**1. Introduction:**

**1.1 Background:**

Liver disease is a significant public health issue in India(M. Banu Priya, 2018). According to the Global Health Estimates by WHO, cirrhosis of the liver ranks as the 9th leading cause of death in India. Liver diseases, such as hepatitis, alcoholic liver disease, and non-alcoholic fatty liver disease, can lead to complications like ascites and liver cancer. The management of liver diseases can impose a significant economic burden on patients(Han et al., 2024). As such, understanding liver disease patterns and treatment outcomes through disease-specific datasets is crucial.

**1.2 Context:**

Tapas Ranjan Baitharua, et al. have developed several classification algorithms including SVM and Random Forest to classify liver disorder. The results show that Multilayer perceptron gives the overall best classification result with the accuracy 71.59% than other classifiers(Tapas RanjanBaitharua, 2016). Anju Gulia, et al. have applied SVM and Random Forest on the liver patient dataset. The results obtained from experiments indicate that Random Forest algorithm outperformed all other techniques with the help of feature selection with an accuracy of 71.8696%(Anju Gulia, 2014).

**1.3 Importance:**

Analysing the Indian Liver Patient Dataset (ILPD) serves multiple purposes. First, it contributes to the understanding of the epidemiology of liver diseases in India. Second, it supports the development of predictive models for disease progression and patient outcomes. Finally, it helps in identifying high-risk groups that may benefit from targeted interventions.

**1.4 Data exploration:**

The dataset is imported from the UCI repository. It includes attributes like albumin and other enzymes levels which are important for diagnosing liver disorders such as hepatitis. All the attributes are shown in Table 1.

Table 1: Variables table

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Role | Type | Description |
| Age | Feature | Integer | Age of the patient. Any patient whose age exceeded 89 is listed as being of age "90". |
| Gender | Feature | Binary | Gender of the patient |
| TB | Feature | Continuous | Total Bilirubin |
| DB | Feature | Continuous | Direct Bilirubin |
| Alkphos | Feature | Integer | Alkaline Phosphotase |
| Sgpt | Feature | Integer | Alamine Aminotransferase |
| Sgot | Feature | Integer | Aspartate Aminotransferase |
| TP | Feature | Continuous | Total Proteins |
| ALB | Feature | Continuous | Albumin |
| A/G Ratio | Feature | Continuous | Albumin and Globulin Ratio |
| Selector | Class | Binary | Indicates whether the patient has liver disease or not. |

Table 2: Descriptive statistics of variables

表格

描述已自动生成

Figure 1: Distribution of all variables.

图表, 箱线图

描述已自动生成

The dataset contains records for 416 patients diagnosed with liver disease and 167 patients without it, clearly indicating an imbalance in the data. Out of the 583 patient records, 441 are male, and 142 are female. The descriptive statistics for each variable are presented above. The distribution of Age, Total Protein (TP), Albumin (ALB), and Albumin-Globulin (A/G) ratio is normal, while other continuous variables show a right-skewed distribution.

Figure 2: Distribution of age for liver patients vs. non-liver patients.

图表, 直方图

描述已自动生成

The age distribution exhibits a peak within the 40-50 age range, indicating that liver disease is most prevalent among the middle-aged population.

Figure 3: Liver disease frequency for gender.

图表, 条形图, 瀑布图

描述已自动生成

The plot shows that men have a higher number and prevalence of liver disease than women.

Figure 4: Visualization of outliers.

图表, 箱线图

描述已自动生成

There are outliers in Alkaline Phosphotase, Alamine Aminotransferase and Aspartate Aminotransferase.

Figure 5: Correlation matrix for ILPD.

图表, 树状图

描述已自动生成

TP and ALB have a high positive correlation, which can be explained as total proteins include albumin; Sgpt and Sgot also have a high positive correlation, suggesting these liver enzymes often increase together, which is common in liver damage. Age has little correlation with other variables. Gender has exceptionally low correlation with liver function tests. The features with higher correlation to the presence of liver diseases are DB, TB, Alkphos, Sgpt, A/G Ratio, ALB.

Figure 6: 2-D Feature space of Alkphos and DB

图表, 散点图

描述已自动生成

This plot suggests that there is a cluster of patients without liver disease that have low levels of both Alkphos and direct bilirubin, while patients with liver disease might have a wider range of these biochemical markers. There are a few 'Liver Disease' data points that have particularly high Alkphos levels, much higher than most other data points, which could indicate cases with severe liver dysfunction.

Figure 7: 3-D Feature space.

图表, 散点图

描述已自动生成

Most data points, particularly those indicating 'No Liver Disease', are clustered near the origin point of the plot, suggesting lower levels of these measurements. The 'Liver Disease' points are more spread out and extend further along the Sgpt axis, which could imply higher levels of Sgpt among some individuals with liver disease.

**2. Methodology:**

**2.1 Data preprocessing:**

The dataset contained four missing values in the "A/G Ratio" attribute. Considering these constitute a small fraction of the dataset (4 out of 583), and after reviewing the four records, the missing values were deemed random and subsequently dropped.

Gender was mapped to a 0/1 numerical datatype, with '0' representing female and '1' representing male. The class label was originally coded 1/2; following data cleansing, this has been recoded to '0' for non-liver-diseased and '1' for liver-diseased patients to facilitate analysis. Normalisation was performed using the Min-Max Scaler function from sklearn package.

The dataset shows a significant class imbalance (167 healthy individuals versus 416 patients with liver disease), which persisted after the records were dropped, with a resulting count of 165 non-liver disease and 414 liver-disease patients. To address this, the RandomOverSampler() from the sklearn package was implemented, the class label 0 was oversampled. The summary of oversampled dataset is shown in Table 3.

Table 3: Summary of oversampled dataset

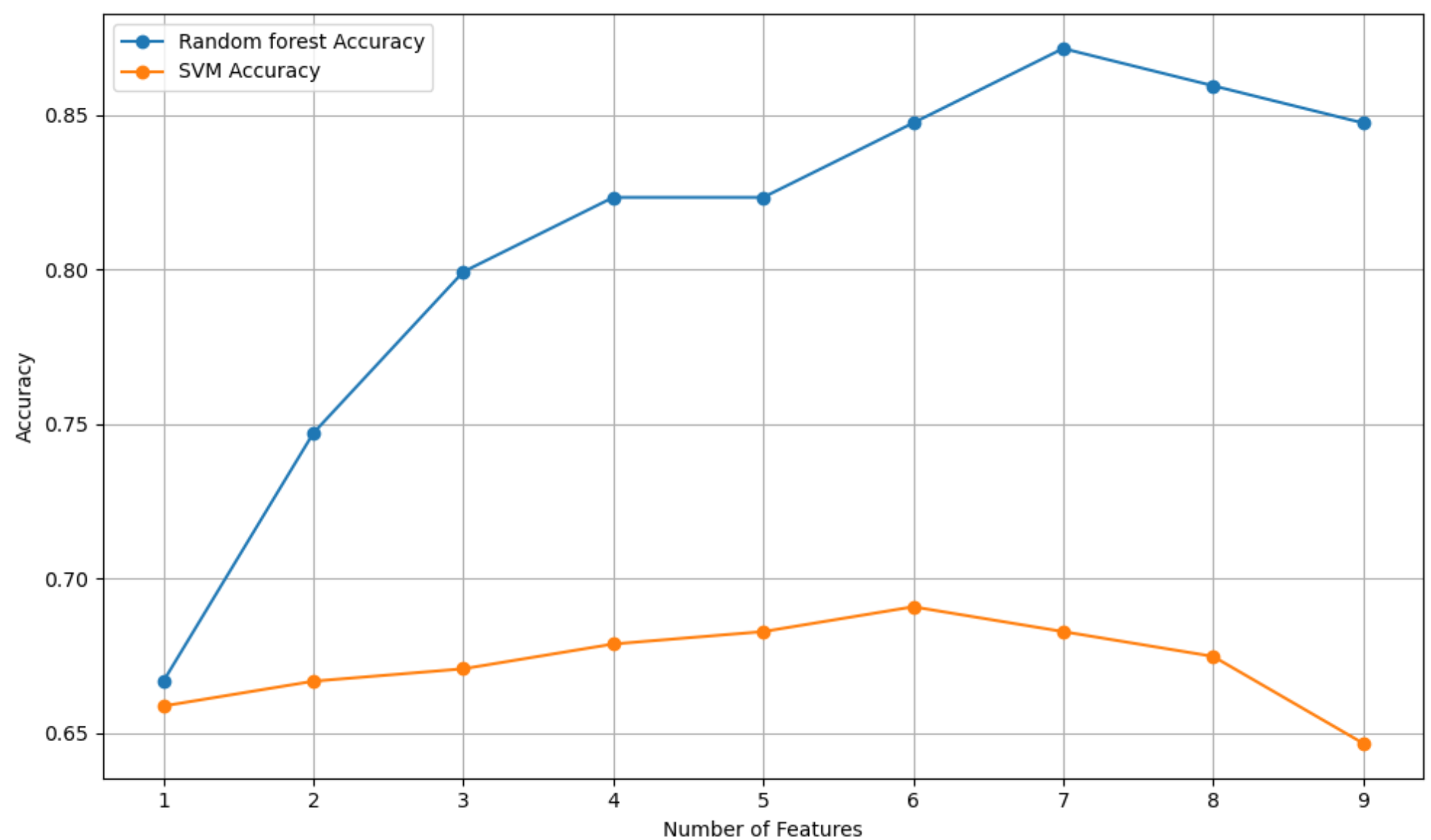


**2.2 Feature selection:**

Feature selection was performed using Sequential Backward Selection (SBS) from the sklearn package. It is a greedy search algorithm that sequentially removes features from the full feature subset until the new feature subspace contains the desired number of features. It can improve computational efficiency and reduce the generalization error of the model by removing irrelevant features.

As shown in Figure 8, using seven features seems to be optimal in terms of maintaining high accuracy without unnecessary complexity for Random Forest classifiers. As for the SVM, the choice of features does not significantly influence performance, which may indicate the feature selection was not optimized for this dataset. The top ranked features for Random Forest (RF) are Age, Gender, TB, DB, Alkphos, Sgpt and Sgot. The top ranked features for support vector machines (SVM) are Age, DB, Alkphos, Sgpt, ALB and A/G Ratio.

Figure 8: Comparison of model accuracy with varying number of features

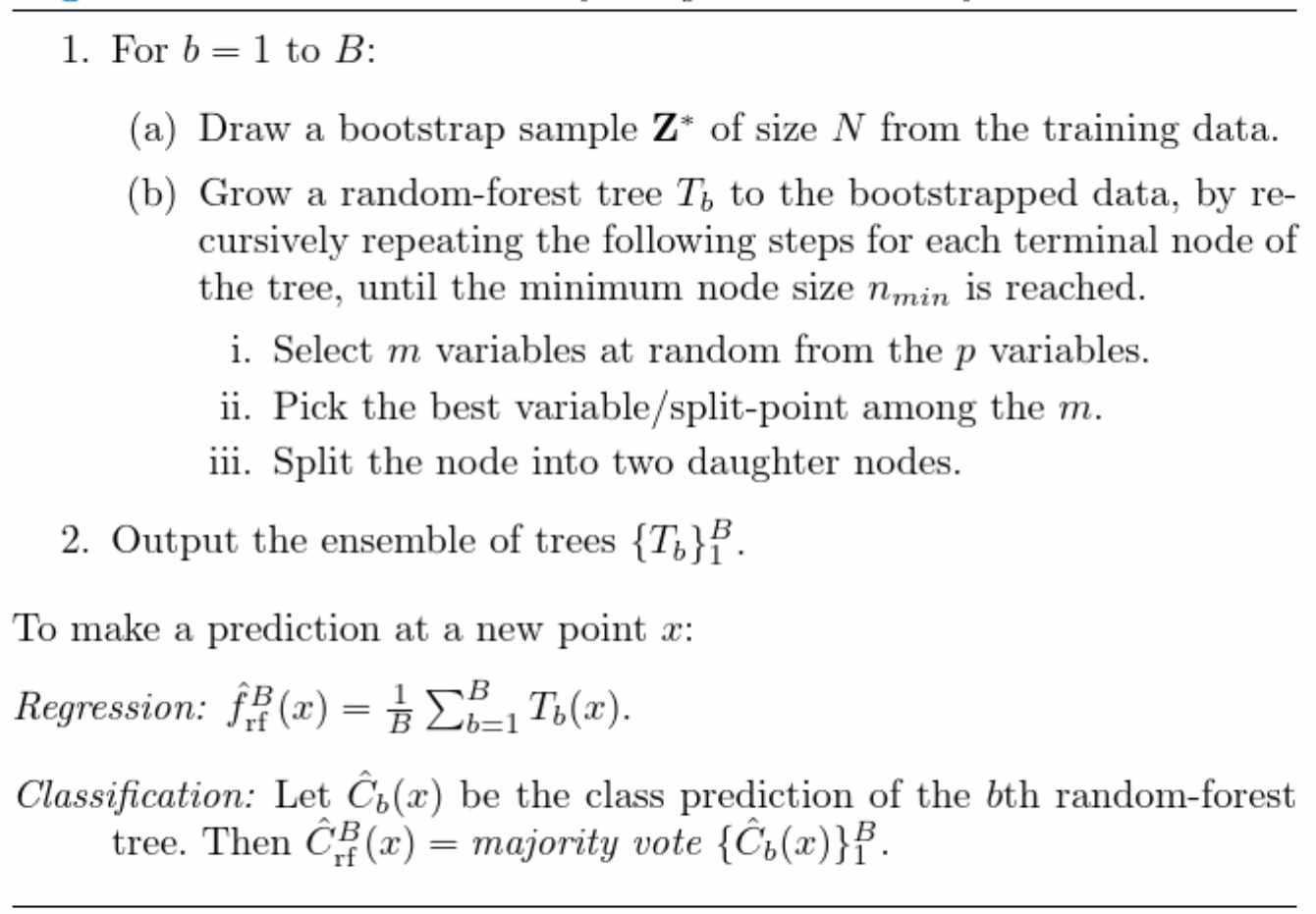


**2.3 Model development:**

**2.3.1 Random Forests (RFs):**

Random Forests (RFs) are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time(Trevor Hastie, 2009). For classification tasks, the output of the RFs is the class selected by most trees.

Figure 9: Pseudo code for the Random Forest algorithm(Trevor Hastie, 2009)



**2.3.2. Support Vector Machines (SVM):**

SVMs are supervised learning models used for classification and regression tasks. The goal of SVM is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. The best hyperplane is the one that has the largest distance to the nearest training data points of any class (functional margin), as this can help reduce generalization error(Trevor Hastie, 2009).

Given a training dataset (xn, yn) where xi are the data points and yi are the class labels, the decision function for a SVM is defined as:

(1)

sgn is the sign function that returns 1 if the argument is positive and 0 if it is negative.

are the Lagrange multipliers obtained from solving the SVM's dual optimization problem.

K(xi ,x) is the kernel function, which computes the inner product of vectors xi and x in the transformed feature space.

b is the bias term, which shifts the decision boundary away from the origin.

Radial Basis Function (RBF) Kernel:

(2)

Table 4: Pseudo code for SVM

|  |
| --- |
| Choose an SVM Kernel:  Based on the dataset, decide which kernel function to use:  Considered training data, use a radial basis function (RBF) kernel.    SVM Model Initialization:  Initialize the SVM classifier with the chosen kernel.  Set the initial values for SVM hyperparameters:  Regularization parameter (C) to control the trade-off between smooth decision boundary and classifying training points correctly.  Kernel-specific parameters (like gamma for the RBF kernel).    Model Training:  Train the SVM classifier using the training dataset:  The algorithm learns the weights that define the decision boundary hyperplane.  The algorithm identifies the support vectors that are critical data points defining the decision boundary. |

**2.4 Hyperparameter tuning:**

Data are divided into test and training subsets (30%/70%). Hyperparameters were tuned using GridSearchCV() from sklearn package, the scoring includes accuracy, AUC, f1, recall and precision.

**2.5 Model evaluation:**

The primary evaluation metrics used to measure the performance of the models included:

1. Accuracy: The proportion of total correct predictions made from all predictions.

2. Precision: The ratio of true positive predictions to the total positive predictions.

of positive instances.

4. F1 Score: The harmonic mean of precision and recall.

5. Confusion Matrix: A detailed breakdown of the model's predictions, allowing us to see the number of true positives, false positives, true negatives, and false negatives.

6. Receiver Operating Characteristic (ROC) Curve: A plot that illustrates the diagnostic ability of the classifier as its discrimination threshold is varied. The curve is created by plotting the true positive rate (recall) against the false positive rate at various threshold settings.

7. AUC score: It refers specifically to the area under the Receiver Operating Characteristic (ROC) curve.

**3.Results:**

**3.1 Results without** **hyper-parameter optimization**

First, SVM and RF were conducted without hyper-parameter tuning. The number of trees for RF is 100. The parameter C equals 0.1, and the kernel is 'rbf' for SVM. The training time for RF is 0.369 seconds, and the training time for SVM is 0.165 seconds. The precisions for the two classifiers are: RF 0.90 and SVM 0.75. The recalls for the two classifiers are: RF 0.79 and SVM 0.43. The F1 scores for the two classifiers are: RF 0.84 and SVM 0.55. The confusion matrices for the two models are shown in Figure 10. The AUC scores are 0.96 and 0.71, respectively, as shown in Figure 11.

Figure 10: Comparison of confusion matrix for two models

图表, 树状图

描述已自动生成

Figure 11: ROC curve comparison

图表, 折线图

描述已自动生成

**3.2 Results for hyper-parameter optimization**

The experiments were performed on a PC equipped with an Intel i7-13700H processor (2.4 GHz base clock speed), 16 GB RAM. The operating system was Windows 11, version 23H2. The machine learning models were implemented using Python 3.11.4 with scikit-learn version 1.4.0.

The number of cross validations is 5.

For SVM, optimal parameters such as C, gamma, and the kernel type for performance metrics are shown in Table 5.

For RF, optimal parameters such as the number of trees, maximum depth, and criterion for performance metrics are shown in Table 6.

Table 5: Optimal Hyperparameters for SVM Classifier

**表格

描述已自动生成**

Table 6: Optimal hyperparameters for Random Forest Classifier

|  |  |  |  |
| --- | --- | --- | --- |
| Scoring | Best Score | Best Parameters | Elapsed Time |
| accuracy | 0.808366 | {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 100} | 36.52239 |
| roc\_auc | 0.926454 | {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 300} | 37.39461 |
| f1 | 0.805769 | {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 100} | 37.51829 |
| recall | 0.807921 | {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 100} | 35.5059 |
| precision | 0.822939 | {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 100} | 35.68107 |

**3.2 Results for feature ranking:**

From the Random Forest output, the feature rankings are: Alkphos (0.170), Sgpt (0.147), , Age (0.140), Sgot (0.132), TB (0.115), TP (0.102), DB (0.100), ALB (0.075) and Gender (0.018).

**3.3 Results for optimal SVM on accuracy without feature selection:**

There is a significant improvement in overall performance compared to SVM without hyper-parameter optimization. The accuracy of optimal SVM on accuracy (C=100, gamma=100, kernel=’rbf’) is 0.823, the F1 score is 0.818, recall is 0.786, precision is 0.853. The confusion matrix is shown in Figure 12. The AUC score is 0.89 as shown in Figure 13.

Figure 12: Confusion matrix for SVM

**图表, 树状图

描述已自动生成**

Figure 13: The ROC plot of optimal SVM

**图表, 折线图

描述已自动生成**

**3.4 Results for optimal RF on accuracy without feature selection:**

There isn't an improvement in overall performance when comparing RF with and without hyper-parameter optimization. This could be attributed to the non-linearity of the data; SVM with a non-linear kernel might better model these complexities after optimization. Conversely, RF may not exhibit significant gains because it might already be performing near its optimum with default hyperparameters on the given dataset. Another contributing factor could be the range and granularity of the hyperparameter grid used for RF. If the hyperparameter space was not thoroughly explored, or if the optimal parameters fall outside the specified grid, then improvements could be negligible.

The accuracy of optimal RF on accuracy (number of trees are 100, criterion is entropy, max depth is 15) is 0.843, the F1 score is 0.834, recall is 0.777, precision is 0.899. The confusion matrix is shown in Figure 14. The AUC score is 0.95 as shown in Figure 15.

Figure 14: Confusion matrix for RF

**图表, 树状图

描述已自动生成**

Figure 15: The ROC plot of optimal RF

**图表, 折线图

描述已自动生成**

**4. Conclusion:**

The study has demonstrated the application of SVM and RF algorithms on the ILPD, reinforces the notion that machine learning can play a pivotal role in disease classification. SVM and RF have both been optimized and evaluated in terms of their performance on the dataset. It has been seen that Random Forest gives better results compare to SVM. And hyperparameters tuning can help build a better-performing model.

**5. Discussion:**

The results of this study align closely with those reported by Tapas Ranjan Baitharu et al., yet my research has realized a slightly higher accuracy in SVM classification in comparison to the studies by Anju Gulia, M. Banu Priya, and Isabel Straw(Tapas RanjanBaitharua, 2016, Straw and Wu, 2022, Anju Gulia, 2014, Baitharu and Pani, 2016). This improvement can likely be attributed to the distinct preprocessing procedures and hyperparameter optimization techniques that were implemented. While the recall metrics are consistent with those found in Isabel Straw’s work, my study demonstrates enhanced precision, AUC scores, and F1 scores. In the case of the Random Forest classifier, my research outperforms that of Selamawit Sileshi Nigatu in terms of accuracy, and it surpasses the overall performance metrics reported in the work by Isabel Straw(Nigatu et al., 2023, Straw and Wu, 2022).

Differences in results among various studies are common due to several factors, which may include but are not limited to: Variations in preprocessing steps, such as normalization and handling of missing values. For example, Selamawit used StandardScaler() to normalise the data and Isabel Straw used mean imputation to address missing values(Straw and Wu, 2022, Nigatu et al., 2023). Furthermore, the choice of hyperparameter range and randomness in training and testing datasets can also yield disparate outcomes.

Expanding upon feature engineering and selection might further refine model performance but is of limited use. Certain variables in the ILPD dataset serve as covariates, displaying some degree of collinearity as depicted in Figure 5. Advanced feature selection could not only enhance performance but also boost model interpretability. Furthermore, ensemble methods beyond RF, such as gradient boosting or stacking, could be explored to improve predictive accuracy(Md et al., 2023). Additionally, deep learning approaches could be investigated to capture complex patterns in the data. Some studies have employed various ML models like ANN, which have shown high effectiveness, with accuracies up to 87%(Nigatu et al., 2023).

Deploying these algorithms could assist healthcare professionals in the early diagnosis of liver diseases, potentially improving patient outcomes and reducing the economic and staffing burdens on healthcare systems. By integrating these models into electronic health records systems, real-time predictive analytics could become a feasible tool for clinicians to screen patients for liver diseases. While the optimized SVM and RF models show promise, it is imperative to validate these findings with larger, external datasets before clinical implementation. Studies by Isabel and Honghan has already examined disparities in algorithmic performance and demonstrated sex differences in data cause high false negative rate(Straw and Wu, 2022).

The dataset focuses on the Indian population and includes only a few features. By collecting more samples from other countries and incorporating additional demographic and disease-related variables, racial biases can be reported, race stratified analysis can be conducted. With these improvements, preventive measures could be more effectively directed toward high-risk groups identified through the analysis. This would enhance diagnostic accuracy and efficiency, leading to better resource allocation and personalized patient care.

**6. Reference:**

1. ANJU GULIA, D. R. V., PRAVEEN RANI 2014. Liver Patient Classification UsingIntelligent Techniques. *International Journal of Computer Science and Information Technologies,* 5**,** 5110.

2. BAITHARU, T. R. & PANI, S. K. 2016. Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset. *International Conference on Computational Modelling and Security (Cms 2016),* 85**,** 862-870.

3. HAN, X., ZHANG, X., LIU, Z., FAN, H., GUO, C., WANG, H., GU, Y. & ZHANG, T. 2024. Prevalence of nonalcoholic fatty liver disease and liver cirrhosis in Chinese adults with type 2 diabetes mellitus. *J Diabetes,* 16**,** e13564.

4. M. BANU PRIYA, P. L. J., P.R. TAMILSELVI 2018. Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms. *International Research Journal of Engineering and Technology (IRJET),* 05.

5. MD, A. Q., KULKARNI, S., JOSHUA, C. J., VAICHOLE, T., MOHAN, S. & IWENDI, C. 2023. Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease. *Biomedicines,* 11.

6. NIGATU, S. S., ALLA, P. C. R., RAVIKUMAR, R. N., MISHRA, K., KOMALA, G. & CHAMI, G. R. 2023. A Comparitive Study on Liver Disease Prediction using Supervised Learning Algorithms with Hyperparameter Tuning. *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT).*

7. STRAW, I. & WU, H. 2022. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform,* 29.

8. TAPAS RANJANBAITHARUA, S. K. P. 2016. Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset. *International Conference on Computational Modeling and Security*.

9. TREVOR HASTIE, R. T., JEROME FRIEDMAN 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*, Springer New York, NY.