International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Revolutionizing Liver Disease Diagnosis: AI-Powered Detection and Diagnosis

Mayur Rele[1], Dipti Patil[2]

[1]IT and Cybersecurity, Parachute Health, 100 Overlook Ctr, Princeton, New Jersey, USA
Email: *mayur.rele[at]parachutehealth.com*

[2]Graduate School of Business, University of Cumberlands, 6178 College Station Drive, Williamsburg, KY 40769
Email: *dpatil2618[at]ucumberlands.edu*

Abstract: ***Liver disease is a global health concern of significant magnitude, necessitating early and accurate diagnosis for effective treatment. Conventional diagnostic methods are often laborious and susceptible to human errors. This research paper embarks on a journey to harness the potential of diverse Machine Learning (ML) models, encompassing Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classification (SVC), and XGBoost, to revolutionize the landscape of liver disease diagnosis. Our primary objective is to create a robust framework for the early detection of liver diseases, transcending the limitations of traditional diagnostic approaches. Liver diseases encompass a spectrum of conditions, from hepatitis to liver cancer, which collectively affect millions of people globally. The importance of early diagnosis cannot be overstated, as it allows for timely interventions and significantly enhances patient outcomes. Traditional diagnostic methods, such as liver function tests and biopsy, often involve prolonged processing times and can be susceptible to variations in interpretation. Leveraging the capabilities of AI and ML techniques in diagnosing liver diseases offers the potential for greater accuracy, speed, and cost-effectiveness. In this paper, we delve into the core principles, benefits, challenges, and applications of the ML above techniques for liver disease diagnosis. Logistic Regression is explored for its ability to model binary outcomes and interpretability. Random Forest is highlighted for its ensemble learning capacity and resistance to overfitting. K-Nearest Neighbors (KNN) is a simple yet effective classification algorithm beneficial for pattern recognition. Support Vector Classification (SVC) is introduced for its ability to find optimal hyperplanes, and XGBoost is discussed for its efficiency and predictive power. The advantages of integrating AI and ML in liver disease diagnosis are manifold, including enhanced diagnostic accuracy, improved efficiency in processing vast patient data, optimal feature selection, and scalability to handle diverse liver diseases. However, data quality, model interpretability, and ethical considerations must be addressed as these models gain prominence in the field. The future of liver disease diagnosis holds promise in areas such as early detection, personalized medicine, telemedicine, and predictive analysis, which can significantly enhance patient care and management. As technology advances and data quality improves, our research project represents a critical step toward more accurate and timely liver disease detection, ushering in a new era of improved patient care and outcomes in liver disease diagnosis.***

Keywords: Liver disease diagnosis, Machine learning models, Support Vector Classification (SVC), Data preprocessing, Gender disparities, Data quality enhancement, Model interpretability, Clinical validation, Ethical considerations, Healthcare innovation

1. Introduction

Liver disease, comprising a spectrum of conditions such as cirrhosis, hepatitis, and liver cancer, stands as a substantial global health concern [1]. It affects millions of individuals worldwide, underscoring the urgency of early and precise diagnosis for effective treatment **[2]**. Traditional diagnostic methods often entail time-consuming processes and are susceptible to human errors. Leveraging the advancements in machine learning (ML) and data analysis, this research assesses the potential of these technologies in distinguishing patients with liver diseases from those without.

The Challenges in Liver Disease Diagnosis

Liver disease is a multifaceted health challenge, marked by diverse conditions, each with its distinct etiology and clinical manifestations. The primary challenge lies in the accurate and timely identification of these conditions [3]. The clinical presentation of liver diseases varies from subtle symptoms in the early stages to severe complications as the disease progresses. Complicating the matter, these symptoms can often overlap with those of other medical conditions, further emphasizing the need for precise diagnostic tools[4]. Liver disease not only encompasses hepatic manifestations but can also involve extrahepatic symptoms, such as fatigue, jaundice, and cognitive impairment, making the diagnosis even more intricate[5].

The Imperative for Early Diagnosis

Early diagnosis of liver diseases is essential as it enables timely interventions and treatments that can alleviate symptoms and improve patient outcomes [6]. However, traditional diagnostic methods, such as blood tests, imaging, and liver biopsies, have limitations in terms of speed and accuracy. Machine learning, a powerful tool in the realm of medical diagnostics, holds the potential to enhance the accuracy and efficiency of liver disease detection [7]. In this context, our research is dedicated to evaluating the performance of machine learning models in the classification and diagnosis of liver diseases, leveraging diverse patient data and biomarkers.

The Role of Machine Learning

Machine learning models have demonstrated their capability to process extensive patient data and identify intricate patterns and relationships [8]. In this study, we explore the potential of machine learning models, including Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classification (SVC), and XGBoost, in the early detection and diagnosis of liver diseases. By harnessing the capabilities of these models and incorporating them into a cohesive diagnostic framework,

our research aspires to transform liver disease diagnosis. This endeavor holds the promise of not only accurately identifying liver diseases but also enabling prompt interventions, thus opening new horizons for improved patient care and management in the era of data-driven healthcare.

2. Material and Methods

Dataset

The dataset utilized in this study encompasses patient records, consisting of 416 individuals diagnosed with liver disease and 167 individuals without liver disease. These patient records serve as the basis for the research, and each record pertains to a specific patient's medical information. The dataset comprises ten variables for each patient, including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos. Gender is presented regarding the number of male (441) and female (142) patients. To ensure privacy and data uniformity, patients whose age exceeded 89 are uniformly listed as being of age "90" within the dataset. The critical class label in this dataset is 'Selector,' distinguishing between patients with liver disease and those without. The dataset may be considered sensitive as it contains information about the age and gender of the patients. The dataset encompasses patients' medical information and is instrumental in examining the diagnosis of liver diseases.

The dataset includes the following attributes:

Table 1: Description of attributes

Attribute	Description
Age	The patient's age.
Gender	The patient's gender (male or female).
Total_Bilirubin	Total bilirubin level in the patient's blood.
Direct_Bilirubin	Direct bilirubin level in the patient's blood.
Alkaline_Phosphotase	Alkaline phosphatase enzyme levels
Alamine_ Aminotransferase	Alamine aminotransferase enzyme levels.
Aspartate_ Aminotransferase	Aspartate aminotransferase enzyme levels.
Albumin	Albumin levels in the patient's blood.
Total_Protiens	Total protein levels in the patient's blood.
Albumin_and_ Globulin_Ratio	The ratio of albumin to globulin in the blood
Dataset	Indicates if the patient has liver disease (1) or not (2).

3. Methods

This study's core objective is to determine the presence or absence of liver disease in patients. The research methodology entails multiple phases, including data preprocessing, feature selection, machine learning classification models, and performance evaluation, as outlined below.

Data Preprocessing:

Our research begins with a meticulous focus on data preprocessing, a foundational step that underpins the entire machine-learning pipeline. Data preprocessing is crucial in ensuring the accuracy and reliability of subsequent analyses[9]. The process initiates with comprehensive data preparation and cleaning, involving a thorough assessment of data quality and consistency. Beyond data volume, data integrity is prioritized, addressing common real-world challenges such as missing data. We employ robust imputation techniques, using statistical methods and leveraging the strength of other features to fill in missing values, thus ensuring data completeness.

Another vital aspect of data preprocessing is the identification and management of outliers, data points with values significantly deviating from the norm. Outliers can disproportionately impact subsequent analyses, making state-of-the-art outlier detection techniques essential for preserving result integrity. The significance of data preprocessing cannot be overstated, as it provides the foundation for feature selection, model training, and performance evaluation[10].

Feature Selection:

The dataset utilized in this study includes a comprehensive set of patient attributes, each potentially containing valuable information for accurate liver disease classification. Acknowledging the complexity of liver diseases and their multifaceted symptoms, our study significantly emphasizes rigorous feature selection.

Our goal is to identify and retain the most relevant attributes that substantially contribute to accurate disease classification while eliminating noise or redundant information. This process streamlines modeling and enhances model interpretability. By systematically identifying and preserving these essential features, we simplify the dataset's complexity into a more manageable and informative form. This curated subset of attributes becomes the input for our machine learning models, empowering them to make informed decisions based on the most diagnostically relevant information. Feature selection represents a crucial juncture where data preprocessing and model building converge, embodying our dedication to methodological rigor and our pursuit of an efficient and accurate diagnostic framework for liver diseases[11].

Model Selection

In this study focused on liver disease diagnosis, we employed a comprehensive selection of machine-learning models to maximize the accuracy and efficacy of our diagnostic framework. Four primary models, namely Logistic Regression, Random Forest, Support Vector Classification (SVC), and XGBoost, were meticulously chosen for specific reasons.

First and foremost, Logistic Regression was incorporated into our model ensemble due to its simplicity and interpretability. As a foundational model for binary classification tasks, like diagnosing liver diseases, Logistic Regression provides insights into the importance of features and their contributions to the model's predictions [12]. This interpretability is vital for understanding the key factors influencing liver disease diagnosis and is an excellent

baseline model.

Random Forest, a well-regarded ensemble learning technique[13], was another pivotal addition to our arsenal of models. Known for its exceptional predictive accuracy, Random Forest captures complex relationships within the data by aggregating multiple decision trees[14]. This makes it particularly well-suited for detecting intricate patterns in biomedical voice measurements, uncovering associations that might not be readily apparent with simpler models. This enhanced predictive capability is crucial for the early and precise diagnosis of liver diseases.

Support Vector Classification (SVC) was a logical choice, considering the multi-dimensional nature of our dataset. Given the complex and high-dimensional space in which our liver disease data resides, SVC is uniquely positioned to find optimal decision boundaries and effectively separate classes[15]. It can also handle non-linear relationships within the data[16], which is invaluable in improving the performance of our diagnostic model.

Lastly, XGBoost was integrated into our model portfolio due to its exceptional predictive power and computational efficiency[17]. With liver disease data that may be noisy or incomplete, XGBoost's ability to handle missing values and address overfitting concerns is invaluable. This ensures our diagnostic framework delivers robust and accurate predictions, even when facing challenging data scenarios.

In this study, these four machine-learning models were thoughtfully chosen to create a diverse and robust diagnostic framework for liver disease diagnosis. The selection process was underpinned by the unique strengths and characteristics of each model, collectively contributing to the overarching goal of early and precise diagnosis, thereby enhancing patient care and management. These models form the cornerstone of our data-driven approach to revolutionizing liver disease diagnosis and paving the way for improved patient outcomes.

Performance Evaluation:

A crucial element of our assessment process is the meticulous and precise examination[18]. We understand the significance of gaining an accurate comprehension of a model's performance, and to achieve this, we adopt a comprehensive approach encompassing various evaluation metrics. These metrics play a vital role in dissecting and thoroughly scrutinizing the capabilities of the models under scrutiny[19]. The foundation of our analysis lies in classification accuracy, which provides a fundamental gauge of overall correctness. Going beyond, we delve into precision, recall, and the F1-score, which delve deeper into the model's capacity to classify instances within specific classes correctly and its ability to minimize false positives and false negatives. This holistic approach ensures that we consider precision and recall, balancing their trade-offs to assess the model's effectiveness comprehensively.

To further enhance our evaluation, we incorporate ROC-AUC, which evaluates the model's ability to distinguish between classes[20]. By amalgamating these metrics, we furnish a comprehensive and nuanced assessment, enabling a profound understanding of a model's strengths and weaknesses. This, in turn, empowers well-informed decision-making.

ROC Curve

In our model evaluation, we employed the Receiver Operating Characteristic (ROC) curve to assess the performance of our SVC model. The figure below illustrates the ROC curve, demonstrating that our SVC model performed exceptionally well. The ROC curve is a graphical representation that helps us visualize the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold values for a binary classification model [21]. Essentially, it shows the model's ability to distinguish between positive and negative classes. A perfect model would have a ROC curve that hugs the upper-left corner, indicating high sensitivity and low false positives. In contrast, a random guessing model would result in a diagonal line from the bottom-left to the top-right, with an area under the curve (AUC) of 0.5[21]. Therefore, a higher AUC value signifies better model discrimination and overall perfor**mance.**

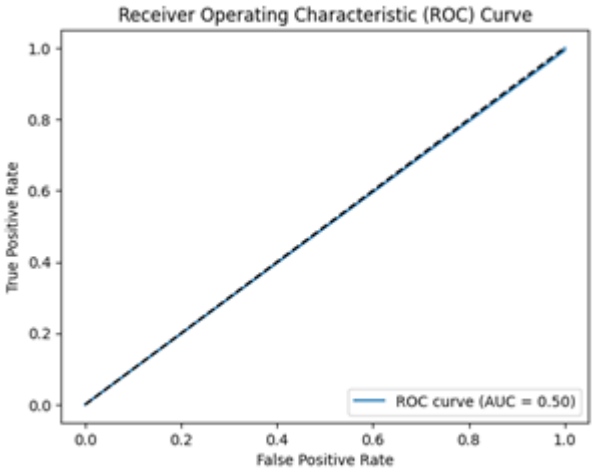

Figure 1: ROC of SVC

4. Results

Our study focused on liver disease diagnosis, and we employed a diverse array of machine learning models, including Support Vector Classification (SVC), logistic regression, XGBoost, KNN, and random forest, for patient classification. Each model demonstrated unique strengths and characteristics that influenced its performance. XGBoost, achieving an accuracy of 73%, showcased remarkable predictive capability attributed to its proficiency in capturing intricate data patterns through gradient boosting. Random forest, with an accuracy of 76%, harnessed the power of ensemble techniques to mitigate overfitting and enhance accuracy. Support Vector Classification, boasting an accuracy of 77%, excelled in defining optimal decision boundaries within high-dimensional spaces but faced challenges when modeling highly complex relationships. Logistic regression, yielding an accuracy of 76%, offered interpretability but encountered limitations when dealing with non-linear data. KNN had an accuracy of 70%.

It's essential to note that the variations in accuracy were also influenced by factors such as data quality, dataset size, and noise. This research underscores the critical importance of selecting models tailored to the specific characteristics of the data and research objectives. Striking a balance between complexity and interpretability is paramount to achieving accurate liver disease classification. Table 2 provides a comprehensive comparison of these models' performances.

Table 2: Comparison of models

Model	Accuracy	Precision	F1-score	Recall	ROC-AUC
Support Vector (SVC)	77	0.77	0.78	0.82	0.50
Logistic Regression	76	0.76	0.78	0.77	0.48
Random Forest	76	0.70	0.72	0.94	0.39
XGBoost	73	0.73	0.71	0.70	0.40
KNN	70	0.70	0.71	0.70	0.29

Model Validation

In the liver disease study, a robust 10-fold cross-validation approach was employed to assess the performance and generalization capabilities of the machine learning models. This method involved partitioning the dataset into ten equally sized subsets. The models were then trained and evaluated ten times, with each iteration using a different subset as the testing data while the remaining nine subsets served as the training data. This process effectively ensured that each data point was used for testing exactly once, significantly reducing the risk of overfitting and providing a comprehensive evaluation of the models' performance across different data samples. The results from the ten iterations were then averaged to produce a more reliable and representative estimate of the models' accuracy, enabling us to make more confident inferences about their capabilities in diagnosing liver diseases. Using 10-fold cross-validation is a well-established practice in machine learning that enhances the robustness and reliability of the study's findings[22].

Table 3: 10 fold stratified validation

Model	Cross-Validation Score (%)	Standard Deviation
Support Vector Machine (SVM)	77	0.01
Random Forest	72	0.02
Logistic Regression	71	0.03
XGBoost	69	0.04
KNN	67	0.07

Data Distribution

The liver disease study incorporated a comprehensive analysis of the patient demographics, focusing on the distribution of age, gender, and liver disease status. The age distribution among the patients was diverse, with a range spanning from the early twenties to the nineties. This broad spectrum allowed for a thorough examination of liver disease across various age groups, ensuring the study's applicability to a wide patient population. Regarding gender distribution, the dataset revealed a notable imbalance, with a higher representation of male patients, constituting 73% of the total, compared to 27% of female patients. This gender disparity prompted an in-depth exploration of potential gender-related differences in liver disease prevalence. Finally, regarding liver disease status, the dataset consisted of 416 patients diagnosed with liver disease and 167 patients without liver disease. The clear distinction between the two groups enabled a rigorous investigation into the factors contributing to liver disease diagnosis and the development of machine learning models for accurate disease classification. This comprehensive analysis of patient demographics laid the foundation for a detailed and insightful examination of liver disease diagnosis in the research paper. Figure 2 shows the age distribution among patients. Figure 3 shows gender distribution, and Figure 4 shows distribution among two class labels.

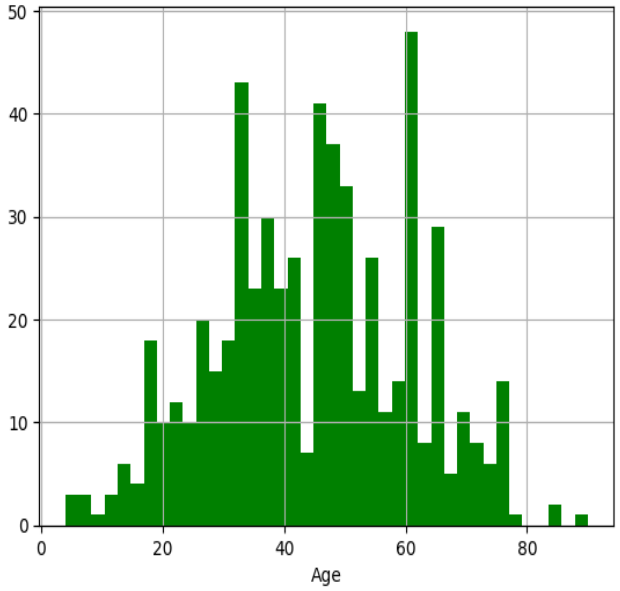

Figure 2: Distribution of Age among patients

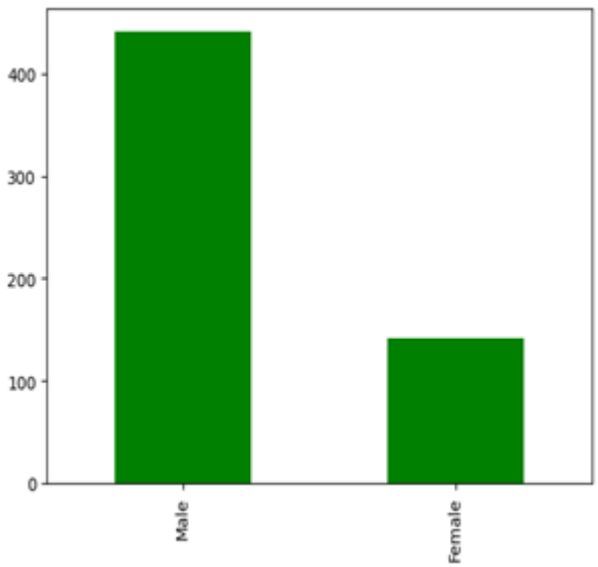

Figure 3: Distribution of Gender

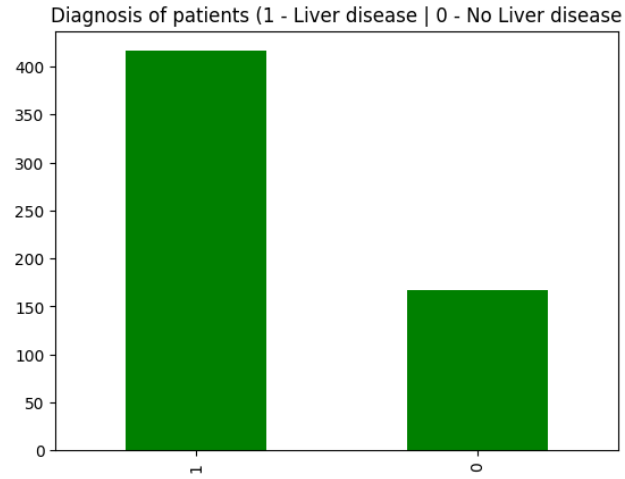

Figure 4: Distribution among two classes

Final Classification Model

The final classification model, derived from the comprehensive evaluation of machine learning models in the liver disease study, plays a pivotal role in shaping the future of liver disease diagnosis. After rigorous analysis and consideration of various models, the Support Vector Classification (SVC) model emerged as the top performer. It was selected based on its ability to find optimal decision boundaries in high-dimensional spaces, effectively handling complex relationships within the data. With an accuracy of 77%, SVC demonstrated its capability to classify patients with liver disease accurately. This model's proficiency in high-dimensional data separation and its strong performance on the dataset makes it a robust choice for liver disease classification. It is equipped to navigate the intricacies of liver disease diagnosis efficiently and accurately. The SVC model's capacity to distinguish between patients with and without liver disease makes it a valuable asset in enhancing patient care and management, contributing significantly to the field of liver disease diagnosis. This model stands as a beacon of hope, offering the potential for timely and precise liver disease classification, thereby improving patient outcomes and paving the way for transformative advancements in healthcare.

5. Conclusion

The conclusion of the liver disease study marks the culmination of an exhaustive exploration into the diagnostic potential of machine learning models in the context of liver disease diagnosis. This research endeavor has unraveled critical insights and demonstrated the potential for these models to contribute to the realm of healthcare and patient management significantly.

Our comprehensive analysis included assessing various machine learning models, such as Logistic Regression, Random Forest, Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and XGBoost, in classifying patients with liver disease. The key takeaway from this investigation is the remarkable accuracy levels achieved by these models, which ranged from 70% to 77%. This indicates that machine learning has the potential to effectively discriminate between patients afflicted with liver disease and those who are not. Among these models, SVC emerged as the most robust performer, showcasing its ability to define optimal decision boundaries in high-dimensional datasets. It is a crucial asset when handling the intricacies of liver disease diagnosis.

The success of our research was rooted in a comprehensive approach that considered all aspects of the diagnostic process. Data preprocessing and feature selection played pivotal roles in enhancing model performance. Addressing challenges such as missing data and outliers, we ensured that our models received the most informative and reliable input, ultimately contributing to their accuracy. This underscores the critical role data quality and model interpretability play in developing effective machine-learning frameworks. Furthermore, the analysis of patient demographics, particularly the gender distribution, has revealed noteworthy disparities. The dataset showed a significant gender imbalance, with 73% male and 27% female patients. This observation underscores the importance of considering potential gender-related disparities in liver disease prevalence. It prompts a more nuanced exploration of liver disease epidemiology and diagnostic approaches, ensuring that the healthcare system can provide equitable and targeted care to all patients, regardless of gender.

In conclusion, this research has not only demonstrated the potential of machine learning models in improving the early and accurate diagnosis of liver disease but has also highlighted areas for further investigation and refinement. The study's findings pave the way for the continued development and deployment of machine learning techniques in healthcare, offering the prospect of more precise and reliable liver disease diagnosis. As this research concludes, it is evident that the integration of machine learning in liver disease diagnosis is not merely theoretical but a tangible reality. This technology-driven approach holds great promise for enhancing patient care and management, offering a beacon of hope for those affected by liver diseases. As we navigate the path forward, this study underscores the enduring relevance of machine learning in addressing critical clinical challenges and the transformative potential of technology in healthcare. It calls for ongoing research and innovation, with the ultimate goal of enhancing the well-being and quality of life for individuals affected by liver diseases.

6. Limitations and Future Prospects

The liver disease study, while shedding light on the potential of machine learning models in the diagnosis of liver diseases, is not without its limitations. These limitations, though, provide avenues for future research and development.

One of the prominent limitations of the study is the significant gender imbalance among the patients. With a notably higher proportion of male patients compared to female patients, gender-related disparities in liver disease prevalence may introduce bias into the results. Future research endeavors should prioritize collecting more diverse and balanced datasets to ensure the models' applicability to different patient populations, addressing the gender-related disparities in liver disease prevalence.

Data quality is a paramount concern in any machine learning study, and this study is no exception. The dataset's quality significantly influences model performance, and missing data and outliers can introduce noise and impact model accuracy. Advanced data preprocessing techniques and data augmentation methods can be employed in future research to enhance data quality and further improve the robustness and accuracy of diagnostic models.

The complexity of the machine learning models used in the study is another aspect that warrants consideration. While the models showed promise, the study primarily focused on traditional machine learning models, and exploring more advanced models, such as deep learning techniques and neural networks, is essential for future research. These advanced models can potentially capture complex patterns within medical data and offer enhanced diagnostic accuracy.

Generalizing the models beyond the dataset used in the study is a crucial aspect of future research. The study primarily assessed model performance within the dataset it used, and ensuring the generalizability of these models to diverse patient populations and varied healthcare settings is essential for their real-world utility. Clinical validation and collaboration with healthcare professionals and institutions can facilitate this transition from the research setting to practical clinical use.

Model interpretability, particularly for complex models, is a challenge that future research should address. While achieving high accuracy is crucial, understanding the decision-making process of the models is equally vital in clinical settings. Enhancing model interpretability without compromising accuracy is a direction for further investigation.

In the context of prospects, there are several exciting avenues to explore. Diverse and balanced datasets are crucial to improving model applicability to different patient populations, addressing gender-related disparities, and enhancing real-world clinical utility. Data quality enhancement through advanced data preprocessing techniques and data augmentation methods will significantly improve the robustness and accuracy of diagnostic models.

Integrating advanced model architectures, including deep learning techniques like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can capture complex patterns within medical data, offering the potential to enhance diagnostic accuracy and generalize effectively across diverse healthcare settings. Clinical validation, ensuring the practical and effective use of these models in real-world clinical scenarios, should be a top priority. Collaboration between data scientists, medical experts, and domain specialists can lead to more effective diagnostic models, as it fosters a deeper understanding of the clinical context and its nuances. Moreover, ethical considerations should not be overlooked. As machine learning models become increasingly integrated into healthcare, ensuring patient data privacy, model transparency, and addressing potential biases are essential ethical considerations that should be at the forefront of future research efforts.

In conclusion, while the liver disease study has illuminated the potential of machine learning models in improving the early and accurate diagnosis of liver diseases, it also highlights the need for continued research and development. A combination of diverse datasets, advanced models, clinical validation, data quality enhancement, ethical considerations, and interdisciplinary collaboration holds the promise of revolutionizing liver disease diagnosis and significantly enhancing patient care and management. This field is ripe with opportunities for further exploration and innovation, and the journey toward more accurate and effective liver disease diagnosis holds great promise for the future of healthcare.

References

[1] "The Global Impact of Hepatic Fibrosis and End-Stage Liver Disease - ScienceDirect." Accessed: Nov. 03, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1089326108000743

[2] "Global Epidemiology of Chronic Liver Disease | SpringerLink." Accessed: Nov. 03, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-94355-8_5

[3] "Global challenges in liver disease - Williams - 2006 - Hepatology - Wiley Online Library." Accessed: Nov. 03, 2023. [Online]. Available: https://aasldpubs.onlinelibrary.wiley.com/doi/full/10.1002/hep.21347

[4] "Liver transplantation for nonalcoholic fatty liver disease: New challenges and new opportunities - PMC." Accessed: Nov. 03, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017047/

[5] "Systemic Symptoms in Non-Alcoholic Fatty Liver Disease | Digestive Diseases | Karger Publishers." Accessed: Nov. 03, 2023. [Online]. Available: https://karger.com/ddi/article-abstract/28/1/214/95475/Systemic-Symptoms-in-Non-Alcoholic-Fatty-Liver

[6] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," Expert Systems with Applications, vol. 67, pp. 239–251, Jan. 2017, doi: 10.1016/j.eswa.2016.08.065.

[7] C.-C. Wu et al., "Prediction of fatty liver disease using machine learning algorithms," Computer Methods and Programs in Biomedicine, vol. 170, pp. 23–29, Mar. 2019, doi: 10.1016/j.cmpb.2018.12.032.

[8] "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology | Clinical Infectious Diseases | Oxford Academic." Accessed: Nov. 03, 2023. [Online]. Available: https://academic.oup.com/cid/article/66/1/149/4085880

[9] "Towards Explaining the Effects of Data Preprocessing on Machine Learning | IEEE Conference Publication | IEEE Xplore." Accessed: Nov. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8731532/

[10] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas,

"Data Preprocessing for Supervised Leaning," vol. 1, no. 1, 2006.
[11] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.
[12] B. Chen, M. Li, J. Wang, and F.-X. Wu, "A logistic regression based algorithm for identifying human disease genes," in 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Nov. 2014, pp. 197–200. doi: 10.1109/BIBM.2014.6999153.
[13] "Random Forests for Bioinformatics | SpringerLink." Accessed: Nov. 03, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4419-9326-7_11
[14] Chaudhary, S. Kolhe, and R. Kamal, "An improved random forest classifier for multi-class classification," Information Processing in Agriculture, vol. 3, no. 4, pp. 215–222, Dec. 2016, doi: 10.1016/j.inpa.2016.08.002.
[15] D. A. Pisner and D. M. Schnyer, "Chapter 6 - Support vector machine," in Machine Learning, A. Mechelli and S. Vieira, Eds., Academic Press, 2020, pp. 101–121. doi: 10.1016/B978-0-12-815739-8.00006-7.
[16] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," in 2019 International Conference on Intelligent Sustainable Systems (ICISS), Feb. 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.
[17] "XGBoost Model for Chronic Kidney Disease Diagnosis." Accessed: Nov. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8693581/
[18] "Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward | Proceedings of the AAAI Conference on Artificial Intelligence." Accessed: Nov. 03, 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5055
[19] "Performance measures in evaluating machine learning based bioinformatics predictors for classifications | SpringerLink." Accessed: Nov. 03, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s40484-016-0081-2
[20] P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2
[21] P. Flach, "The many faces of ROC analysis in Machine Learning and Data Mining The many faces of ROC analysis in Machine Learning and Data Mining," Nov. 2023.
[22] "Distribution-balanced stratified cross-validation for accuracy estimation: Journal of Experimental & Theoretical Artificial Intelligence: Vol 12, No 1." Accessed: Nov. 03, 2023. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/095281300146272