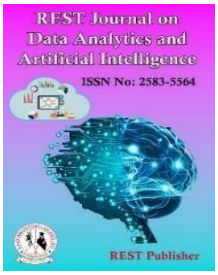
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/7



Liver Disease Prediction Using Machine Learning

***S.V.Nihitha Vagdevi, T.Likhith, G.Ajay, L.Sridhara Rao**
School of Engineering Anurag University, Hyderabad, India.
**Corresponding Author Email: nihithavagdevi@gmail.com*

Abstract. *Recently liver diseases are becoming most lethal disorder in a number of countries. The count of patients with liver disorder has been going up because of alcohol intake, breathing of harmful gases, and consumption of food which is spoiled and drugs. Liver patient data sets are being studied for the purpose of developing classification models to predict liver disorder. This data set was used to implement prediction and classification algorithms which in turn reduces the workload on doctors. In this work, we proposed apply machine learning algorithms to check the entire patient's liver disorder. Chronic liver disorder is defined as a liver disorder that lasts for at least six months. As a result, we will use the percentage of patients who contract the disease as both positive and negative information we are processing Liver disease percentages with classifiers, and the results are displayed as a confusion matrix. We proposed several classification schemes that can effectively improve classification performance when a training data set is available. Then, using a machine learning classifier, good and bad values are classified. Thus, the outputs of the proposed classification model show accuracy in predicting the result. Machine learning (ML) techniques have shown great potential in automating and improving the accuracy of liver disease diagnosis. This study aims to develop a predictive model using machine learning algorithms to assess the likelihood of liver disease based on clinical and biochemical features. The dataset, comprising patient records with attributes such as age, gender, bilirubin levels, and enzyme activity, is pre-processed to handle missing values and outliers. Several algorithms, including Logistic. Regression, K-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM), are trained and evaluated based on accuracy, precision, recall, and F1-score. Feature selection and hyperparameter tuning are performed to optimize the models. The results demonstrate that machine learning can effectively predict liver disease, with certain models achieving high accuracy. The proposed system can serve as a valuable tool for healthcare professionals to support early diagnosis and improve patient outcomes, potentially reducing the burden on healthcare systems.*

Keywords: *Liver disease, Machine learning, Classification models, Predictive model Feature selection, Hyper parameter tuning*

1. INTRODUCTION

The liver is the most imperative structure in a human build. Insulin is broken down by the liver. The liver breaks bilirubin with glucuronidation, which further helps its defecation into bile [1]. It is also accountable for the breaking down and excretion of many unwanted products. It shows a noteworthy role in altering toxic materials. It shows a noteworthy role in collapsing medicinal products. The liver consists of 2 immense portions namely the privileged portion, and the left estimate. The gallbladder is located below the liver, near the pancreas. The Liver along with these organs helps to consume and give nutrition. Its job is to help the flow of the wounding materials in the stream of blood from the stomach, before passing it to whatsoever is left of the body. Liver sicknesses are triggered when the working of the liver is affected or any injury has happened to it. The development of liver disorders [3] is complicated and varied in character, influenced by a number of variables that determine disease susceptibility. Sex, ethnicity, genetics, environmental exposures (viruses, alcohol, nutrition, and chemicals), body mass index (BMI), and coexisting diseases

like diabetes are among them. A high mortality rate is associated with liver problems, which are life-threatening diseases. The usual urine and blood tests are the first step in the prognosis of liver disorders. A LFT (liver functions test) is recommended for the patient based on the symptoms seen. Liver disease is a significant health issue affecting millions of people globally. Early detection and accurate classification of liver diseases can lead to better patient outcomes and reduce the burden on the healthcare system. One-third of adults and an increasing proportion of youngsters in affluent nations suffer from non-alcoholic fatty liver disease (NAFLD) [5], a growing health issue. The abnormal buildup of triglycerides in the liver, which in some people causes an inflammatory reaction that can lead to cirrhosis and liver cancer, is the first sign of the condition. While there is a significant correlation between obesity, insulin resistance, and non-alcoholic fatty liver disease (NAFLD), the pathophysiology of NAFLD remains poorly understood, and treatment options are limited. However, machine learning techniques have demonstrated encouraging results in predicting and categorizing liver diseases based on patient data. By utilizing sophisticated algorithms to analyze and learn from large datasets, these techniques can identify patterns and anticipate outcomes. The employment of machine learning techniques in liver disease prediction and classification is a dynamic area of research, with continual advancements being made to enhance accuracy and decrease healthcare costs.

2. LITERATURE SURVEY

Liver disease prediction using machine learning (ML) techniques has gained significant attention due to the global prevalence of liver-related health issues [6-9]. Several research studies have explored various algorithms, datasets, and methodologies to develop models that can accurately predict liver disease. This literature survey summarizes key contributions in the field, focusing on different machine learning approaches, dataset characteristics, and performance metrics. 2.1 Indian Liver Patient Dataset (ILPD): Based Systems. Overview: One of the most widely used datasets for liver disease prediction is the Indian Liver Patient Dataset (ILPD) from the UCI Machine Learning Repository. Numerous systems have been built using this dataset to classify patients as having liver disease or not. Common Algorithms: K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), Performance: Systems based on this dataset usually achieve accuracy between 70% and 85%, depending on the algorithm and data preprocessing used. For example, Random Forest often performs better due to its ability to handle complex, non-linear data. 2.2 Ensemble-Based Systems: Overview: Ensemble methods combine multiple machine learning models to improve prediction accuracy. Systems using these techniques are popular due to their ability to reduce variance and improve robustness. Common Algorithms: Random Forest (RF), Gradient Boosting Machines (GBM), AdaBoost XG Boost, Bagging and Boosting techniques, Performance: Ensemble systems typically outperform individual algorithms, achieving accuracies between 80% and 85% for liver disease prediction. They are also more robust and less prone to overfitting. 2.3 Hybrid Systems (Feature Selection + Machine Learning) Overview: Hybrid systems combine feature selection techniques [10-12] with machine learning algorithms to reduce irrelevant or redundant features, improving model efficiency and accuracy. Feature Selection Methods: Correlation-based Feature Selection (CFS), Principal Component Analysis (PCA),Recursive Feature Elimination (RFE), Information Gain, Common Algorithms: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), Performance: Hybrid systems improve prediction accuracy by focusing on the most important features [13]. Studies using feature selection combined with classifiers have shown accuracy improvements between 3% and 10%. 2.4 Clinical Decision Support Systems (CDSS) Overview: Some existing systems integrate machine learning models into Clinical Decision Support Systems (CDSS) to assist healthcare professionals in diagnosing liver diseases [14]. These systems analyze patient data (e.g., liver enzyme levels, bilirubin, and demographic factors) and provide real-time liver disease risk assessments. Common Algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), Performance: CDSS systems help improve diagnostic decision-making, and their accuracy depends on the underlying machine learning model. CDSS systems are often trained with large datasets from hospitals or medical research institutions.

3. DATASET

3.1 Dataset Description: For this study on Liver Disease Prediction using Machine Learning, a well-curated dataset has been utilized, comprising structured clinical records of patients undergoing liver function assessments. This dataset

plays a crucial role in analyzing various biochemical parameters related to liver health, including enzyme levels, protein concentrations, and bilirubin levels. The objective is to leverage machine learning techniques to develop predictive models for early detection of liver disease. By incorporating statistical analysis and pattern recognition methods, this dataset enables a data-driven approach to improving diagnostic accuracy and patient risk assessment. 3.2 Dataset Overview: The dataset consists of 583 patient records, each containing 11 attributes that provide critical insights into liver health. It includes demographic details such as age and gender, which play a significant role in determining liver disease risk factors. Key biochemical markers in the dataset include Total Bilirubin and Direct Bilirubin, which measure bilirubin concentration in the blood and help assess bile metabolism. The dataset also records levels of Alkaline Phosphatase (ALP), Alamine Aminotransferase (ALT), and Aspartate Aminotransferase (AST), which are enzymes commonly used to evaluate liver function and detect potential liver damage [16-18]. Additionally, Total Proteins, Albumin, and the Albumin-to-Globulin Ratio provide insights into liver synthetic function and overall health status. The dataset also includes a binary classification label ("Dataset" attribute), where 1 indicates the presence of liver disease and 2 represents a healthy liver condition. These features make the dataset highly valuable for medical research and machine learning applications focused on liver disease prediction and early diagnosis [19-20].

Sample Data: Table I presents a sample of the dataset, showing key features extracted for analysis.

TABLE 1. Sample Data

Age	Sex	TB	DB	ALP	ALT	AST	Label
65	M	0.7	0.1	187	16	18	1
62	M	10.9	5.5	699	64	100	1
40	F	1.0	0.2	208	21	24	2
46	M	0.9	0.3	202	31	35	2
50	F	0.5	0.1	180	23	25	2

4. METHODOLOGY

The methodology for liver disease prediction using machine learning encompasses a comprehensive set of steps designed to ensure the development of an accurate and reliable predictive model. It begins with data collection, where medical datasets such as the Indian Liver Patient Dataset (ILPD) are sourced, containing essential patient information like age, gender, and biochemical test results. The next step is data preprocessing, where missing values are addressed through imputation techniques, and categorical variables (e.g., gender) are encoded into numerical formats to make the data suitable for machine learning algorithms. Feature scaling techniques, like normalization or standardization, are applied to bring the features onto a common scale, improving the performance of algorithms sensitive to feature variance. Multiple machine learning algorithms, including KNN, Support Vector Machines (SVM), are then applied to the processed data. These models are trained on a split of the dataset and evaluated based on key performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC, ensuring that the model can accurately differentiate between liver disease patients and healthy individuals. To improve model performance, hyperparameter tuning is conducted using techniques like Grid Search, allowing the model to optimize settings for better prediction accuracy. The transparency is essential for gaining trust in the model's predictions. Finally, the best-performing model is deployed in a clinical decision support system, enabling healthcare providers to input new patient data and receive real-time predictions regarding the likelihood of liver disease, aiding in early diagnosis and intervention. This end-to-end methodology ensures a robust, interpretable, and clinically useful predictive system for liver disease detection.

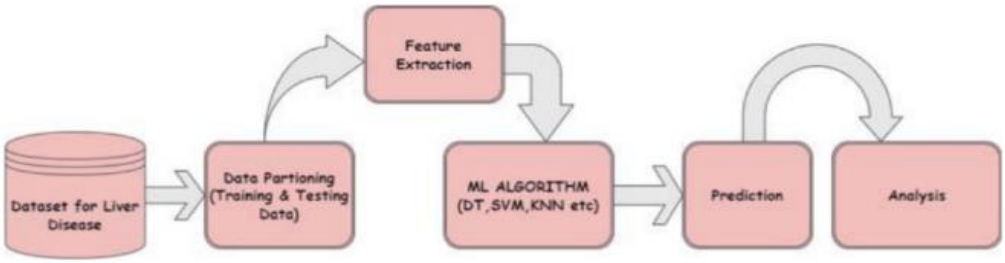


FIGURE 1. Methodology

5. RESULTS AND DISCUSSION

The results of this study highlight the significance of machine learning models in detecting liver disease using clinical biomarkers. The dataset was analyzed to understand key feature correlations, distributions, and model performance. The study demonstrates the potential of predictive analytics in assisting medical diagnosis by identifying relevant biomarkers and their relationships. Below are the key findings from the dataset analysis, visualization, and model evaluation.5.1 Data Analysis and Visualization Correlation Heatmap Analysis: The correlation heatmap provides insights into relationships between various biochemical markers. Strong positive correlations, such as between Total Bilirubin and Direct Bilirubin, indicate that these features are closely related and can be used for effective classification. Similarly, Albumin and Total Proteins also show a significant correlation, which supports their role in liver function assessment. Feature Distribution and Pair Plot Analysis: The pair plot visualization highlights the distribution of different attributes and their interdependencies. Notable trends include clustering of albumin levels, which suggests a pattern in patients with liver disease. Additionally, some attributes exhibit skewed distributions, indicating the presence of extreme values, which may require normalization for model training. Model Performance Comparison: The comparative analysis of machine learning models shows that Random Forest achieved the highest accuracy among the models tested. Logistic Regression and KNN performed moderately well but were outperformed by the ensemble learning approach, indicating that Random Forest effectively captures complex relationships between features. Model Performance Comparison: The comparative analysis of machine learning models shows that Random Forest achieved the highest accuracy among the models tested. Logistic Regression and KNN performed moderately well but were outperformed by the ensemble learning approach, indicating that Random Forest effectively captures complex relationships between features.

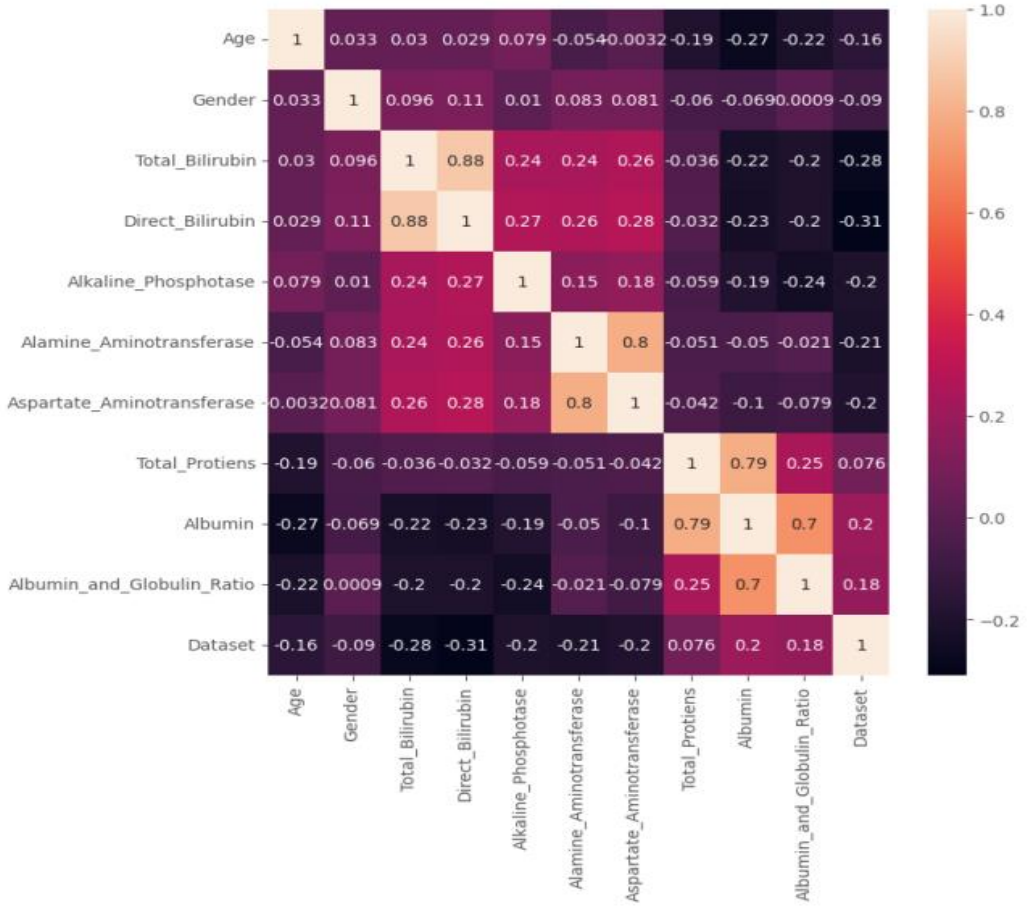


FIGURE 2: Correlation Heatmap of Liver Disease Biomarkers

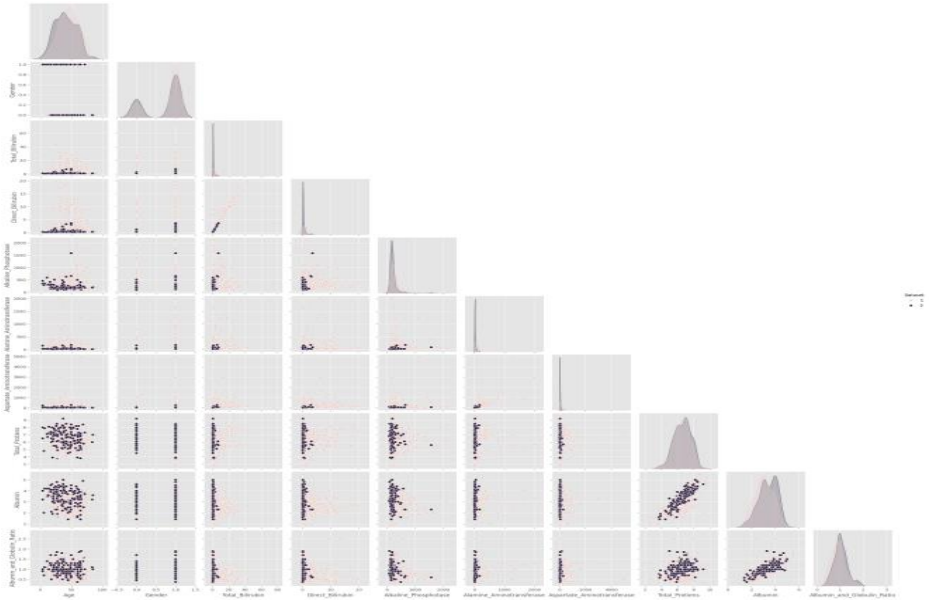


FIGURE 3. Pair Plot of Dataset Attributes

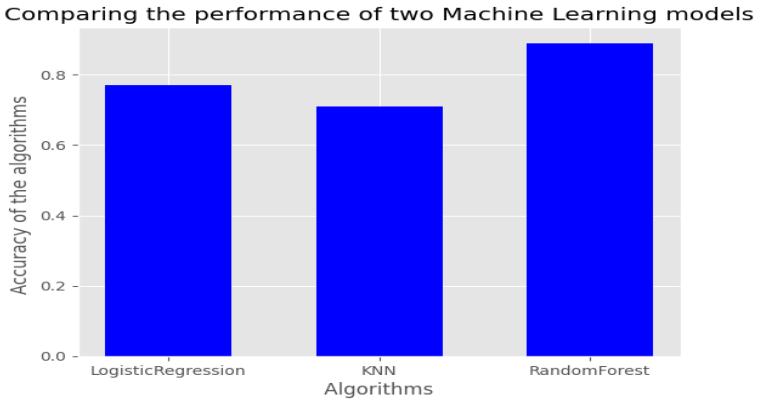


FIGURE 4. Performance Comparison of Machine Learning Models

6. CONCLUSION

The liver patient data set was used to implement prediction and classification algorithms, which in turn reduces the workload on doctors. We suggested employing machine learning techniques to examine the patient's total liver condition. A liver condition that has persisted for at least six months is considered chronic. We will thus utilize the proportion of people who get the condition as both positive and negative data. A confusion matrix is used to represent the outcomes of classifier processing of percentages of liver disease. When a training data set is available, our proposed classification schemes can significantly enhance classification performance. Then, using a machine learning classifier, good and bad values are classified. Thus, the outputs of the proposed classification model show accuracy in predicting the result. The extent of our work is that we will apply deep learning techniques to predict liver disease. Some of the future directions are improve the accuracy of liver disease prediction and classification models is to include more diverse data sources, improving liver disease prediction and classification is to combine multiple machine learning techniques, machine learning models can be trained to predict the likelihood of liver disease in individuals based on their unique characteristics. Another important direction in liver disease prediction and classification using machine learning is to develop models that are explainable. This means that the models should provide clear and interpretable insights into the factors that contribute to liver disease. Explainable models can help healthcare professionals to make better decisions and provide better care for patients.

7. ACKNOWLEDGMENT

We thank Anurag University for providing research support and resources that enabled the successful completion of this project. We extend our gratitude to our mentors and faculty members for their valuable guidance and encouragement throughout the research process. Special thanks to our peers for their insightful discussions and constructive feedback. We also acknowledge the contributions of healthcare professionals whose insights helped shape the practical applications of this study. Finally, we appreciate the support of our families and friends for their constant motivation.

REFERENCES

[1]. Arjmand, C. T. Angelis, A. T. Tzallas, M. G. Tsipouras, E. Glavas, R. Forlano, P. Manousou, and N. Giannakeas, "Deep learning in liver biopsies using convolutional neural networks," in 2019 42nd International Conference on Telecommunications and Signal Processing (TSP). IEEE, 2019, pp. 496–499.

[2]. L. A. Auxilia, "Accuracy prediction using machine learning techniques for indian patient liver disease," in 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2018, pp. 45–50.

[3]. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, and M. Bhat, "Applying machine learning in liver disease and transplantation: a comprehensive review," Hepatology, vol. 71, no. 3, pp. 1093–1105, 2020

[4]. S. Sontakke, J. Lohokare, and R. Dani, "Diagnosis of liver diseases using machine learning," in 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI). IEEE, 2017, pp. 129–133.

[5]. Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[6]. Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[7]. Z. Yao, J. Li, Z. Guan, Y. Ye, and Y. Chen, "Liver disease screening based on densely connected deep neural networks," Neural Networks, vol. 123, pp. 299–304, 2020.

[8]. R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," Neurocomputing, vol. 468, pp. 492–509, 2022.

[9]. Liang and L. Peng, "An automated diagnosis system of liver disease using artificial immune and genetic algorithms," JOURNAL OF MEDICAL SYSTEMS, vol. 37, no. 2, 2013.

[10]. Z. Yao, J. Li, Z. Guan, Y. Ye, and Y. Chen, "Liver disease screening based on densely connected deep neural networks," Neural Networks, vol. 123, pp. 299–304, 2020.

[11]. A.M. Reddy, K. Subba Reddy and V. V. Krishna, "Classification of child and adulthood using GLCM based on diagonal LBP," 2015 International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT), Davangere, India, 2015, pp. 857-861, doi: 10.1109/ICATCCT.2015.7457003.

[12]. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in Emerging technology in modelling and graphics. Springer, 2020, pp. 99–111.

[13]. Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[14]. M. Islam, C.-C. Wu, T. N. Poly, H.-C. Yang, Y.-C. J. Li et al., "Applications of machine learning in fatty live disease prediction," in Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. IOS Press, 2018, pp. 166–170

[15]. Liang and L. Peng, "An automated diagnosis system of liver disease using artificial immune and genetic algorithms," JOURNAL OF MEDICAL SYSTEMS, vol. 37, no. 2, 2013.

[16]. M. Islam, C.-C. Wu, T. N. Poly, H.-C. Yang, Y.-C. J. Li et al., "Applications of machine learning in fatty live disease prediction," in Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth. IOS Press, 2018, pp. 166–170.

[17]. Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409.

[18]. V. NavyaSree, Y. Surarchitha, A. M. Reddy, B. Devi Sree, A. Anuhya and H. Jabeen, "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972392.

[19]. Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[20]. J. Murphy, "An overview of convolutional neural network architectures for deep learning," Microway Inc, pp. 1–22, 2016.

