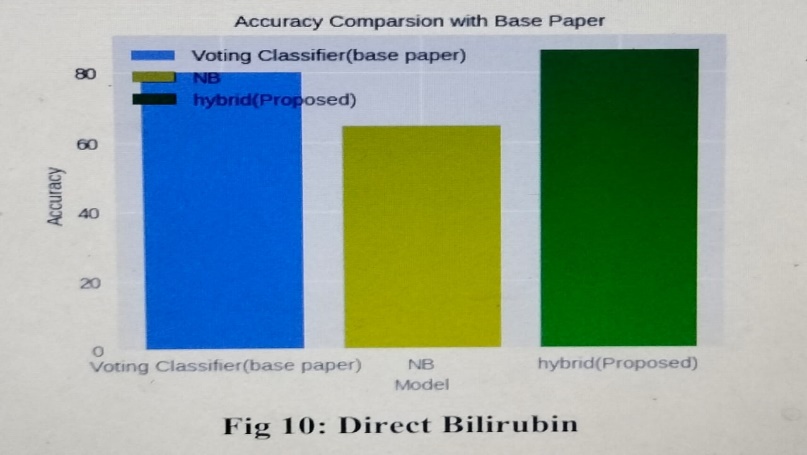
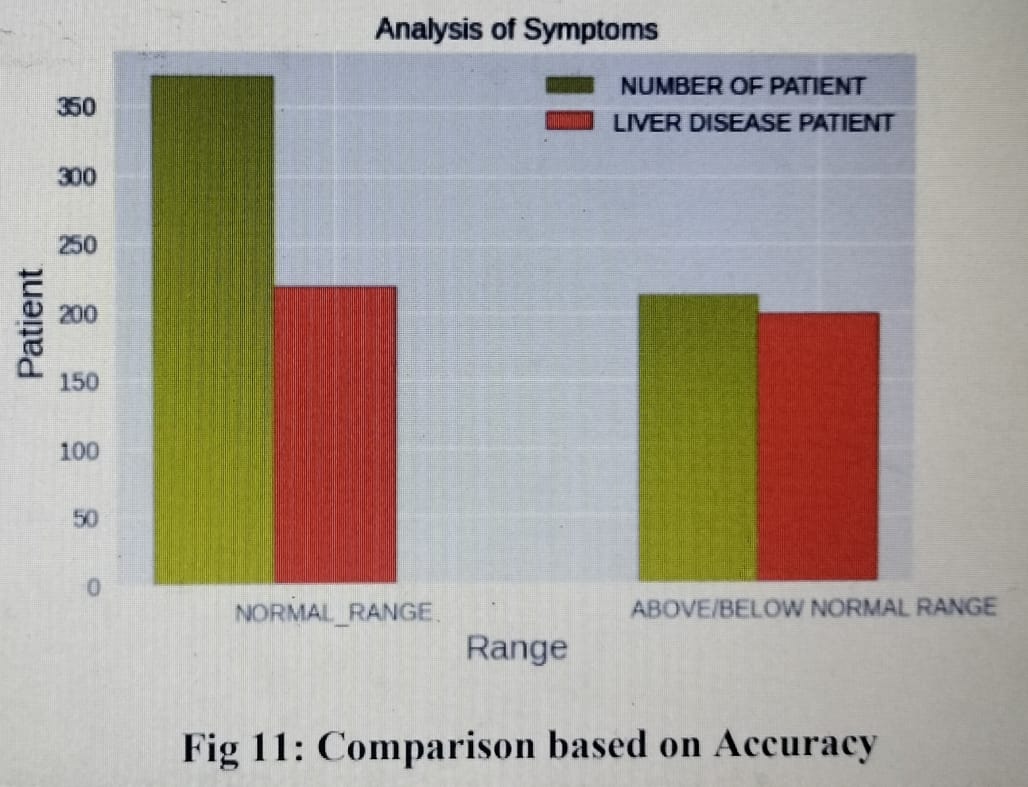
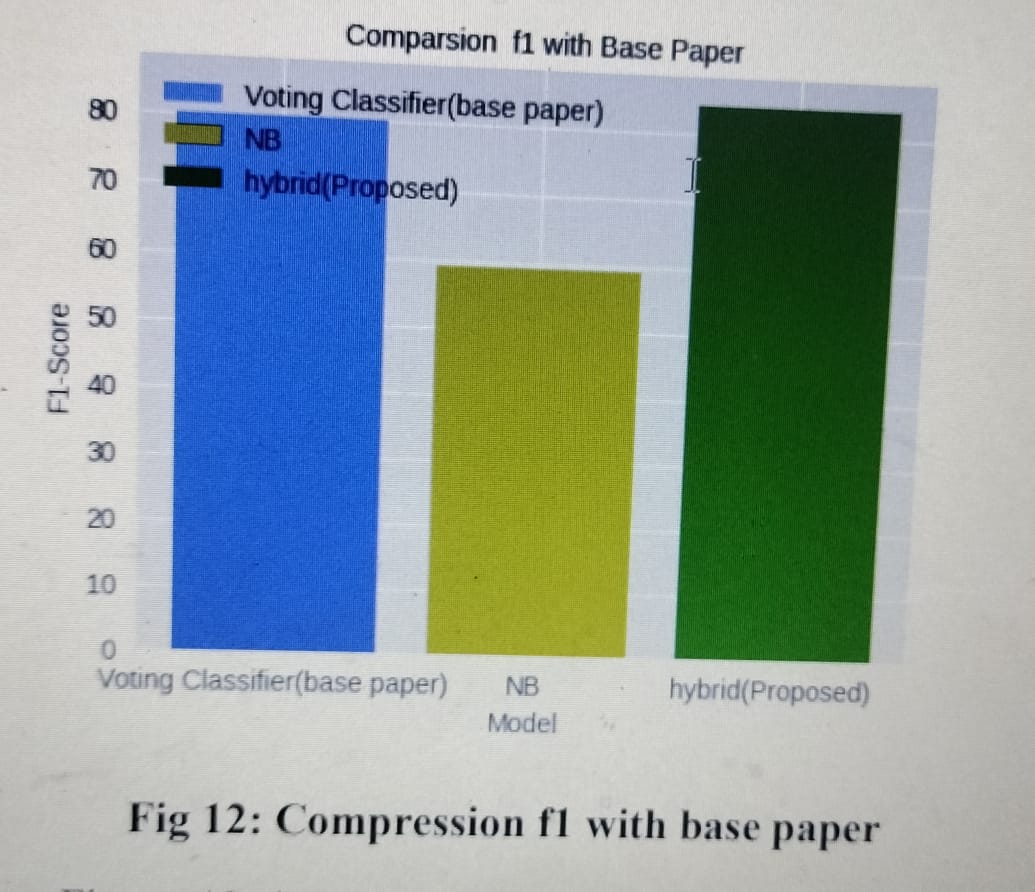


Figure 11 shows that when three models are compared, a hybrid proposed model performs somewhere between the main paper model (80% accuracy) and the Naive Bayes model (65% accuracy) (86 % accuracy).

4. RESULTS AND DISCUSSION: This work demonstrates the promise of supervised learning algorithms, and in particular the random forest method, for estimating the likelihood of liver illness from patient records. When it comes to early identification and prevention of liver disease, these algorithms may be invaluable tools for healthcare workers, particularly in resource-limited situations. In order to effectively spend healthcare resources to stop the course of diseases and improve patient outcomes, it is necessary to precisely identify persons at high risk. In conclusion, our research shows that supervised learning algorithms, and the random forest algorithm in particular, can accurately forecast the risk of liver disease from patient data. The findings highlight the potential public health uses of data-driven methods to illness risk prediction.



As can be seen in the green area, the total number of patients whose total bilirubin levels are within the normal range is 374. Figure 9 shows 271 people have liver disease (indicated in red). 58% of individuals in the database have liver illness even if their total bilirubin levels are within the normal range, according to the statistics. There are a total of 209 patients in the data set whose total bilirubin levels are either high or low relative to the normal range. When the total bilirubin result is either high or too low, 93% of the 209,185 individuals with liver disease fall into the latter category Naive Bayes model and the basic paper model (80% accuracy) (65 percent accuracy). Overall, the Hybrid Proposed Model outperforms both the original paper model and the Naive Bayes model, with an accuracy of 86% achieved. This demonstrates that the suggested model's hybrid method has improved the accuracy of liver disease diagnosis, since it outperforms the other two models in terms of predictive performance .



From Figure 12, it appears that comparing the F1-scores of three different models: a base paper model (80% F1-score), a Naive Bayes model (58% F1-score), and a Hybrid Proposed Model (F1-score above 86%).

1. Base Paper Model (80% F1-Score): It has an F1- score of 80%. The F1-score is a statistic that takes into account both the accuracy and the reliability of a model. The model strikes a fair mix between accuracy and recall on the liver disease detection task, as shown by an F1-score of 80%
2. Naive Bayes Model (58% F1-Score): The Naive Bayes model is another classification algorithm that assumes independence between features. In this case, the Naive Bayes model achieved an F1-score of 58%. This suggests that the model's performance in terms of both precision and recall is lower compared to the base paper model. It may indicate that the Naive Bayes model struggles with certain aspects of the liver disease detection task
3. Hybrid Proposed Model (F1-Score above 86%): The Hybrid Proposed Model is a different model that outperforms both the base paper model and the Naive Bayes model. It achieves an F1-score above 86%, indicating a high level of precision and recall for liver disease detection. The hybrid approach used in the proposed model incorporates various techniques, approaches, or features to improve the overall performance, leading to a higher F1-score compared to the other two models.

The Hybrid Proposed Model demonstrates superior performance compared to both the base paper model (80% F1- score) and the Naive Bayes model (58% F1-score). It achieves an F1-score above 86%, suggesting a strong balance between precision and recall in detecting liver disease. This indicates that the hybrid approach used in the proposed model is effective in improving the F1-score and overall performance for liver disease detection.

CONCLUSION: Artificial intelligence and machine learning are revolutionizing the prediction and management of liver cirrhosis, offering promising solutions for early detection, risk stratification, and personalized treatment. These advancements have the potential to significantly improve patient outcomes and reduce the burden of liver disease