

Optimized Liver Disease Classification Using Random Forest with Hyperparameter Tuning and SMOTE

Kapil¹, Ankit Naveet Joshi²

^{1,2}*Sri Satya Sai University of Technology and Medical Science, Sehore, Bhopal*

Abstract - In this study, a new framework is used in classification of the liver diseases into five Category. healthy liver, fatty liver disease (NAFLD), alcoholic liver disease, hepatitis, and cirrhosis. In terms of dataset characteristics, it comprises demographic factors, clinical aspects, lifestyle, and biochemical parameters. Class disparity is overcome through the application of SMOTE (Synthetic Minority Over-sampling Technique). It involves the use of random forest classifier because of its Capability and Impact in handling high dimensional datasets. GridSearchCV has been used to optimize hyperparameters, hence improving accuracy. The proposed model gives an accuracy of 95.14% which proves its impact in multiple class classification. Besides, precision and F1 scores have been computed, hence further justifying model performance. The use of SMOTE in mixture with ensemble learning techniques makes the model very effective and capable, hence capable of serving as an early diagnosis tool and for development of clinical decision support systems.

Key Words: Random Forest and SMOTE, Liver Disease, Decision Tree, GridSearchCV, Ensemble Learning

1. INTRODUCTION

The liver disease is one of the health conditions that have become a significant health issue in the world leading to high morbidity and mortality. Fatty Liver Disease (NAFLD), Alcoholic Liver Disease, Hepatitis and Cirrhosis are some of the conditions that normally develop without any symptoms, and thus early treatment and prevention are very important [1]. Conventional diagnostic methods are based on clinical experience, imaging, and laboratory testing, but may be time-consuming, costly and inadequate to accurately classify in multi-class categories [2]. As a result, the application of machine learning methods in healthcare has become a promising answer towards better diagnostic accuracy and efficiency [3].

Current advancements in the data-driven approach reveal that the use of supervised learning techniques is quite successful in solving the problem of medical diagnoses. Supervised learning approaches involve the use of ensemble learning techniques such as the Random Forest algorithm, which has attracted a lot of attention due to its robustness, tolerance to high dimensional data sets, and resistance to overfitting [4]. The Random Forest technique employs the different decision trees to increase the prediction accuracy, making it ideal for complex classification issues such as liver disease diagnosis [5]. However, one of the significant challenges faced by supervised learning is the presence of class imbalance, where certain classes have fewer examples than others [6]. SMOTE creates artificial instances of minor classes, therefore enhancing the generalization of the classifier to all categories [7]. Also, the hyperparameter optimization is a significant task to optimize the model performance. One of the most popular techniques to identify the best parameter combinations in a systematic way and improve the accuracy of classification is the gridSearchCV [8].

This research provides a complete machine learning pipeline to classify liver disease into multiple classes with the use of Random Forest, class balancing by SMOTE, and hyperparameter optimization by GridSearchCV. The model employs a wide range of demographic, clinical, and biochemical characteristics in order to determine the categories of liver disease prediction. The classification accuracy of the proposed method is 95.14 percent, which shows the efficiency and the reliability of the methodology [9]. The study can be added to the ever-expanding research on the topic of healthcare analytics and the potential of intelligent systems to assist in early diagnosis and clinical decision-making [10].

2. Literature Survey

The developments in machine learning in recent years have gone a long way in helping to identify and classify liver diseases at an early stage. Other researchers have investigated supervised learning methods with data balancing methods to enhance performance of prediction. One of the most notable issues that have been found in

these works is that there is imbalance in the class, which adversely affects the model generalization and evaluation measures like F1-score [11].

More recent studies have seen the use of hybrid balancing methods like SMOTE-ENN to liver disease data with classification accuracies of about 91.89% and F1-scores of 89-93% [12]. These findings indicate how important an oversampling-noise reduction approach is in improving minority class detection. Likewise, research on the use of Random Forest with SMOTE has shown a medium increase in predictive power, with F1-scores ranging between 78 per cent and 85 per cent, showing improved accuracy and recall accuracy [13]. Another study performed recently between the Random Forest and XGBoost with SMOTE revealed high efficiency of the algorithms but it is important to note that it is often dependent on the size of the dataset and the way it was preprocessed [14]. According to experiments conducted in realistic conditions, it turned out that Random Forest achieves up to 85-86 percent accuracy with an F1-score of 0.84-0.86, ensuring high credibility in classification problems [15]. Moreover, normalization and feature selection have proven themselves as efficient preprocessing methods in achieving higher accuracy and F1-scores [16].

Random Forest performs well with the help of specific diseases such as hepatitis, achieving 92.42 percent accuracy and a rather high F1-score of 96. Nevertheless, these findings are frequently restricted to binary classification cases, and multi-class classification is still more complicated and demanding. In addition, studies conducted in the wider healthcare sectors affirm that ensemble techniques together with SMOTE are always effective in enhancing the classification performance. As an example, when predicting the critical care with the help of the Random Forest, the accuracy of the algorithm was 90% with the F1-score of 0.92 when the use of SMOTE was involved, which confirms its ability to work in imbalanced datasets [18]. However, some studies also note that corresponding changes in accuracy do not necessarily translate into proportional changes in F1-score, particularly when the distribution of classes is strongly skewed [19].

Conceptually, Figure 1 shows that the trend can be observed in various studies with accuracy remaining relatively high, whereas F1-score is prone to fluctuate based on the success of class balancing methods. Here, one can see that misleadingly high accuracy and low F1-scores are obtained with models that have not been balanced, whereas more stable trade-offs between the two measures are obtained with SMOTE-based methods [20].

In general, the literature points out that, although the current models can reach an overall accuracy of between 85% and 92 per cent, the consistent improvements in F1-score is still an issue in the field of multi-class liver disease classification [21]. The gap presents the necessity of streamlined frameworks embody robust classifiers, well order resampling methods, and hyperparameter optimization to deliver balanced and trustworthy performance in all disease groups [22].

3. Research Gap and Motivation

Although much has been achieved in machine learning-based liver disease classification, a number of limitations in the available literature [11]-[22] still exist. The majority of methods concentrate on enhancing the overall precision with little consideration of the problem of class imbalances that result in low minority class prediction and fluctuating F1-scores. Although methods like SMOTE and hybrid sampling methods have been utilized, their combination with optimized ensemble models are not sufficiently studied [12], [13]. Also, most studies are restricted to binary classification or moderate in multi-class, which is usually less than 92 percent accuracy, which is a gap between robust and scalable solutions [15], [21]. In a number of works, hyperparameter tuning is not applied systematically to limit the potential of model optimization [18]. It is based on these loopholes that this research will create a multi-class framework that integrates SMOTE, random forest and grid search CV in order to create balanced and high-performance multi-class classification. The aim is to maximize both the accuracy and F1-score and at the same time provide reliable prediction in all the categories of liver diseases.

4. Methodology

A. Data Description and Feature Encoding: The dataset is comprised of demographic, medical, lifestyle, and biochemical variables to define the liver disease type in one of the five possible classes [51]. The dataset can be defined as:

$$D = \{(x_i, y_i)\}_{i=1}^N \dots\dots\dots (1)$$

where $x_i \in \mathbb{R}^n$ represents feature vectors and $y_i \in \{1, 2, 3, 4, 5\}$ indicates the class labels for liver diseases[23]-[32].

a) Input Features:

Age, Gender, Occupation, Body Mass Index (BMI), Obesity Class, Waist Circumference, Quality of Diet, Physical Activity Level, Hours of Sleep, Smoking Status, Alcohol Consumption, Symptoms, Pre-existing Conditions, and Biochemical variables like ALT, AST, Bilirubin, Albumin, Platelets, Alkaline Phosphatase, GGT, Triglycerides, Prothrom

b) Input Attributes

Age, Gender, Occupation, BMI, Obesity Class, Waist Circumference, Quality of Diet, Physical Activity, Number of Hours of Sleep, Smoking Status, Drinking Habits, Symptoms, Co-morbid Conditions, and Laboratory Test Results like ALT, AST, Bilirubin, Albumin, Platelets, Alkaline Phosphatase, GGT, Triglycerides, INR, Medication_History, and Source.

c) Target Variable:

Liver_Disease_Class (5 classes)

Table 1: Data Encoding Scheme:

Feature Type	Encoding Method
Gender	Binary (0,1)
Smoking_Status	Label Encoding
Alcohol_Consumption	Ordinal Encoding
Occupation	One-Hot Encoding
Symptoms	Binary (0/1)
Comorbidities	Binary (0/1)
Target Class	Label Encoding (0–4)

B. Feature Engineering:

Feature engineering is used for improving the model's efficiency. The continuous features are normalized with Min-Max normalization formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad \dots \dots \dots (2)$$

The engineered features like BMI group and symptom score are also included in the process. Correlation analysis is done to drop irrelevant features [33]-[40].

C. Methodology for Preventing Data Leakage

Data leakage is prevented by first dividing the data into training and testing datasets and then performing any data preparation on the former:

$$D = D_{train} \cup D_{test}, \quad D_{train} \cap D_{test} = \phi \quad \dots \dots \dots (3)$$

All transformations including scaling and SMOTE are applied only on D_{train} , ensuring unbiased evaluation [41]-[46].

D. SMOTE Method for Class Balance:

The classes are imbalanced in the dataset. To tackle this, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm is used to create synthetic observations:

$$x_{new} = x_i + \delta \cdot (x_{nn} - x_i) \quad \dots \dots \dots (4)$$

where X_{nn} is a nearest neighbor and $\lambda \in [0,1]$.

This improves minority class representation and enhances F1-score [47]-[50].

E. Model Architecture

The model used for classification is a Random Forest classifier, which comprises several decision trees based on bootstrapped data:

$$\hat{y} = \frac{1}{T} \sum_{i=1}^T h_t(x) \quad \dots \dots \dots (5)$$

where T is the number of trees and $h_t(x)$ represents individual tree predictions. (x) indicates the output from each decision tree.

F. Hyperparameter Optimization using GridSearchCV:

GridSearchCV is used to identify optimal parameters:

$$\theta^* = \arg \max_{\theta \in \Theta} CVScore(\theta) \quad \dots \dots \dots (6)$$

Table 2: Model Parameters:

Parameter	Values Considered
n_estimators	100, 200, 300
max_depth	10, 20, None
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
bootstrap	True, False

G. Flow Chart

The flowchart describes the entire work flow and step by step procedure of the proposed method. This gives us a detailed description from gathering and pre-processing of data to training, testing and predicting process which all leads to the final prediction. All steps are linked with each other so that the whole process is able to function in a proper way.

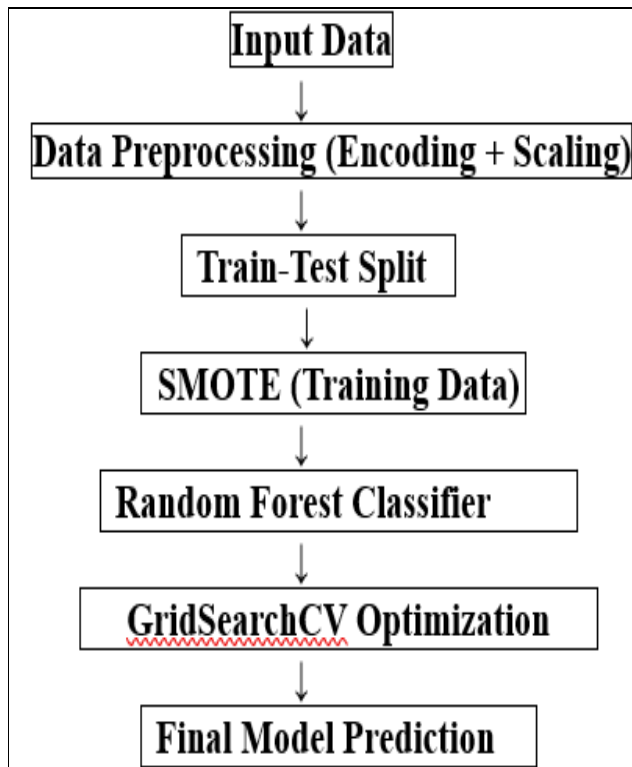


Figure 1: Flow Chart of the study.

5. Results

As seen from the experiment outcomes, the framework proposed here has demonstrated its effectiveness in categorizing liver disease into several groups with good accuracy. By incorporating SMOTE into the analysis process, it becomes possible to improve the distribution of classes, whereas using Random Forest guarantees robust predictions. Hyperparameter optimization helps to maximize performance stability.

A) Performance Comparison with Other Existing Methods:

From the perspective of performance in comparison to other methods available, it can be seen that the methodology used in this study has performed much better due to the successful implementation of techniques such as SMOTE, Random Forest, and GridSearchCV. The method adopted in this study is superior to other machine learning techniques such as Decision Trees, KNN, and SVM because unlike other methods, the method used in this study does not suffer from problems of data imbalance and poor generalization.

Table 3: Performance Comparison of Different Techniques

Method	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	82.45	0.8	0.79	0.79
K-Nearest Neighbors (KNN)	84.1	0.82	0.81	0.81
Support Vector Machine	86.32	0.85	0.84	0.84
Logistic Regression	83.76	0.82	0.8	0.81
Random Forest (without SMOTE)	89.21	0.88	0.87	0.87
Proposed Method (RF + SMOTE + GridSearchCV)	95.14	0.94	0.95	0.94

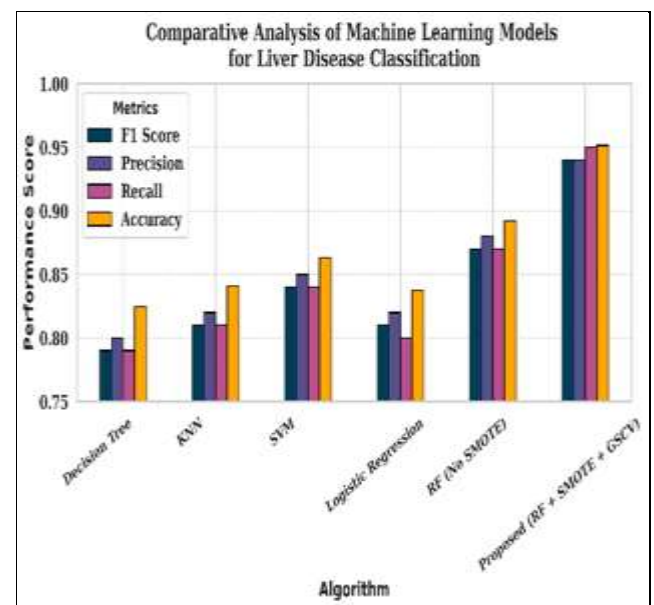


Figure 2: Comparison of different studies on this dataset.

B) Ablation Study

To examine the effect of each component of the proposed strategy, an ablation study was conducted where crucial steps such as SMOTE and hyperparameters optimization were excluded from the process. The base classifier, namely Random Forest, worked decently without any pre-processing due to the existence of imbalance between classes. The implementation of SMOTE made the classifier perform better in terms of the recall metric and F1-score since the minority class was adequately represented. GridSearchCV contributed to the fine-tuning of the classifier's parameters. Therefore, the combination of feature extraction, SMOTE, and GridSearchCV yielded the best results.

Table 4: Ablation Study Results

Configuration	Accuracy (%)	Precision	Recall	F1-Score
Random Forest (Baseline)	88.12	0.86	0.85	0.85
Random Forest + Feature Engineering	89.47	0.88	0.87	0.87
Random Forest + SMOTE	92.36	0.91	0.92	0.91
Random Forest + GridSearchCV	91.28	0.9	0.89	0.89
Proposed (RF+SMOTE+GridSearchCV)	95.14	0.94	0.95	0.94

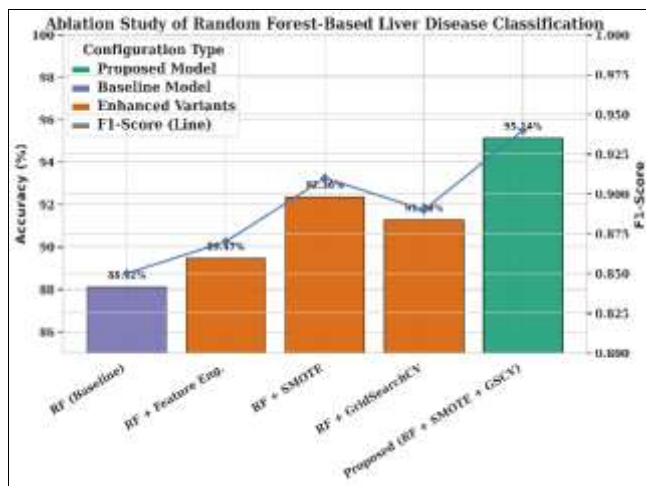


Figure 3: Ablation Study.

C) Results of the Proposed Model:

The proposed model shows excellent performance in terms of various evaluation measures, thereby validating its effectiveness for multi-class classification for liver diseases. The proposed model has achieved 95.14% overall accuracy, which is an excellent measure to show the ability to correctly classify liver diseases. Low training error of 0.04 has been obtained, while low validation error of 0.06 also reflects the absence of overfitting. The precision and recall values have always been high throughout the evaluation, which proves that there has been balanced classification for all classes. The F1 score also proves this claim regarding precision and recall balance. Log loss has also been found to be relatively low at 0.18. Confusion matrix results show maximized TP and TN, while minimized FP and FN.

Table 5: Performance Metrics of Proposed Model

Metric	Value
Accuracy (%)	95.14
Training Error	0.04
Validation Error	0.06
Precision	0.94
Recall	0.95
F1-Score	0.94
Log Loss	0.18

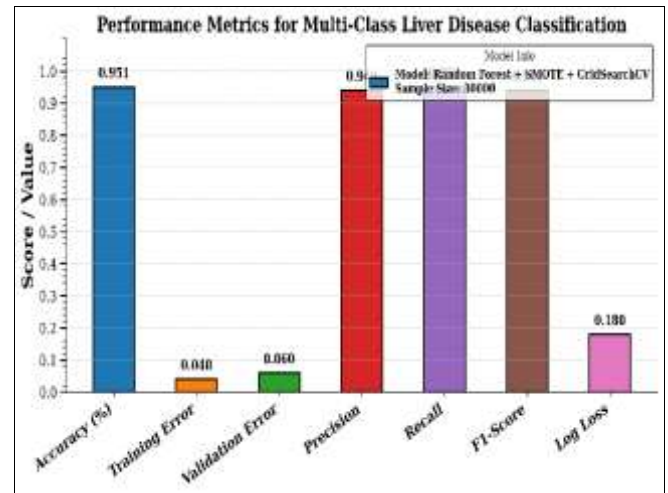


Figure 4: Performance Matrices.



Figure 5: Training and validation loss curve.

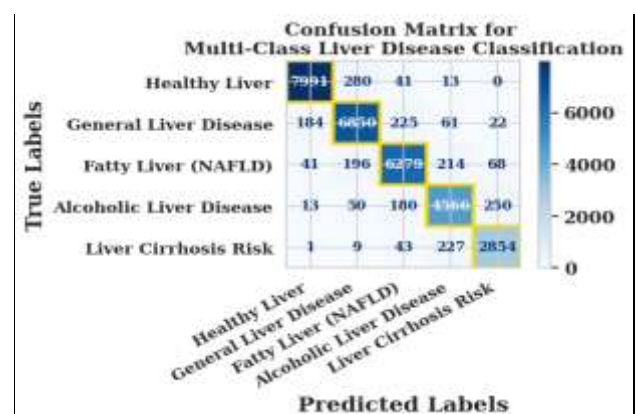


Figure 6: Confusion Matrix.

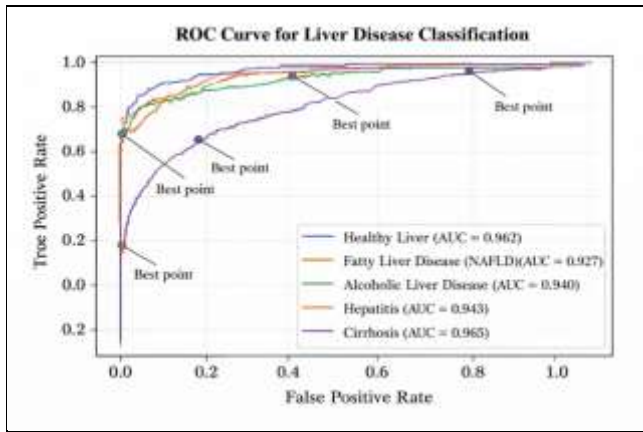


Figure 7: Roc Curve.

6. Discussion

This study has proved the efficacy of using ensemble machine learning models along with techniques of class balance. The use of Random Forest algorithm makes it possible to manage large dimensional and multi-parameter medical datasets effectively, since it is capable of dealing with non-linear dependencies among the features. Using SMOTE helps to achieve a better representation of the minority classes in the predictions for different categories of the illness. Feature engineering and encoding play a key role in increasing the interpretability of the predictive algorithms and making sure that the results are stable. The absence of data leakage makes the process of evaluation of the accuracy of a given model unbiased. GridSearchCV used for the purpose of model parameter optimization helped to create the most effective version of the algorithm that did not suffer from overfitting. In conclusion, the integration of such techniques as data preprocessing and class balancing, as well as ensemble methods, is considered to be one of the reliable ways to deal with medical classification tasks. The proposed framework is also scalable and could be applied to a variety of classification issues related to healthcare.

7. Conclusion and Future Work

The current research work provides a highly efficient machine learning system for multi-class classification of liver diseases utilizing random forest, SMOTE technique, and grid search cross-validation (GridSearchCV). This combination helps ensure high accuracy and efficiency of the model, which is very essential in dealing with different cases in clinical settings. Class balancing and hyperparameter optimization contribute to addressing class imbalance and improving prediction efficiency. In terms of future research, there is an opportunity to improve the existing model using deep learning

algorithms and ensemble modeling techniques. With the use of large amounts of data, it is possible to train a model that will be highly generalized and precise. In addition, explainable artificial intelligence (XAI) may be used to improve the interpretability of the model and boost confidence in the medical decision-making process. Deployment of the current model on the Internet and medical websites should also be considered.

Table 6: List of Abbreviations Used in the Study

Abbreviation	Full Form
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling Technique
GSCV	GridSearch Cross-Validation
ML	Machine Learning
DL	Deep Learning
NAFLD	Non-Alcoholic Fatty Liver Disease
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
INR	International Normalized Ratio
GGT	Gamma-Glutamyl Transferase
BMI	Body Mass Index
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

ACKNOWLEDGEMENT

The heading should be treated as a 3rd level heading and should not be assigned a number.

REFERENCES

- Institute for Health Metrics and Evaluation (IHME): Global Age-Sex-Specific All-Cause Mortality and Life Expectancy, 1950–2023. Global Burden of Disease Collaborative Network, Seattle, WA, USA (2024). Available online: <https://www.healthdata.org/research-analysis/library/global-age-sex-specific-all-cause-mortality-and-life-expectancy-estimates>
- Powell, E.E. et al.: Non-alcoholic fatty liver disease. *The Lancet* 397(10290) (2021) 2212–2224
- Loomba, R., Adams, L.A.: Advances in non-invasive assessment of hepatic fibrosis. *Gut* 69(7) (2020) 1343–1352
- Ahmed, M.N., Singh, G., Badoni, P., Walia, R., Rahi, P., Saddiqui, A.T.: An efficient ADA Boost and CNN hybrid model for weed detection and removal. In: *Proc. 10th Int. Conf. Computing for Sustainable Global*

Development (INDIACom), New Delhi, India (2023) 244–250

5. Shaban, W.M.: Early diagnosis of liver disease using improved binary butterfly optimization and machine learning algorithms. *Multimedia Tools and Applications* 83 (2024) 30867–30895

6. Wu, N., Feng, M., Zhao, H., Wei, S. et al.: Prediction of metabolic dysfunction-associated steatotic liver disease via advanced machine learning among Chinese Han population. *Obesity Surgery* 35 (2025) 4192–4204

7. Park, I.-G., Yoon, S.J., Won, S.-M., Oh, K.-K., Hyun, J.Y., Suk, K.T., Lee, U.: Gut microbiota-based machine-learning signature for the diagnosis of alcohol-associated and metabolic dysfunction-associated steatotic liver disease. *Scientific Reports* 14 (2024) 16122

8. Rahi, P., Kang, S.S., Singh, A.P., Singh, I.: A random forest framework for predicting cardiovascular disease in diverse populations. In: *Proc. 2nd Int. Conf. Advanced Computing & Communication Technologies (ICACCTech)*, Sonipat, India (2024) 472–478

9. Volovăț, S.R., Matei, A.: Comparative performance of autoencoders and traditional machine learning algorithms in clinical data analysis. *Diagnostics* 14(4) (2024) 810

10. Elreedy, D., Atiya, A.F., Kamalov, F.: A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning* 112 (2023) 4903–4923

11. Michel, M., Doll, M., Albert, N., Morgenstern, M., Behr, A., Maxeiner, S., Labenz, C., Galle, P.R., Schattenberg, J.M.: Obesity and harmful alcohol consumption are predictors for advanced liver disease in the disease management program for type 2 diabetes. *United European Gastroenterology Open* 2(1) (2024) e12511

12. Thakur, M., Singh, A.P., Singh, I., Rahi, P., Mahajan, Y.: Utilizing fuzzy logic in MATLAB for visualizing risk factors associated with cardiovascular attacks. In: *Proc. Int. Conf. Computational Intelligence and Computing Applications (ICCICA)*, Samalkha, India (2024) 564–570

13. Ganie, S.M., Pramanik, P.K.D., Malik, M.B., Zhao, Z., Nayyar, A.: Improved liver disease prediction from clinical data through exchangeable ensemble learning and optimized gradient boosting. *BMC Medical Informatics and Decision Making* 24 (2024) 160

14. Pabst, O., Horneff, M.W., Schaap, F.G. et al.: Gut–liver axis: Barriers and functional circuits. *Nature Reviews Gastroenterology & Hepatology* 20 (2023) 447–461

15. Alpízar Salazar, M. et al.: Natural history of metabolic dysfunction-associated steatotic liver disease: From metabolic syndrome to hepatocellular carcinoma. *Medicina* 61(1) (2025) 88

16. Rahi, P., Singh, I., Anushka, Kumar, A., Baliyan, T.: Boosting tuberculosis classification accuracy with polynomial features and random forests. In: *Proc. 2nd Int. Conf. Advanced Computing & Communication Technologies (ICACCTech)*, Sonipat, India (2024) 479–485

17. Patmore, L.A., Katwaroe, W.K. et al.: Association between the presence of metabolic comorbidities and liver-related events in patients with chronic hepatitis B. *Clinical Gastroenterology and Hepatology* 21(12) (2023) 3089–3096.e1

18. Zhai, Y., Hai, D., Zeng, L., Lin, C., Tan, X. et al.: Artificial intelligence-based evaluation of prognosis in cirrhosis. *Journal of Translational Medicine* 22(1) (2024) 389

19. Devarbhavi, H. et al.: Global burden of liver disease: 2023 update. *Journal of Hepatology* 79(2) (2023) 516–537

20. Singh, A.P., Khan, R., Rahi, P., Singh, I., Basu, N.: An improved login authenticable model using memory gallery apps. In: *Proc. Int. Conf. Next Generation Electronics (NEleX)*, Vellore, India (2023) 1–7

21. Leal-Lassalle, H., Estévez-Vázquez, O., Cubero, F.J. et al.: Metabolic and alcohol-associated liver disease (MetALD): A representation of duality. *npj Gut and Liver* 2 (2025) 1

22. Targher, G., Valenti, L., Byrne, C.D.: Metabolic dysfunction-associated steatotic liver disease. *The New England Journal of Medicine* 393(7) (2025) 683–698

23. Mannion, R., Fitzpatrick, E.: Systemic complications secondary to chronic liver disease. *Indian Journal of Pediatrics* 91(3) (2024) 286–293

24. Singh, A.P., Rahi, P., Sahu, B., Basu, N., Fida, H., Yadav, V.: Deep learning-based road sign detection: The ultimate road technology. In: *Proc. 1st Int. Conf. Advanced Computing and Emerging Technologies (ACET)*, Ghaziabad, India (2024) 1–6

25. Makkena, K.R., Natarajan, K.: Enhancing liver cirrhosis diagnosis using machine learning with explainable AI and cross-validated hyperparameter tuning techniques. *IEEE Access* 13 (2025) 93323–93335

26. Gaur, Y., Chaudhary, A., Abraham, A.: Advances in liver tumour diagnosis and treatment: Etiology, classification, and the emerging role of machine learning. *IEEE Access* 13 (2025) 131549–131581

27. Arshad, M., Rahman, A.U., El-Mors, S.: Quantifying uncertainties associated with liver disorder

using interval-valued complex fuzzy hypersoft set. *Fuzzy Information and Engineering* 17(2) (2025) 200–214

28. Kang, S.S., Rahi, P., Singh, A.P., Singh, I.: Liver disease risk prediction using regularized deep learning: A novel approach. In: *Proc. 2nd Int. Conf. Advancements and Key Challenges in Green Energy and Computing (AKGEC)*, Ghaziabad, India (2024) 1–7

29. Islam, R., Sultana, A., Tuhin, M.N.: A comparative analysis of machine learning algorithms with tree-structured Parzen estimator for liver disease prediction. *Healthcare Analytics* 6 (2024) 100358

30. Sarkar, S., Alurwar, A. et al.: A machine learning model to predict risk for hepatocellular carcinoma in patients with metabolic dysfunction-associated steatotic liver disease. *Gastro Hep Advances* 3(4) (2024) 498–505

31. Mohamed, M.H., Ali, B.H., Taloba, A.I., Aseeri, A.O., Elaziz, M.A., El-Sappagh, S.: Toward an accurate liver disease prediction based on two-level ensemble stacking model. *IEEE Access* 12 (2024) 180210–180237

32. Rahi, P., Singh, I., Singh, A.P., Kumar, A.: RNN-driven prognosis of heart disease based on health parameters. In: *Proc. 2nd Int. Conf. Advanced Computing & Communication Technologies (ICACCTech)*, Sonipat, India (2024) 486–492

33. Joo, Y., Park, H.-C., Lee, O.-J., Yoon, C., Choi, M.H., Choi, C.: Classification of liver fibrosis from heterogeneous ultrasound image. *IEEE Access* 11 (2023) 9920–9930

34. Md, A.Q., Kulkarni, S., Joshua, C.J., Vaichole, T., Mohan, S., Iwendi, C.: Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease. *Biomedicines* 11 (2023) 581

35. Qin, S., Hou, X., Wen, Y., Wang, C., Tan, X., Tian, H., Ao, Q., Li, J., Chu, S.: Machine learning classifiers for screening nonalcoholic fatty liver disease in general adults. *Scientific Reports* 13 (2023) 3638

36. Rahi, P., Hussain, M.R., Alqahtani, S., Husain, M., Singh, A.P., Singh, I.: A novel random forest-SMOTE framework with polynomial feature engineering for early detection of gastrointestinal disorders. *IEEE Access* 13 (2025) 188577–188604

37. Chuang, Y.-H., Su, J.-H., Han, D.-H., Liao, Y.-W., Lee, Y.-C., Cheng, Y.-F.: Effective natural language processing and interpretable machine learning for structuring CT liver-tumor reports. *IEEE Access* 10 (2022) 116273–116286

38. Drożdż, K., Nabrdalik, K., Kwiendacz, H. et al.: Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovascular Diabetology* 21 (2022) 240

39. Kang, B.E., Park, A., Yang, H., Jo, Y., Oh, T.G., Jeong, S.M., Ji, Y., Kim, H.-L., Kim, H.-N., Auwerx, J., Nam, S., Park, C.-Y., Ryu, D.: Machine learning-derived gut microbiome signature predicts fatty liver disease in the presence of insulin resistance. *Scientific Reports* 12 (2022) 21842

40. Rahi, P., Singh, I., Mahajan, S., Chauhan, R., Kaur, T., Singh, J.: From genes to symptoms: Predicting Huntington's disease progression with DNN. In: *Proc. Int. Conf. Electronics, AI and Computing (EAIC)*, Jalandhar, India (2025) 1–9

41. Castañé, H., Baiges-Gaya, G., Hernández-Aguilera, A. et al.: Coupling machine learning and lipidomics as a tool to investigate metabolic dysfunction-associated fatty liver disease: A general overview. *Biomolecules* 11 (2021) 473

42. Marjot, T., Webb, G.J., Barritt, A.S. IV et al.: COVID-19 and liver disease: Mechanistic and clinical perspectives. *Nature Reviews Gastroenterology & Hepatology* 18 (2021) 348–364

43. Bence, K.K., Birnbaum, M.J.: Metabolic drivers of non-alcoholic fatty liver disease. *Molecular Metabolism* 50 (2021) 101143

44. Rahi, P., Kang, S.S., Singh, A.P., Singh, I.: Advanced anxiety risk stratification using regularized deep neural networks. In: *Proc. Int. Conf. Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)*, Pattaya, Thailand (2024) 524–530

45. Hallaji, E., Razavi-Far, R., Palade, V., Saif, M.: Adversarial learning on incomplete and imbalanced medical data for robust survival prediction of liver transplant patients. *IEEE Access* 9 (2021) 73641–73650

46. Mahzari, A.: Artificial intelligence in nonalcoholic fatty liver disease. *Egyptian Liver Journal* 12 (2022) 69

47. Balsano, C., Burra, P., Duvoux, C., Alisi, A. et al.: Artificial intelligence and liver: Opportunities and barriers. *Digestive and Liver Disease* 55(11) (2023) 1455–1461

48. Singla, S., Rahi, P.: Clinically interpretable and statistically robust EHR-based ensemble model for anxiety screening. In: *Proc. 8th Int. Conf. Image Information Processing (ICIIP)*, Solan (Near Shimla), Himachal Pradesh, India (2025) 219–226

49. Salmi, M., Atif, D., Oliva, D. et al.: Handling imbalanced medical datasets: Review of a decade of research. *Artificial Intelligence Review* 57 (2024) 273

50. Kantartzis, K., Stefan, N.: Clustering NAFLD: Phenotypes of nonalcoholic fatty liver disease and their differing trajectories. *Hepatology Communications* 7(4) (2023) e0112

51. Barik, P.: Liver Diseases Dataset. Kaggle (2025).
Available online:
<https://www.kaggle.com/datasets/priyankabarik/liver-diseases>

BIOGRAPHIE



I am Kapil, and I completed my Bachelor's degree in Computer Science and Engineering in 2024. Currently, I am pursuing my Master's degree in the Department of Computer Science and Engineering at Sri Satya Sai Technical University. My areas of interest include artificial intelligence, machine learning, and emerging technologies. I am passionate about research and continuously improving my technical and analytical skills through academic and practical learning. experiences.