Empowering Diagnostics: An Ensemble Machine Learning Model for Early Liver Disease Detection

Abdulrahman Ahmed Jasim[1], **Hajer Alwindawi**[2], **Layth Rafea Hazim**[3]

[1]Dept. of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey
Email: abdulrahman.jasim@alumni.altinbas.edu.tr

[2]Dept. of Artificial Intelligence Engineering, Bahçeşehir University, Istanbul, Turkey
Email: hajir.alwindawi@bahcesehir.edu.tr

[3]Dept. of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey
Email: layth.hazim@alumni.altinbas.edu.tr

Article History

Received: Feb. 16, 2025

Revised: Apr. 30, 2025

Accepted: Jun. 02, 2025

Abstract

Early and accurate detection of liver disease is critical to improving patient outcomes yet remains challenging due to class imbalance and noisy clinical data. In this study, we present a robust ensemble learning framework applied to the Indian Liver Patient Dataset, incorporating systematic data cleaning, normalization, and Synthetic Minority Over-Sampling (SMOTE) to address missing values, outliers, and class skew. We then perform correlation-based feature reduction before training a stacking classifier that combines Random Forest, XGBoost, and ExtraTrees base learners with an ExtraTrees meta-learner. Using stratified 10-fold cross-validation on the balanced cohort ($n = 792$), our ensemble achieves 91.6 % accuracy, 92 % F1-score, and a high area under the ROC curve, outperforming individual models and prior published approaches. These results demonstrate the potential of heterogeneous ensembles for clinical decision support in hepatology and lay the groundwork for prospective validation in diverse patient populations.

Keywords- Telemedicine, Artificial Intelligence (AI), Machine Learning, Stacking Classifier, Liver Diseases.

I. INTRODUCTION

The incidence of liver disease in India is very high and continues to rise. This indicates a significant public health crisis. The liver is the largest intra-abdominal organ in the human body and is vital in the execution of important metabolic functions and therefore, the health of the liver is essential to overall well-being [1]. Hepatic illness is a leading cause both for morbidity and mortality worldwide and the rise in incidence in India is rapid with the incidence impacting one in five adults. Annually, liver disease in India claims 268,580 deaths and accounts for 3.17% of total deaths and 18.3% of global deaths due to liver diseases [2]. Hepatitis B and C pose significant public health burdens because of their potential to progress to chronic states, cirrhosis, and hepatocellular carcinoma. In the Indian scenario alone, it has been estimated that more than 40 million people are infected with HBV and 4.7 to 10 million are estimated to be carriers of HCV infection [3]. Despite the seriousness of the problem, control measures implemented, such as screening of blood products, observance of safe injection practices, and inclusion of the hepatitis B vaccine in the Universal Immunization Program, have resulted in significant progress in the control of these infections[4]. However, the pursuit of a 'zero-risk' goal, as in the case of moving towards eradication, continues to be an elusive dream [5].

The country is faced with the problem of HEV waterborne epidemicsdemanding a vigorous public health effort that focuses on supplying pure drinking water and safely handling sewage[6], [7]. With the availability of effective HEV vaccines, there is optimism for managing this pervasive virus [8]. Alcoholic liver disease, alongside non-alcoholic fatty liver disease, compounds the already susceptible status of liver function. The latter is a key factor in cirrhosis and liver cancer, while the former has become the leading liver ailment, affecting almost two billion people across the globe. NAFLD has a central role at the juncture where liver function overlaps with other primary determinants of morbidity and the strong emphasis for the necessity for alterations in lifestyle in conjunction with a research focus on potential therapies [9]. In this backdrop of desolation, liver transplantation presents itself as an icon of hope, with India performing around 1800 transplants every year. However, the procedure is plagued by hurdles in the form of economic impediments, restricted donor supply, and a large disparity between demand and supply of transplants. This calls for the creation of novel strategies in accordance with the diverse cultural and economic terrain of India [10]. Early diagnosis is crucial for managing life-threatening diseases, and ML techniques play a vital role in healthcare, particularly for the diagnosis and classification

http://doi.org/10.58564/IJSER.4.2.2025.314

of illnesses based on distinctive information [11], [12]. These systems assist clinicians in making accurate decisions for patient care. However, the raw feature space often contains irrelevant information and exhibits high dimensionality, posing challenges in conventional machine learning systems[13] [12].

Medical data often faces class imbalance, posing challenges for Projection-based scaling methods and machine learning algorithms may lead to potential overfitting [1]. The liver disease dataset faces a number of challenges, including having missing values, outliers, and a considerable class imbalance in which the positive class comprises almost twice the occurrences of the negative class. To overcome these challenges, several data preparation procedures are executed in our computer-aided diagnostic system. The recommended ensemble method seeks to improve the precision of liver disease detection, minimize misdiagnosis instances, and eventually increase patient survival rates[13].

The study utilized diverse advanced machine learning techniques on the Indian Liver Patient Dataset with the objective of creating an effective predictive model. To further increase the effectiveness and usability of the dataset in real-life scenarios, the study carried out extensive preprocessing procedures that included removing duplicate records and correcting missing data instances. In addition, the preprocessing steps addressed the problem of class imbalance commonly found in medical data sets through the use of Random Over-Sampling so that the class balance became equalized, a substantial enhancement in the model's applicability in practical situations.

II. LITERATURE REVIEW

Dritsas et al. [14] Employed Stochastic methods for feature extraction applied to identical datasets, using four machine learning algorithms in conjunction with an ensemble voting classifier. The proposed approach showed considerable potential, achieving an area under the curve (AUC) of 88.4%, so exceeding the performance of the other classifiers assessed in the research. Conversely, Straw et al.[15] Employed an alternative methodology for liver condition diagnosis, including a gender-stratified investigation, while using the SMOTE technique for data balance and Recursive Feature Elimination for feature selection.

Their investigation consisted of classifiers such as RF, LR, SVM, and GNB, emphasizing that RF and LR demonstrated superior performance, particularly for gender variations. Kumar and colleagues[16] focus on the Indian Liver Dataset, using neighbor-weighted K-NN classifiers. Their novel Variable-neighbor weighted fuzzy K nearest neighbor (NWFKNN) algorithm attained an accuracy of 87.71%, demonstrating the efficacy of their methodology in predicting liver illness. Gan and associates [17] initiated the use of four distinct classification methodologies: AdaC-TANBN, TANBN, BN, and SVM. Their distinctive methodology, namely the integrated TANBN using a cost-sensitive strategy (AdaC-TANBN), attained an accuracy of 69.03%, exceeding prior methods in their study. Sreejith and his associates[18] analyzed the liver illness dataset, Thoracic Surgery, and Pima Indian Diabetes datasets, focusing on the comparative classification performance pre- and post-feature selection.

The authors' approach, employing the Synthetic Minority Over-sampling Technique and Chaotic Multi-Verse Optimization for feature selection, resulted in a notable increase in accuracy. They attained an accuracy rate of 69.43% with a random forest classifier without employing SMOTE and CMVO, and 82.62% accuracy while utilizing SMOTE alone, therefore illustrating the impact of class balance and feature selection techniques. In a similar vein, Kuzhippallil et al. [19] conducted an extensive investigation aimed at comparing and enhancing the categorization efficacy of chronic liver disease. A variety of data preparation strategies were examined, including imputation of missing values, outlier identification, removal using isolation forest, and elimination of duplicate values. Furthermore, several feature selection methodologies were employed. Utilizing an array of classifiers, such as MLP, KNN, LR, DT, RF, Gradient Boosting, AdaBoost, XGBoost, LightGBM, and Stacking an estimator, they attained classification accuracies for liver patients ranging from 76% to 88% following the implementation of their proposed technique. Anagaw, A. et al. [20] introduced a technique termed complement naive Bayesian classification and evaluated its performance against the naive Bayes classifier and other models. The outcomes of their suggested methodology were promising, with an accuracy of 71.36%, exceeding the performance of the other classifiers analyzed in their study.

III. METHODOLOGY

This study utilizes a thorough methodological framework to create a strong predictive model for hepatic disease using the Indian Liver Dataset. This methodology is explicitly formulated to address critical issues in medical data analysis, including class imbalance and the need for effective feature selection, aiming to provide models for prediction with improved accuracy.

Figure 1 illustrates a detailed schematic of our methodical approach, showcasing the transition from fundamental dataset management to the final construction of the predictive model. The flowchart delineates essential phases in our technique, encompassing data pretreatment, feature extraction, model construction, and assessment. An examination of the graphic will familiarize readers with the sequential progression and the interrelation among the many phases of this study.

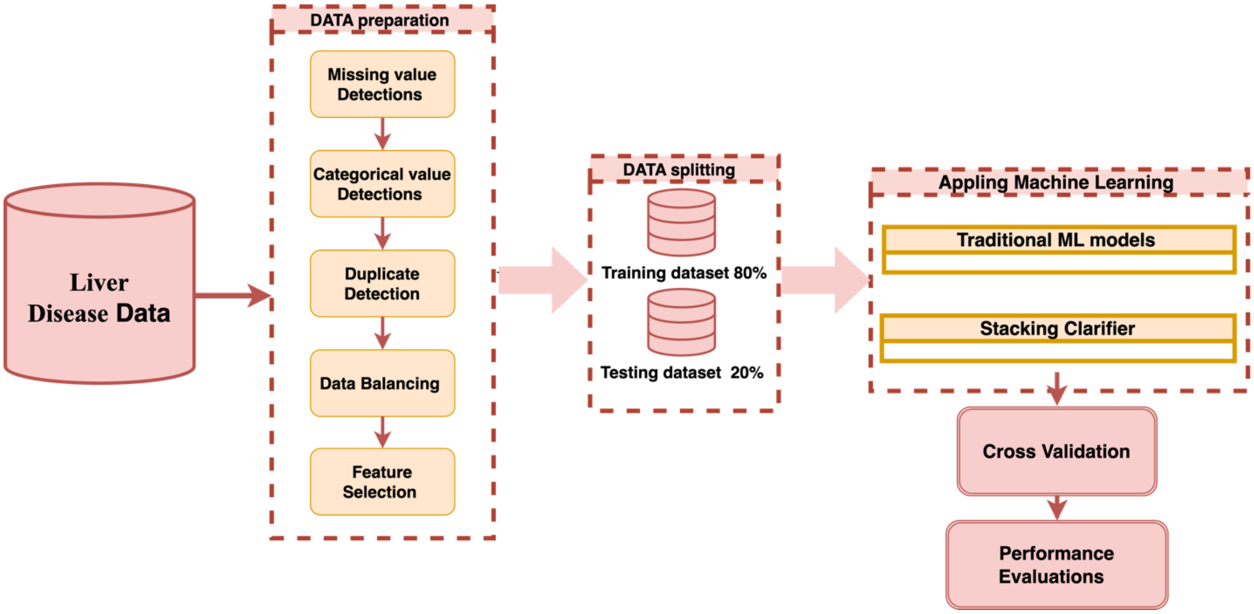

Figure 1. Flowchart of proposed Approach

A. Outliers

Figure 2 shows some examples of outliers in our dataset. These outliers can be due to data entry errors. Removing these outliers can help increase the efficacy of the prediction model. In the current study, a normalization process was applied to the dataset with the aim of scaling its values to a specific range of [0,1] to try to solve the problem of outliers.

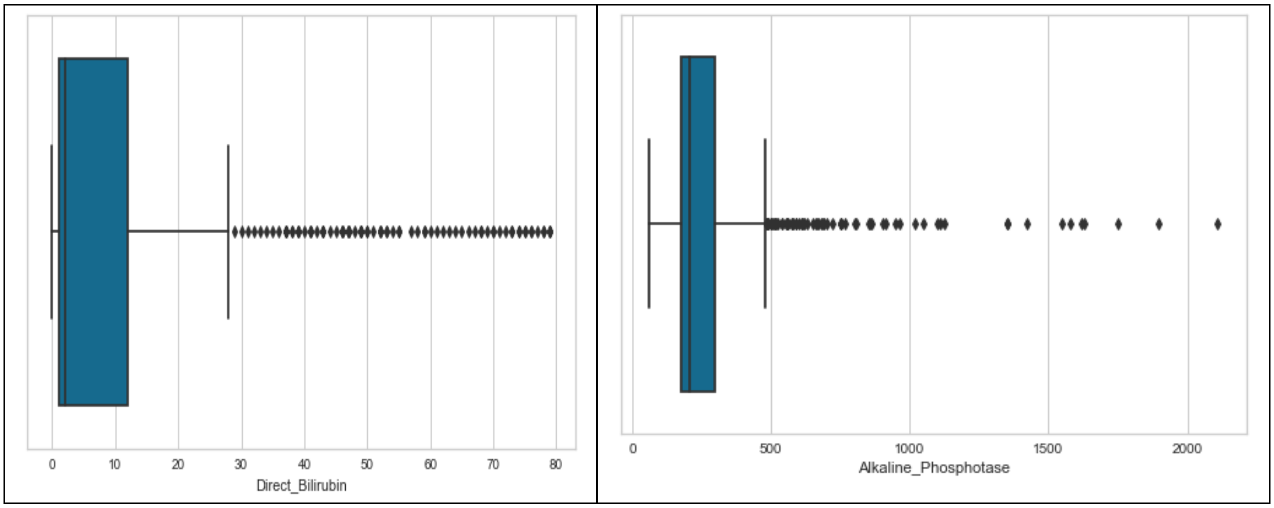

Figure 2. Data Outliers

B. Synthetic Minority Oversampling Technique (SMOTE)

Due to the considerable disparity seen across the classes, as seen in Figure 1, the implementation of tailored class balancing techniques is crucial. This study employs a substantial class balancing technique to tackle this particular issue. In all instances of class balancing, the modified dataset exhibits a class distribution ratio of 50:50, as seen in Figure 3. The Synthetic Minority Over-Sampling Technique is a technique designed to augment the sample size of the minority class. The production of synthetic data, essential to SMOTE, is performed using the K-nearest neighbor's approach. This technique facilitates the enhancement of examples in the minority class by the generation of synthetic data inside the feature space.

C. Dimensionality Reduction via Correlation Analysis

After addressing the class imbalance issue, an analysis of feature importance was conducted to assess the significance of each feature in predicting liver disease. Dimensionality reduction is a technique employed to eliminate variables that exert minimal influence on predictive outcomes. In this study, dimensionality reduction was achieved by conducting a correlation analysis on the features of the input data and then removing those with high variance. Figure 4 is a heat map used to measure the correlation of the features in the dataset. An interesting correlation was found between the attributes of "Total_Bilirubin," "Direct_Bilirubin," "Alamine_Aminotransferase," and "Alkaline_Phosphotase," as all these properties provide vital information related to liver pathology.

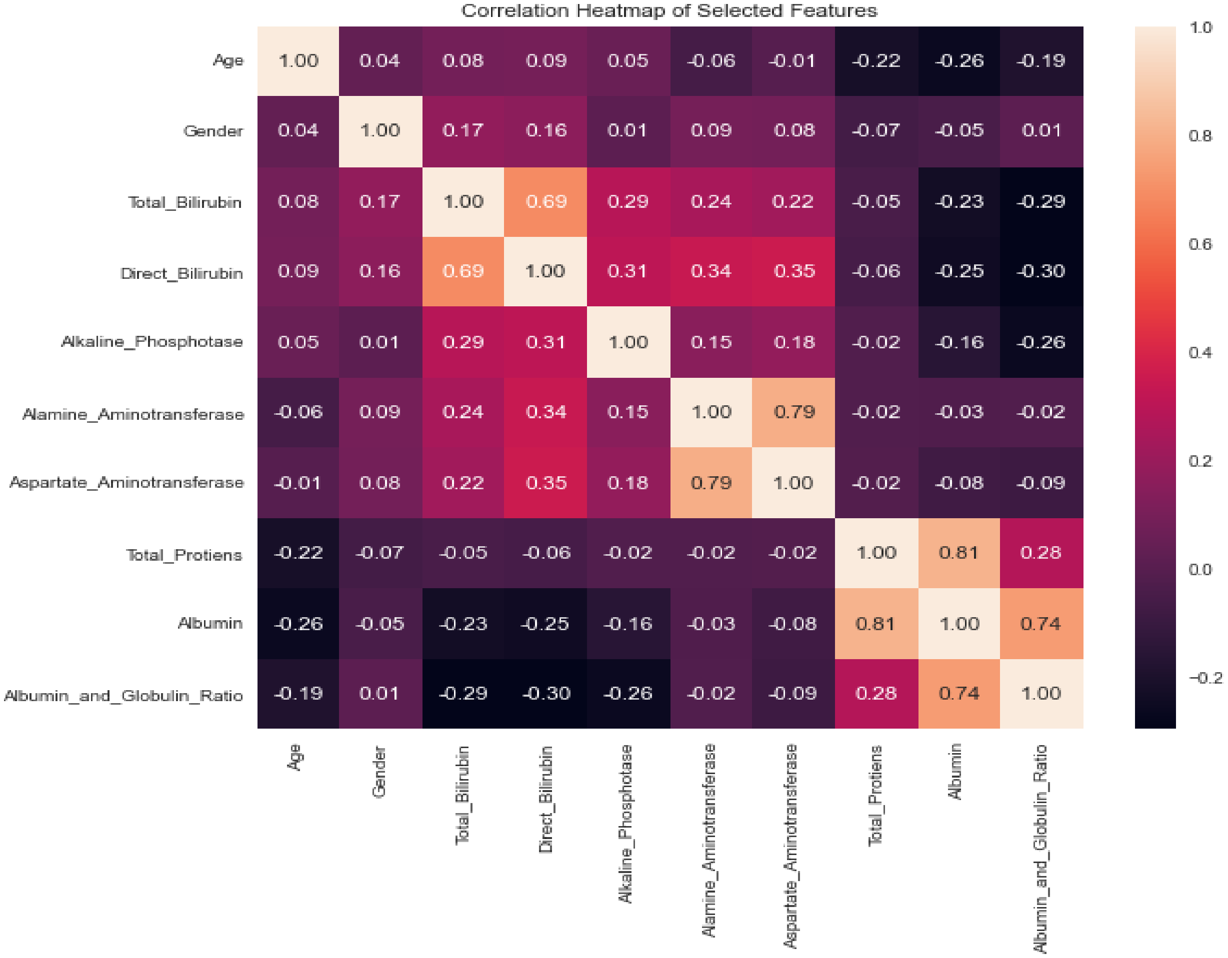

Figure 3. Correlation heatmap.

D. *Machine Learning Models*

In our effort to come up with a solid classification model, we carefully chose a line of machine learning algorithms, each with competence in handling classification issues. The rationale for employing any specific model lies in its distinctive characteristics and recorded efficacy in like tasks.

1. *Random Forest Model*

Random forest algorithm is recognized as a regression technique under the field of machine learning. During the training process, it creates decision trees based on liver data, ultimately leading to individual trees created using the partitioning process [19]. Its effectiveness is considered superior compared to other algorithms. The method can handle large datasets with many input variables without leaving any variable behind, thus classifying outputs efficiently and providing estimates of important variables related to the process of classification. Random Forests generate a high number of classification trees. Upon introducing a new liver sample as an input vector, it goes through each tree in the forest. Each tree assigns a class and votes for the specific class. The final classification is determined based on the majority of the trees' votes in the forest.

2. *XGBoost Model*

XGBoost is based on the gradient boosting principles and employs a boosting approach using a set of DT, like the boosted tree algorithm [21]. In this system, every tree learns from its previous trees and adjusts the residual errors, hence allowing the subsequent tree to learn from these adjusted residuals. XGBoost incorporates several advantages, such as the use of normalization, overfitting prevention mechanisms, and greater efficiency in handling sparse datasets.

3. *ExtraTrees Model*

The Extra Trees (ET) method produces A multitude of decision trees is generated by randomly picking subsets according to characteristics and input cases. The algorithm then amalgamates the forecasts derived from these several decision trees to formulate a unified prognosis[22]. A unique feature of ET is the use of randomness in the process of tree construction, as it uses random splits at each node, which helps in reducing variance and improving model accuracy compared to conventional decision tree methods. This method also helps in reducing overfitting, thus making ET a very robust and reliable machine learning algorithm. Especially in situations with a high feature dimensionality and large datasets, ET proves to be both efficient and easy to use in meeting analytical needs.

4. *Stacking Model*

Stacking is the process of combining predictions from the best of multiple basic algorithms for ML to form a new meta-learning algorithm. Its advantage is that it combines the ability of a group of high-performing approaches to provide forecasts that outperform any single model. The authors of [14] concluded that we may use stacking techniques to increase the performance of prediction models in the investigated circumstances. In this study 4 ML models (RF, DT, XGBoost, ExtraTrees) used as basic classifiers, for the meta-learner, several classifiers have been tested and performed. Finaly, ExtraTrees classifier was selected as the meta-learner[4].

IV. RESULTS AND DISCUSSION

Five artificial intelligence models, Random Forest (RF), XGBoost, Decision Tree (DT), Extra Trees, and Stacking, were evaluated using several performance measures, including accuracy, precision, recall, F1-score, and AUC. In the Stacking technique, the Random Forest, Decision Tree, XGBoost, and Extra Trees models acted as base classifiers, whilst the Extra Trees model served as the meta-classifier. The evaluation of model performance utilized a 10-fold cross-validation method on a balanced dataset of 792 cases following the implementation of the SMOTE approach. Following the identification of best features via correlation analysis, shown in the Heatmap, Table 1 illustrates the evolutionary performance of several machine learning algorithms, comparing their efficacy with and without the use of the SMOTE method. The deployment of machine learning models showed substantial enhancements attributable to the SMOTE approach, particularly when integrated with 10-fold cross-validation, as seen by the pronounced impact of the class balancing technique on recall and accuracy metrics. It is crucial to highlight here that uniform stratification played a part in contributing towards increased accurate detection of both conditions, with maintaining average performance at high levels. This confirms that the models in question retained their effectiveness in detecting patients with precision and accuracy.

The ET model surpassed the other models in performance without the use of the SMOTE approach, achieving accuracy, precision, recall, and F1-score values of 71%, 77%, 87%, and 81%, respectively. Conversely, the stacking model exceeded all evaluated metrics when the SMOTE approach was used, with all metrics over 90%.

TABLE I ILLUSTRATES THE EFFICACY OF THE MACHINE LEARNING MODELS USING 10-FOLD CROSS-VALIDATION, BOTH BEFORE TO AND AFTER TO THE USE OF SMOTE.

ML Models	Before Balancing Dataset				After Balancing Dataset			
	Accuracy	Precision	Recall	F1_score	Accuracy	Precision	Recall	F1_score
ET	**71**	**77**	**87**	**81**	**88**	**92**	**82**	**87**
RF	**69**	**75**	**83**	**79**	**86**	**94**	**77**	**84**
XGB	**67**	**76**	**80**	**78**	**86**	**94**	**76**	**84**
Stacking	**67**	**74**	**82**	**78**	**91.6**	**90**	**92**	**92**

In addition, Table 2 displays the models proposed in line with current research based on the same data set. The model proposed by us, named Stacking, outperforms previous attempts with the achievement of an accuracy rate of 91.6%. Recent studies, for example, saw the introduction of SMOTE, in conjunction with Chaotic Multiverse Optimization, constitutes the feature selection methodology presented by Sreejith et al. [14]. Kumar et al. [12] used Tomek-linked and redundancy-based under-sampling (TL_RUS) with the

Variable-Neighbor Weighted Fuzzy K-Nearest Neighbor method (Variable-NWFKNN). Straw and colleagues [11] used SMOTE with recursive feature elimination (RFE). Dritsas et al. [10] used SMOTE for resampling the Liver Disease dataset.

TABLE II. COMPARATIVE EXAMINATION WITH PRIOR RESEARCH ON LIVER DISEASE DIAGNOSIS.

Previous Studies	**Splitting**	**Model**	**Accuracy**
[18]	**Train-Test (30% / 70%)**	**RF**	**82.46%**
[16]	**10-fold cross-validation**	**Variable-NWFKNN**	**87.71%**
[15]	**Train-Test (30% / 70%)**	**SVM**	**83.30%**
[14]	**10-fold cross-validation**	**Voting**	**80.1%**
Proposed	**10-fold cross-validation**	**Stacking**	**91.6%**

V. CONCLUSION

This work advances liver disease diagnostics by integrating rigorous preprocessing, targeted oversampling, and a novel stacking ensemble to deliver superior predictive performance on the Indian Liver Patient Dataset. Our model's 91.6 % accuracy and strong F1-score confirm that combining complementary algorithms can effectively mitigate class imbalance and enhance generalizability. Beyond outperforming existing single-model and voting-based methods, this framework emphasizes reproducibility through stratified cross-validation and transparent feature selection. Future efforts will focus on external validation with independent cohorts, incorporation of clinical risk factors, and deployment in real-time telemedicine settings to assess the impact on early intervention and patient management.

REFERENCES

[1] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms," *Inform Med Unlocked*, vol. 36, p. 101155, Jan. 2023, doi: 10.1016/J.IMU.2022.101155.

[2] H. R. Singh and S. Rabi, "Study of morphological variations of liver in human," *Translational Research in Anatomy*, vol. 14, pp. 1–5, Mar. 2019, doi: 10.1016/J.TRIA.2018.11.004.

[3] M. P. Manns *et al.*, "Hepatitis C virus infection," *Nature Reviews Disease Primers 2017 3:1*, vol. 3, no. 1, pp. 1–19, Mar. 2017, doi: 10.1038/nrdp.2017.6.

[4] A. Ahmed Jasim, L. Rafea Hazim, H. Alwindawi, and O. Ata, "Optimizing Prediction of Cardiac Conditions Using Hyper-Adaboost-Integrated Machine Learning Models," *Journal for Scientific Engineering Research*, vol. 3, no. 3, 2024, doi: 10.58564/IJSER.3.3.2024.220.

[5] M. F. Yuen *et al.*, "Hepatitis B virus infection," *Nature Reviews Disease Primers 2018 4:1*, vol. 4, no. 1, pp. 1–20, Jun. 2018, doi: 10.1038/nrdp.2018.35.

[6] H. Mohammedqasim, A. A. Jasim, R. Mohammedqasem, and B. A. Ozturk, "Advancing Parkinson's Disease Detection: Integrating Machine Learning with Enhanced Feature Selection and Data Augmentation," *Lecture Notes in Networks and Systems*, vol. 1085 LNNS, pp. 451–465, 2024, doi: 10.1007/978-981-97-6726-7_36.

[7] A. A. Jasim, O. Ata, and O. H. Salman, "AI-Driven Triage: A Graph Neural Network Approach for Prehospital Emergency Triage Patients in IoMT-Based Telemedicine Systems," *2024 International Symposium on Electronics and Telecommunications (ISETC)*, pp. 1–7, Nov. 2024, doi: 10.1109/ISETC63109.2024.10797314.

[8] M. S. Khuroo, "Discovery of Hepatitis E and Its Impact on Global Health: A Journey of 44 Years about an Incredible Human-Interest Story," *Viruses 2023, Vol. 15, Page 1745*, vol. 15, no. 8, p. 1745, Aug. 2023, doi: 10.3390/V15081745.

[9] F. Idalsoaga, A. V. Kulkarni, O. Y. Mousa, M. Arrese, and J. P. Arab, "Non-alcoholic Fatty Liver Disease and Alcohol-Related Liver Disease: Two Intertwined Entities," *Front Med (Lausanne)*, vol. 7, p. 556724, Aug. 2020, doi: 10.3389/FMED.2020.00448/BIBTEX.

[10] "Kashmir in Sickness and in Health - Gulzar Mufti - Google Books." Accessed: Jan. 20, 2024. [Online]. Available: https://books.google.com.tr/books?hl=en&lr=&id=AoAcAgAAQBAJ&oi=fnd&pg=PT4&dq=Liver+diseases+in+India:+hope+and+despair.+Greater+Kashmir&ots=6zwOqEOS0i&sig=6yfuRFLWkXsysU-96cqwTW9GyVM&redir_esc=y#v=onepage&q&f=false

[11] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," *Inform Med Unlocked*, vol. 17, p. 100255, Jan. 2019, doi: 10.1016/J.IMU.2019.100255.

[12] S. Ullah, M. D. Awan, and M. Sikander Hayat Khiyal, "Big Data in Cloud Computing: A Resource Management Perspective," *Sci Program*, vol. 2018, 2018, doi: 10.1155/2018/5418679.

[13] S. S. Mohsin *et al.*, "AI-Powered IoMT Framework for Remote Triage and Diagnosis in Telemedicine Applications," *Al-Iraqia Journal for Scientific Engineering Research*, vol. 4, no. 1, pp. 61–76, Mar. 2025, doi: 10.58564/IJSER.4.1.2025.294.

[14] E. Dritsas and M. Trigka, "Supervised Machine Learning Models for Liver Disease Risk Prediction," *Computers 2023, Vol. 12, Page 19*, vol. 12, no. 1, p. 19, Jan. 2023, doi: 10.3390/COMPUTERS12010019.

[15] I. Straw and H. Wu, "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction," *BMJ Health Care Inform*, vol. 29, no. 1, p. 100457, Apr. 2022, doi: 10.1136/BMJHCI-2021-100457.

[16] P. Kumar and R. S. Thakur, "Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach," *Multimed Tools Appl*, vol. 80, no. 11, pp. 16515–16535, May 2021, doi: 10.1007/S11042-019-07978-3/FIGURES/5.

[17] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," *Comput Ind Eng*, vol. 140, p. 106266, Feb. 2020, doi: 10.1016/J.CIE.2019.106266.

[18] S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection," *Comput Biol Med*, vol. 126, p. 103991, Nov. 2020, doi: 10.1016/J.COMPBIOMED.2020.103991.

[19] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients," *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, pp. 778–782, Mar. 2020, doi: 10.1109/ICACCS48705.2020.9074368.

[20] A. Anagaw and Y. L. Chang, "A new complement naïve Bayesian approach for biomedical data classification," *J Ambient Intell Humaniz Comput*, vol. 10, no. 10, pp. 3889–3897, Oct. 2019, doi: 10.1007/S12652-018-1160-1/TABLES/8.

[21] M. Banu Priya, P. Laura Juliet, and P. R. Tamilselvi, "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms," *International Research Journal of Engineering and Technology*, 2018, Accessed: Jan. 21, 2024. [Online]. Available: www.irjet.net

[22] T. Sumathi, V. Pragadeeswaran, and C. Chethankumar, "Prediction of Chronic Liver Cirrhosis Using Ensemble Classification Approach," *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, 2023, doi: 10.1109/ICCCNT56998.2023.10307455.