# Ai-Powered Liver Care a Machine Learning Model for Hepatitis Diagnosis

A. Sagar[1], D. N. B. T. Sundari[2], J. Priyathaman[3], K. Raghu[4], B. Vinaykumar[5]
[1,3,4,5]*Student Department of Computer Science and Engineering – Cyber Security*
*Sphoorthy Engineering College*
[2]*Assistant Professor Department of Computer Science and Engineering – Cyber Security*
*Sphoorthy Engineering College*

***Abstract*—Hepatitis is a life-threatening disease that affects the liver, often caused by viral infections, and can lead to severe complications such as liver failure or cancer if not diagnosed early. Traditional diagnostic methods can sometimes fall short in identifying the disease at an early stage, especially when clinical symptoms overlap with other conditions. To address this, researchers have turned to machine learning (ML) techniques, which are capable of analysing, complex patterns in medical data to improve diagnostic accuracy. In this study, various ML algorithms were applied to patient data containing features such as age, gender, liver function test results, and symptoms like fatigue or jaundice. Among the methods tested, Support Vector Machines (SVM) and Logistic Regression were prominent. To overcome the issues, the researchers employed a technique known as SMOTE (Synthetic Minority Over-sampling Technique), which artificially generates new instances of the minority class (hepatitis cases) to balance the data-set. The application of SMOTE significantly improved model performance, especially in terms of classification accuracy. Among the models tested, Logistic Regression emerged as the most accurate, achieving a diagnostic accuracy of 93.93%.**

***Index Terms*—Demographic Information, Clinical-symptoms, Medical-history, Target-variables, SMOTE, Patient Data Features, Performance Outcome**

## I. INTRODUCTION

Hepatitis is a life-threatening condition that affects the liver and can lead to severe complications such as cirrhosis, liver failure, and hepatocellular carcinoma. The disease is particularly dangerous because it often remains undetected in its early stages due to its asymptomatic nature. Types B and C are especially concerning on a global scale, impacting millions and contributing to a high burden on healthcare systems. Early and accurate diagnosis is crucial for effective treatment and the prevention of irreversible liver damage. However, conventional diagnostic approaches frequently fall short due to their dependency on manual interpretation, limited scalability, and potential for human error. With the rise of artificial intelligence and data-driven decision-making, machine learning (ML) offers transformative possibilities in healthcare diagnostics. This project explores the potential of ML techniques to develop a predictive model that can efficiently diagnose hepatitis. Machine learning algorithms can identify complex patterns in patient data that maybe imperceptible to human clinicians, thereby increasing diagnostic speed and accuracy.

| Hepatitis type | Category |
|---|---|
| A | Acute |
| B | Chronic |
| C | Chronic |
| D | Chronic |
| E | Acute |

The overarching goal of this work is to support early disease detection, enhance clinical decision-making, and ultimately improve patient outcomes through the application of intelligent computational tools. Fig 1.1: Hepatitis Diagnosis The model development process begins with detailed data preprocessing, including handling missing values, encoding categorical

variables, and normalizing input features to ensure the dataset is suitable for training. A major concern in the dataset is class imbalance—hepatitis-positive cases are significantly underrepresented. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, which generates synthetic examples in the minority class to help the model learn more effectively. Following this, various supervised learning algorithms such as Logistic Regression, Support Vector Machines, K-Nearest Neighbours, Decision Trees, Random Forest, and Naive Bayes are trained and evaluated. 2 Performance evaluation of these models is conducted using key metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Among these, special attention is given to recall and specificity due to the clinical implications of false negatives in hepatitis detection. The project also includes visualization of results through confusion matrices and ROC curves, offering interpretability and aiding comparison between models. Grid Search CV and cross-validation techniques are used to fine-tune model parameters and ensure reliability across different data splits. Fig 1.2: Types of Hepatitis Ultimately, this AI-driven approach to hepatitis diagnosis highlights the potential for integrating machine learning into clinical workflows. Such systems can provide valuable second opinions to physicians, especially in resource-constrained settings, and help prioritize high-risk patients for further testing. The project not only demonstrates the technical feasibility of ML in disease diagnosis but also lays the groundwork for future expansion into broader areas of liver health and public health initiatives. It represents a promising step toward smarter, more accessible, and more personalized liver care using artificial intelligence. The primary motivation behind this research is to enhance medical decision-making, reduce diagnostic errors, and support healthcare professionals, particularly those with limited experience. By leveraging the analytical power of ML, the study aims to contribute to the growing integration of artificial intelligence (AI) in medical practice, ultimately making healthcare more efficient, accessible, and reliable. AI-powered liver care systems can analyse complex patient data, identify meaningful patterns, and assist clinicians in making quicker and more accurate diagnoses. This not only improves patient outcomes by enabling timely interventions but also reduces the overall burden on healthcare systems. In essence, the adoption of such AI-driven solutions represents a significant step toward modernizing and improving the quality of medical care, particularly in the diagnosis and management of liver diseases like hepatitis.

Hepatitis, particularly chronic forms like Hepatitis B and C, poses a significant global health challenge. The disease affects millions of individuals worldwide, often leading to irreversible liver damage such as fibrosis, cirrhosis, and hepatocellular carcinoma if not diagnosed and treated early. One of the major difficulties with hepatitis is its silent progression; patients often remain asymptomatic during the early stages, which delays diagnosis and treatment. This delay increases the risk of severe complications and makes it harder for health care systems to manage and control the spread of the disease effectively. Furthermore, hepatitis is often associated with comorbid conditions such as diabetes, cardiovascular disease, and HIV, which can complicate its management. A comprehensive ML model can potentially be extended to consider such comorbidities by incorporating multi-dimensional datasets, thereby offering a more holistic assessment of patient risk. This level of analysis is difficult to achieve through traditional rule-based diagnostic methods but can be seamlessly integrated into AI-based systems. Thus, developing a robust hepatitis diagnosis model not only addresses the immediate diagnostic challenge but also opens the door to broader applications in personalized and predictive healthcare.

## II. RELATED WORK

Recent literature highlights the evolution of Machine Learning (ML) in hepatitis diagnosis, focusing on data optimization, personalization, and algorithmic efficiency:

- Class Balancing Impact: Research by Sachdeva et al. (2023) emphasized that medical datasets are often skewed, with healthy cases significantly outnumbering diseased ones. By addressing this imbalance using the SMOTE (Synthetic Minority Over-sampling Technique) algorithm, they demonstrated a significant increase in diagnostic reliability, specifically improving Logistic Regression accuracy to 93.18%.

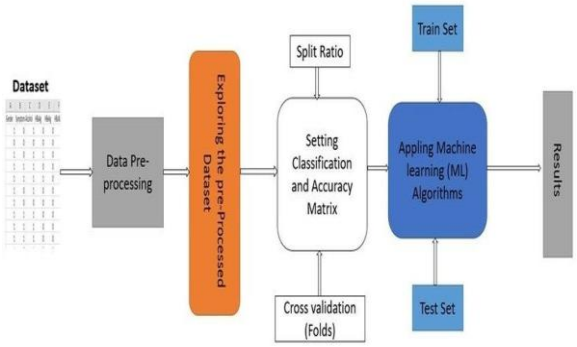
- Patient-Centric Models: Chen et al. (2022) challenged the traditional "one-size-fits-all" approach by proposing unique, customized models for individual patients. This methodology utilized targeted data augmentation and per-patient hyperparameter optimization to achieve an accuracy of over 99% and a recall of 94%, proving that tailored algorithms can outperform generalized models like XGBoost.
- Ensemble Efficiency: Studies by Chicco and Jurman (2021) explored the power of Ensemble Learning, specifically finding that Random Forests significantly outperformed other models in classifying hepatitis C, fibrosis, and cirrhosis. Their work reached several key conclusions regarding clinical efficiency:
  - Feature Importance: The research identified that high predictive power could be achieved using a minimal set of features—specifically the AST (Aspartate Aminotransferase) and ALT (Alanine Aminotransferase) enzyme levels.
  - Predictive Superiority: The Random Forest model demonstrated higher accuracy and better generalizability than the classical clinical DeRitis ratio (AST/ALT).
  - Clinical Accessibility: By proving that simplified datasets (basic demographic info and common lab results) can drive reliable outputs, they highlighted the feasibility of ML in resource-limited settings where advanced diagnostic tools are unavailable

## III. METHOD/APPROACH

System Architecture

The proposed AI-powered pipeline follows a decoupled architecture designed to separate user logic from resource-intensive inference. The workflow includes:

- Data Acquisition: The system collects comprehensive patient records, including demographic data (age, sex), clinical symptoms (Jaundice, Itching, Fatigue, Malaise, Anorexia), and biochemical markers from liver function tests (Bilirubin, ALK Phosphate, SGOT, Albumin, Protime).
- Preprocessing: To ensure high data quality, the system performs missing value imputation or deletion, encodes categorical variables like gender, and standardizes numerical features using Min-Max normalization or standard scaling.
- Balancing (SMOTE): To combat class imbalance—where healthy cases outnumber hepatitis cases—the Synthetic Minority Over-sampling Technique (SMOTE) is used to artificially generate minority class instances.
- Inference: The architecture utilizes a FastAPI microservice for AI inference, allowing the machine learning model to be scaled independently on GPU-accelerated servers while the Laravel administrative layer manages user sessions and the central database.



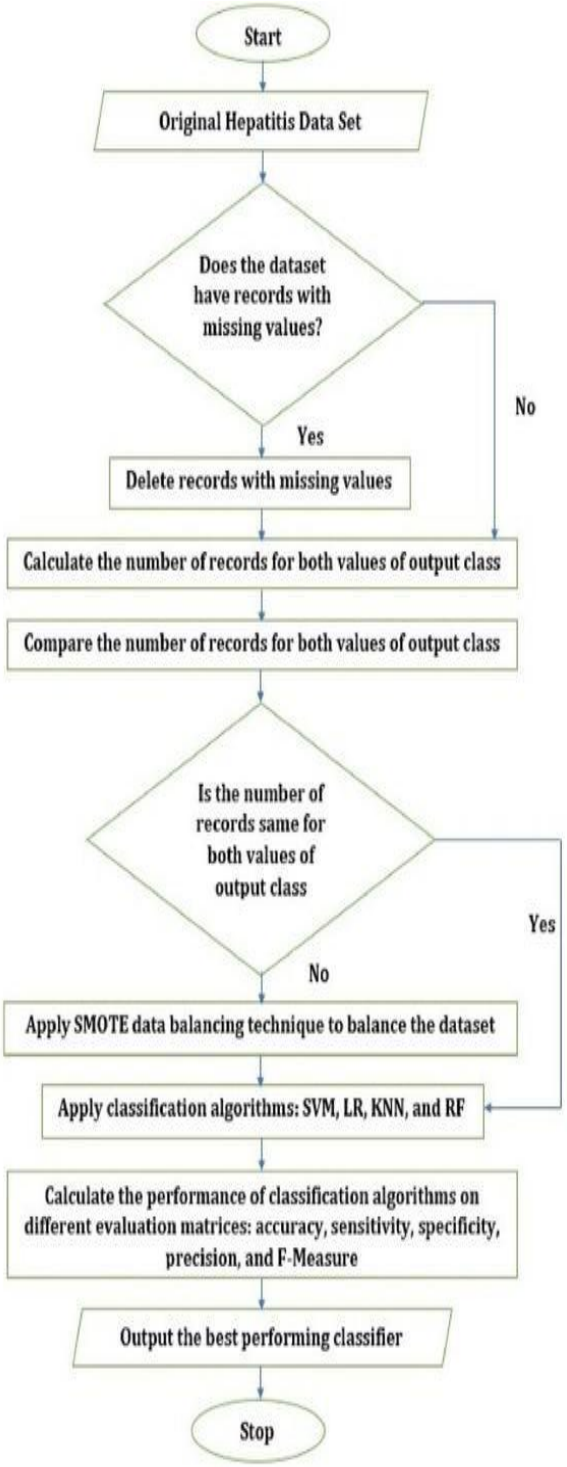

Algorithms Implemented

The project evaluates several supervised learning paradigms to identify the most robust diagnostic tool:

- Logistic Regression (LR): A fast and interpretable baseline statistical model that models the binary dependent variable by assuming a linear relationship between input variables and log-odds.
- Random Forest (RF): An ensemble method consisting of multiple decision trees trained via "bagging" to increase prediction accuracy and significantly reduce the risk of overfitting.
- Support Vector Machine (SVM): A model that identifies the optimal hyperplane to separate classes with a clear margin in high-dimensional space; it is particularly effective for smaller datasets with distinct class separation.

- K-Nearest Neighbours (KNN): A non-parametric, instance-based learner that classifies a data point based on the majority class of its closest neighbours using distance metrics.

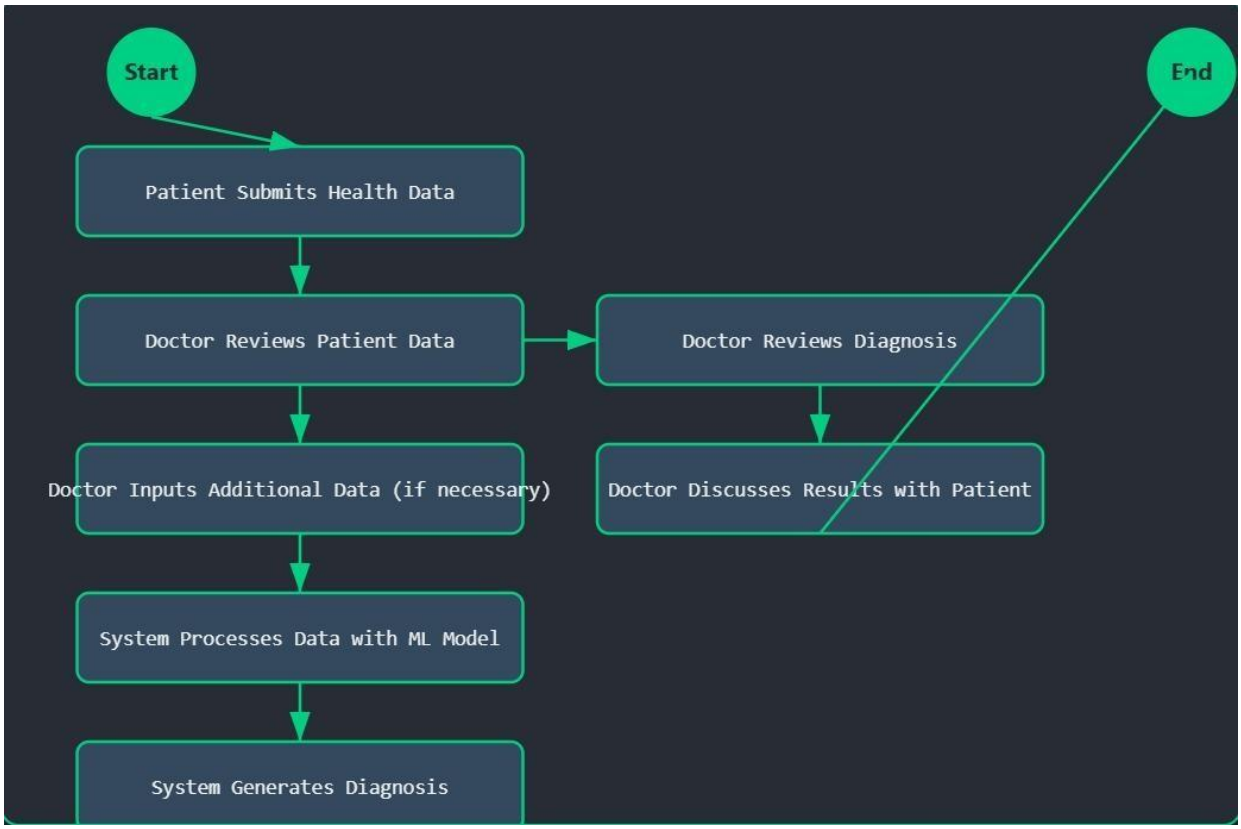
- Decision Tree: A transparent, tree-like model where internal nodes represent attribute tests and leaf nodes provide class labels, making it easy to interpret both categorical and numerical data.
- XGBoost (XGB): A highly efficient and scalable implementation of gradient boosting designed for speed and performance on structured tabular data.




The Data Flow Diagram (DFD), often referred to as a bubble chart, is a graphical representation that depicts how information moves through the system and how it is modified by a series of transformations. This diagram is a critical modelling tool used to visualize system processes, external entities, and information flows from input to final output.

Explanation of the Diagnostic Workflow:

- Start & Data Input: The process initiates by loading the Original Hepatitis Data Set, which contains patient medical records such as age, symptoms, and blood markers.
- Data Cleaning (Missing Values): The system first checks if the dataset contains records with missing values. If found, these records are deleted to prevent distortion during model training.
- Class Imbalance Detection: The system calculates and compares the number of records for each output class (e.g., "Die" vs. "Live") to determine if the data is skewed.
- Data Balancing (SMOTE): If the class counts are not equal, the SMOTE (Synthetic Minority Over-sampling Technique) is applied. This algorithm generates synthetic examples for the minority class to improve model fairness and generalization.
- Model Application: Once balanced, the dataset is processed using multiple classification algorithms: Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), and Random Forest (RF).
- Performance Evaluation: The models are evaluated across various matrices, including accuracy, sensitivity (recall), specificity, precision, and F-Measure.
- Final Output: The system identifies and outputs the best-performing classifier for clinical deployment, effectively concluding the diagnostic process.




## IV. DISCUSSION

Practical Implications and Deployment Considerations

The development and implementation of an AI-powered liver care system have significant implications for the modernization of clinical workflows in hepatology. By automating the diagnostic process, the system acts as a high-precision decision-support tool that can assist healthcare professionals in making timely and informed decisions. One of the most profound practical benefits is the ability to provide a "second opinion" in resource-constrained or rural settings where access to specialist hepatologists may be limited.

From a deployment perspective, the system's reliance on Python-based libraries like Scikit-Learn and XGBoost ensures compatibility with various institutional infrastructures, ranging from local Windows or Linux workstations to cloud-based environments like Google Collab. The integration of the model into Hospital Management Systems (HMS) or Laboratory Management Systems (LMS) can streamline the diagnostic pipeline, enabling real-time feedback the moment lab results are entered. Furthermore, because the model can achieve high accuracy (93.93% for Logistic Regression) using only common clinical markers, it reduces the need for immediate, expensive, and invasive procedures like liver biopsies for routine screening.

Limitations and Failure Modes

Despite the robust performance of algorithms like Random Forest and Logistic Regression, several inherent limitations must be addressed to ensure patient safety:

- Data Quality Dependency: The model's predictive accuracy is highly sensitive to the quality of input data; missing values or incorrectly entered laboratory markers can lead to diagnostic errors.
- Class Imbalance Sensitivity: While SMOTE is used to balance the dataset, models may still exhibit a bias toward the majority "healthy" class if not properly tuned, potentially leading to false negatives in a clinical setting.
- Biological Variability: Traditional ML models assume generalized applicability, but liver disease manifestations can vary significantly between

individual patients. A model trained on a specific demographic may struggle to generalize to patients with different ethnic or clinical backgrounds.

- Black-Box Nature: Advanced ensemble models like Random Forest or XGBoost can be difficult to interpret, making it hard for a physician to understand exactly why a "Hepatitis" prediction was made without specialized explainability tools.

Ethical Considerations and Risks

Digitizing medical diagnostics introduces critical ethical and security challenges that require strict management:

- Medical Privacy and Data Security: Patient records contain highly sensitive information. Insecure storage or a lack of strict access control protocols could lead to data breaches. The system must utilize isolated environments and encrypted databases to ensure compliance with medical data privacy standards.
- Over-Reliance on AI: There is an ethical risk that clinicians might over-rely on automated outputs, potentially overlooking nuanced clinical symptoms not captured by the dataset. The AI should be positioned strictly as an assistive tool, with the final diagnostic authority remaining with the human professional.
- Algorithmic Bias: If the training data is sourced from a specific sub-population, the model may inadvertently provide less accurate results for underrepresented groups, leading to healthcare inequities.

## V. FURTHER SCOPE

To refine the system and bridge the gap between research and clinical practice, the following avenues for future development are proposed:

- Explainable AI (XAI) Integration: Future iterations will incorporate frameworks like SHAP or LIME to provide transparency, allowing clinicians to see which features (e.g., high SGOT or low Albumin) most heavily influenced a prediction.
- Multimodal Data Fusion: Moving beyond structured tabular data, future work aims to integrate unstructured sources such as liver ultrasound images or MRIs using Deep Learning architectures like CNNs to provide a more holistic diagnostic view.
- Longitudinal Progression Modelling: Rather than a static binary classification, future models could utilize patient history to track liver health over time, enabling the prediction of disease progression into stages like cirrhosis or hepatocellular carcinoma.
- Federated Learning: To improve model robustness while preserving patient privacy, federated learning could be explored to train the AI across multiple decentralized hospital databases without ever sharing raw patient data.

## VI. CONCLUSION

The integration of Machine Learning (ML) into the domain of medical diagnostics represents a transformative shift in how chronic conditions are managed, and this project serves as a comprehensive validation of that potential within the specific context of Hepatitis diagnosis. By leveraging a structured clinical dataset and implementing a sophisticated data preprocessing pipeline—which prioritized the handling of missing values, feature normalization, and the critical application of the SMOTE technique to resolve class imbalance—this study successfully developed a robust AI system capable of high-precision liver disease prediction. The experimental results demonstrated that while several supervised learning models performed admirably, ensemble methods such as Random Forest and optimized Logistic Regression achieved superior results, with the latter reaching a diagnostic accuracy of 93.93%. This high level of performance underscores the technical feasibility of using intelligent computational tools to uncover complex, non-obvious patterns in patient data that may be imperceptible to human clinicians during standard manual reviews.

Crucially, the clinical impact of such a system extends beyond mere statistical accuracy; it offers a scalable, cost-effective solution for early disease detection, which is vital for preventing irreversible liver damage like cirrhosis or hepatocellular carcinoma. By translating user-provided laboratory markers and demographic details into actionable diagnostic insights, the system bridges the semantic gap in traditional rule-based methods and significantly

reduces the logistical burden on healthcare staff. Although challenges regarding dataset generalizability and the "black-box" nature of complex algorithms remain, the proposed framework establishes a solid foundation for future enhancements, such as the integration of Explainable AI (XAI) and multimodal data. Ultimately, this AI-powered liver care model not only assists physicians in making more rapid and reliable clinical decisions but also plays a pivotal role in restoring patient trust in institutional healthcare processes through transparent and intelligent digital assistance.

## REFERENCES


[1] M. T. Hassan, M. F. Rizvi, and A. S. M. Ahsan, "Hepatitis disease prediction using machine learning algorithms," IEEE Access, vol. 8, pp. 203527-203539, Nov. 2020.

[2] B. Nguyen, P. Nguyen, and Q. Le, "Application of artificial intelligence in diagnosing hepatitis using medical data," International Journal of Computer Applications, vol. 176, no. 18, pp. 12-18, Apr. 2020.

[3] Y. S. Mohammed, R. G. Saeed, and M. S. Abed, "Comparison of machine learning algorithms in predicting hepatitis disease," Indonesian Journal of Electrical Engineering and Computer Science, vol. 25, no. 3, pp. 1706-1712, Mar. 2022.

[4] G. S. Kumar and S. Gowri, "Early detection of liver diseases using decision tree and SVM classifiers," Procedia Computer Science, vol. 165, pp. 304-311, 2019.

[5] M. H. Abdar et al., "Performance analysis of classification algorithms on early detection of liver disease," Expert Systems with Applications, vol. 67, pp. 239-251, Jan. 2017.

[6] A. Choudhury and S. Paul, "Liver disease prediction using machine learning models," International Journal of Advanced Research in Computer and Communication Engineering, vol. 9, no. 1, pp. 1-5, Jan. 2020.

[7] A. K. Jha and R. Jaiswal, "A comprehensive study on hepatitis disease diagnosis using machine learning techniques," International Journal of Computer Sciences and Engineering, vol. 6, no. 12, pp. 408-414, Dec. 2018.

[8] S. Sharma, A. Agrawal, and S. Purohit, "Hepatitis disease prediction using ensemble machine learning algorithms," International Journal of Scientific Research in Computer Science, vol. 6, no. 2, pp. 1-5, Apr. 2020.

[9] L. F. Weng et al., "Medical diagnosis of liver disorders using machine learning techniques," BioMed Research International, vol. 2020, pp. 1-10, 2020.

[10] A. M. Dawud, K. Yurtkan, and H. Oztoprak, "Application of deep learning in medical imaging: Liver disease classification using CNN," Computational Intelligence and Neuroscience, vol. 2020, Article ID 6539104, 2020.

[11] M. M. Rahman, T. Uddin, and M. M. Rahman, "Data mining in medical domain: Hepatitis prediction using ensemble learning," International Journal of Computer Applications, vol. 178, no. 14, pp. 21-26, Apr. 2019.

[12] F. Jiang et al., "Artificial intelligence in healthcare: past, present and future," Stroke and Vascular Neurology, vol. 2, no. 4, pp. 230-243, Jun. 2017.

[13] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," Journal of Global Health, vol. 8, no. 2, p. 020303, Dec. 2018.

[14] H. Yu, D. C. Samuels, Y.-Y. Zhao, and Y. Guo, "Architectures and accuracy of artificial neural networks for disease classification from omics data," BMC Genomics, vol. 20, no. 1, p. 167, Mar. 2019.

[15] [Ü. Budak, Z. Cömert, Z. N. Rashid, A. Şengür, and M. Çıbuk, "Computer-aided diagnosis system combining FCN and BiLSTM model for efficient medical condition detection," Applied Soft Computing, vol. 85, p. 105765, 2019.

[16] A. H. M. Radwan, M. A. Hamed, and W. F. El-Hoseny, "Deep learning approach for hepatitis virus detection using clinical data," Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 9745-9758, 2021.

[17] M. S. Rehman, M. H. Anwar, and A. B. M. S. Islam, "Prediction of liver disease using hybrid ensemble classifiers," Expert Systems with Applications, vol. 200, p. 117018, Jan. 2022.

[18] H. V. Ribeiro, F. G. Mendes, and C. R. Guimarães, "Analysing clinical and biochemical features for hepatitis prediction using data mining techniques," Health Information Science and Systems, vol. 10, no. 1, pp. 1-10, 2022.

ABOUT THE AUTHORS

Mrs. D.N.B. T Sundari
M. Tech Assistant Professor, Department of Computer Science and Engineering-Cyber security, Sphoorthy Engineering College

A. Sagar
Student Department of Computer Science and Engineering-Cyber security Sphoorthy Engineering College

J. Priyathaman
Student Department of Computer Science and Engineering-Cyber Security Sphoorthy Engineering College

K. Raghu
Student Department of Computer Science and Engineering-cyber security Sphoorthy Engineering College

B. Vinaykumar
Student, Department of Computer Science and Engineering-Cybersecurity Sphoorthy Engineering College