IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

Supervised Machine Learning Approach for Liver Disease Prediction and Classification

Ankitha D M, Thanmayee V, Vaishnavi A, Vaishnavi D, Varsha K Kulkarni

Assistant Professor, Department of ECE, Vidyavardhaka College of Engineering, Mysuru, India

Department of ECE, Vidyavardhaka College of Engineering, Mysuru, India

Department of ECE, Vidyavardhaka College of Engineering, Mysuru, India

Department of ECE, Vidyavardhaka College of Engineering, Mysuru, India

Department of ECE, Vidyavardhaka College of Engineering, Mysuru, India

ABSTRACT: Liver disease is a major global health issue. Various factors like obesity and hepatitis can damage the liver, leading to complications. Diagnosing this disease can be complex. This project explores using artificial intelligence (AI) and supervised learning algorithms (Logistic Regression, Random Forest, SVM) to analyze patient data (Kaggel database) and predict future outcomes for liver disease patients. The project demonstrates that these AI models achieve good accuracy after feature selection.

KEYWORDS: Machine learning, liver ailment, categorization, supervised learning, LR (Logistic Regression), RF (Random Forest), SVM (Support Vector Machine).

I. INTRODUCTION (HEADING 1)

Liver:

The Liver organ resides in the upper right quadrant of the abdomen and stands as the second largest bodily organ after the integument. It assumes a triangular form and functions as the primary gland, generating hormones. It undertakes over 500 roles within the human anatomy and upholds the functionality of numerous vital organs essential for existence. In mature individuals, its mass accounts for roughly 2% of total body mass, spanning from 1.4 to 1.8 kilograms in males and 1.2 to 1.4 kilograms in females, while infants' livers weigh around 150 grams.

The liver executes the subsequent tasks:

- It excretes bile and glycogen.
- It manufactures serum protein lipids.
- It purges the bloodstream of both internal and external substances, encompassing toxins, medications, and alcohol.
- It stockpiles vitamins D, A, K, E, and B12.

Liver Disease:

- Liver Inflammation of the liver may stem from harmful substances, infections, or genetic conditions, impairing normal liver function. It Plays an essential role in digestion and bacterial elimination. Typically affecting individuals aged 40 to 60, liver diseases primarily impact males. In India, around one million individuals receive liver disease diagnoses annually, with 140,000 succumbing to it each year. Globally, liver ailments claim two million lives annually, posing a significant health challenge. Viral infections (such as hepatitis), excessive alcohol intake, and other factors contribute to liver impairment. Various liver diseases include hepatitis, cirrhosis, liver tumors, and cancer. Liver ailments and cirrhosis are the principal causes of mortality.
- Machine learning (ML), a subset of artificial intelligence (AI), mimics human intelligence by programming machines to reason and emulate human actions. It facilitates knowledge acquisition without explicit programming. In healthcare, ML enhances precision in processing and medical diagnosis using classification methodologies. Liver disease symptoms can be elusive in the initial stages due to the organ's continued functionality despite partial damage. Timely detection of liver issues improves patient survival rates. ML models (such as logistic regression, stochastic forests, and SVM) can predict liver disease occurrence using patient datasets. Addressing prior research oversights enhances predictive accuracy. ML significantly impacts biomedical domains, improving liver disease prognosis and diagnosis objectivity.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

II. BACKGROUND STUDY

In the past few years, the utilization in supervised machine-learning methods involved significant traction in the realm of liver disease prediction and classification. This novel method entails training computational models to recognize patterns within datasets, assisting in identifying and categorizing liver diseases. Through the employment of supervised learning algorithms, such as logistic regression, random forest, and SVM (support vector machines), researchers aim to develop predictive models capable of accurately diagnosing liver conditions based on patient data.

This emerging methodology marks a departure from traditional diagnostic methods, which often rely on manual interpretation and may lack the precision required for timely and effective disease detection. Through the utilization of machine learning, healthcare professionals and researchers can harness extensive Information regarding the patient, encompassing demographic particulars, medical records, laboratory tests, and imaging results, to improve diagnostic precision and enable tailored treatment approaches.

In essence, the capability of supervised machine-learning in liver disease prediction and classification represents a promising avenue for advancing healthcare practices, offering the potential for earlier detection, improved diagnostic accuracy, and enhanced patient outcomes.

III. METHODOLOGY

Creating an ML model for liver disease classification entails a series of crucial steps. Here's a comprehensive methodology to navigate through the process:

1. Data Collection and Pre-processing:

a) **Data Sources:** Compile an extensive dataset comprising patient information, encompassing clinical records, laboratory test outcomes, and Data derived from medical imaging where accessible.
b) **Data Cleaning:** Preprocess the dataset provided by addressing absent data points, anomalies, and irregularities, while ensuring the data are present and formatted appropriately for the purpose of machine learning tasks.
c) **Data Split:** Partition the splitting of the dataset into training and testing subsets to accurately assess the effectiveness of models, allocating, for instance, 70% throughout the training stage and 30% for testing.

2. Feature Engineering:

a) **Feature Selection:** Choose relevant attributes that are prone to influence diagnosing liver disease. Medical experts' input can be invaluable in this step.
b) **Feature Scaling:** Standardize numerical features to ensure uniform scales, utilizing frequently employed methods like Min-Max scaling or standardization using Z-scores.
c) **Encoding Categorical:** Transforming categorical variables into numerical ones representations, for example, through techniques like one-hot encoding.

3. Model Selection:

a) **Choose Appropriate Algorithms:** Select suitable ML algorithms for classification tasks, considering options such as logistic regression, decision trees, random forests, and support vector machines, and advanced deep learning techniques such as neural networks.

4. Model Training:

a) Train the Models: Fit the chosen ML models to the training dataset using the selected features.

5. Model Evaluation:

a) **Performance Metrics:** Assess the efficacy of the models during the testing phase dataset using a range of classification metrics, which may include correctness, exactness, completeness, and F1-score.
b) **Confusion Matrix:** Generate a confusion matrix to visually represent the dispersion of correctly classified positives, correctly classified negatives, incorrectly classified positives, and incorrectly classified negatives.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

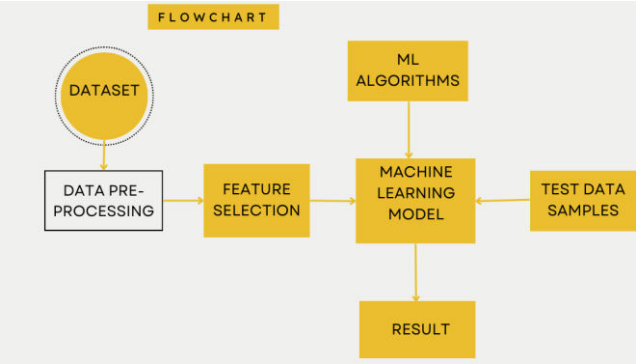

Figure 1: Block Diagram of ML model

IV. IMPLEMENTATION

Data Collection

Gather a comprehensive dataset containing relevant information such as patient demographics (age, gender), medical history, symptoms, lifestyle factors (alcohol consumption, smoking), and various laboratory test results (e.g., liver function tests, blood chemistry panels). Publicly available datasets, including the Indian Liver Patient dataset from Kaggel, can be useful.

Data Preprocessing

- Managing missing values: Fill in missing data using methods like mean imputation, median imputation, and predictive imputation using information from other features, or removal of the particular row of the missing value.
- Outlier detection and removal: Identify and deal with outliers that might have adverse effects on model performance.
- Feature scaling and normalization: Normalize numerical features to a similar range to prevent certain features from exerting a disproportionate influence on others during model training.
- Feature Selection- Perform exploratory data analysis (EDA) to comprehend the associations between features and the target variable, which in this case is the presence or absence of liver disease.
- Employ techniques for ranking feature importance, such as decision trees or ensemble models, to identify the most informative features.
- Optionally, consider applying dimensionality reduction methods like Principal Component Analysis (PCA) to decrease the number of features while retaining a significant portion of the variance in the dataset.

Model Selection

Selecting the right classification algorithms depends on the problem's nature and dataset characteristics. For binary classification tasks such as predicting liver disease, commonly employed algorithms include logistic regression, random forests, support vector machines (SVM), and gradient boosting algorithms. Experimenting with multiple algorithms helps determine which one performs best for the specific dataset.

Model Training

During model training, it's crucial to divide the dataset into training and testing subsets using methods such as train-test split or k-fold cross-validation to guarantee impartial evaluation. Train the chosen models on the training data while adjusting hyperparameters as needed. Continuously monitor model training for convergence and make adjustments to hyperparameters if necessary.

Model Evaluation

In the model evaluation phase, assess the trained models on the testing dataset using pertinent evaluation metrics like accuracy, precision, recall, F1-score, and area under the ROC curve (ROC-AUC). Compare the performance of various models and choose the one with the most favorable overall performance metrics. Conduct additional analyses such as confusion matrices to gain insight into the model's strengths and weaknesses in predicting liver disease.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

V. RESULT

This section presents the comprehensive findings of the experiment assessing different machine-learning models for liver disease prediction and classification. The primary dataset was sourced from Kaggle, a treasure trove of real-world data. To increase the robustness of the model, the dataset was supplemented with manually collected samples, culminating altogether 5000 instances. This enriched dataset was meticulously curated to bolster the model's generalizability and address any inherent partiality within the original Kaggle data.

The features considered were directly related to liver function, providing a more comprehensive grasp of the intricate interplay of physiological indicators. Additionally, a binary target variable labeled "Dataset" succinctly indicated the absence or presence of liver disease. The findings promise to illuminate the path toward more precise and dependable liver disease prediction models.

VI. DATA PREPROCESSING

Data Cleaning:
The dataset was meticulously inspected regarding missing values and inconsistencies. Since no outliers were identified, the focus was on handling missing data. We opted to eliminate rows containing missing values as a result due to the relatively small number of missing entries (18) specifically in "Albumin_and_Globulin_Ratio" feature. This method guarantees the integrity pertaining to the dataset and avoids introducing biases through imputation techniques, which might not be suitable for all missing data patterns.
Number of patients diagnosed with liver disease: 3677
Number of patients not diagnosed with liver disease: 1305

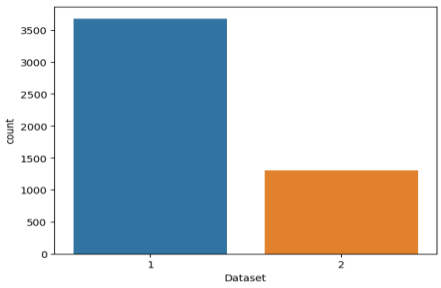

Figure 2: Patients diagnosed with (represented by 1) and without (represented by 2) liver disease

Feature Engineering:

Label Encoding:
The categorical "Gender" feature was encoded to numerical values for compatibility with machine-learning algorithms. Here, "Female" was converted to 0 and "Male" to 1. It's a typical encoding strategy for categorical features with binary values.
Number of patients that are male: 3045
Number of patients that are female: 1937

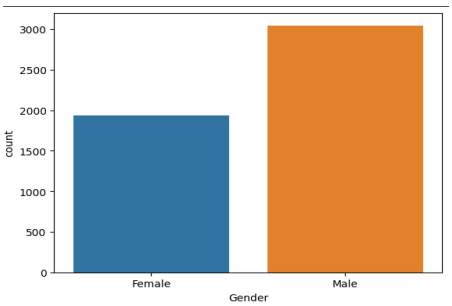

Figure 3: Gender distribution

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

Feature Scaling:
To optimize model performance and minimize computational overhead, the exclusive approach employed was Min-Max scaling. This technique normalizes attributes for a particular range (typically 0 to 1). By doing so, it ensures that features with varying scales do not disproportionately influence the model during training. The "Gender" attribute, already numerical, remained untouched in this process.

Following the preprocessing steps, the data underwent division into testing and training sets using a stratified sampling approach. This approach guaranteed that both datasets maintained an equivalent class distribution-specifically, the absence or presence of liver disease. The resulting sizes were as follows: the training dataset (denoted as X_train) consisted of 3487 samples, each with 10 features. Similarly, the testing dataset (X_test) comprised 1495 samples, also with the same 10 features. Notably, the "Dataset" feature was clearly and deliberately omitted from this count, leaving a total of 9 relevant features for model evaluation.

This approach guarantees a robust assessment of model performance, accounting for the inherent variability in the data distribution across different classes.

VII. MACHINE-LEARNING MODEL EVALUATION

Three distinct machine-learning algorithms underwent assessment regarding their performance in liver disease prediction: Random Forest, Logistic Regression, and SVM (Support Vector Machine). These algorithms represent diverse approaches to classification problems, offering a comprehensive comparison.

Random Forest:
Random Forest, an ensemble learning method, harnesses the collective wisdom of multiple decision trees. By combining their predictions, it constructs a robust and accurate model. This approach excels in handling intricate interactions among features and the response variable. Whether involving intricate information or nonlinear patterns, Random Forest emerges as a reliable ally in the realm of machine learning.

Logistic Regression:
Logistic Regression, a linear classification model, estimates the probability of a data point belonging to a particular class. Its simplicity, interpretability, and efficiency mark it as a widely adopted algorithm. Whether predicting customer churn or diagnosing diseases, Logistic Regression remains a stalwart choice.

Table 1: Comparision between Random Forest and Logistic Regression based on Accuracy

Sl no.	Model	Score	Test Score
1	Random Forest	100.00	94.52
0	Logistic Regression	87.81	87.29

SVM (Support Vector Machine):
SVM, a powerful machine learning algorithm, seeks to discover the optimal hyperplane that efficiently divides the individual data values of different classes. Its versatility extends to high-dimensional data, and it gracefully handles non-linear relationships using kernel functions. When confronted with intricate choice boundaries, SVM emerges as a formidable contender.

In the liver disease prediction study, three machine learning algorithms-Random Forest, Logistic Regression, and SVM underwent rigorous evaluation. Each algorithm brought unique strengths to the table.

To assess performance, a suite of evaluation metrics was employed, including accuracy, F1-score, precision, recall, and the Area Under the Curve (AUC). These metrics provided a comprehensive perspective on the performance of each model on the dataset.

Now, let's delve into the outcomes and findings and observe how the results of these algorithms stack up against each other:

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

Table 2: Machine Learning Model Evaluation

Model	Accuracy	F1-Score	Precision	Recall	AUC
Random Forest	0.9452	0.9642	0.9389	0.991	0.9699
Logistic Regression	0.8729	0.9144	0.9194	0.9095	0.9233
SVM	0.802	0.8543	0.9476	0.7778	0.9209

From the performance table, it's evident that the Random Forest model outshone the others across all evaluation metrics. Its superiority in predicting liver disease within this dataset is evident. Specifically:

- The model accomplished an impressive test accuracy of 94.52%, signifying a high success rate in correctly classifying patients.
- Additionally, the F1-score of 0.9642 reflects a well-balanced performance, considering both precision and recall.

In summary, Random Forest emerges as a robust contender for accurate liver disease predictions.

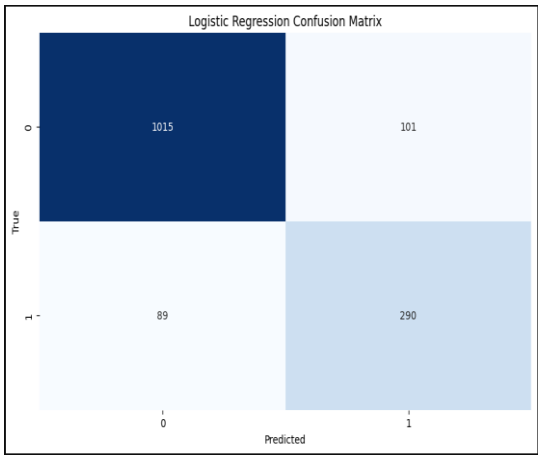

Figure 4: Confusion matrix of Logistic Regression

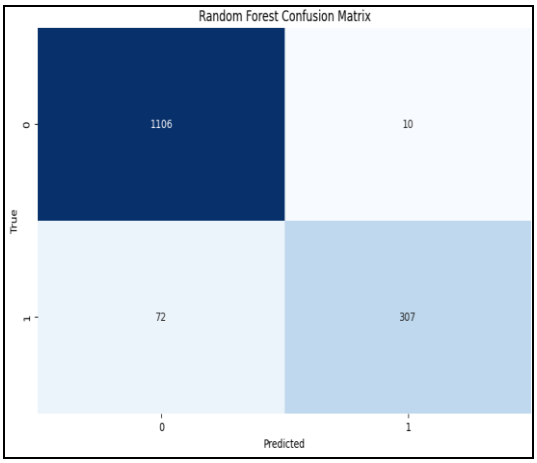

Figure 5: Confusion matrix of Random Forest

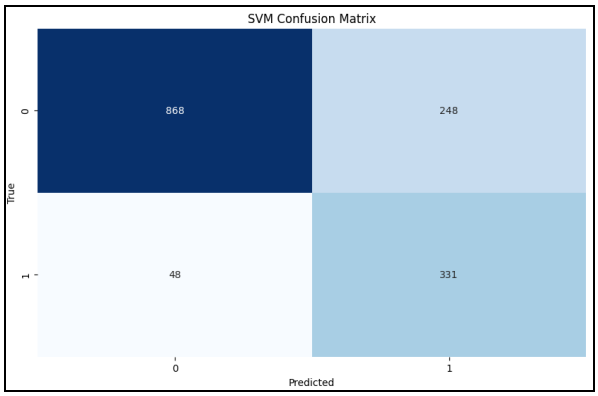

Figure 6: Confusion matrix of SVM

Web Development:

The web application integrates cutting-edge machine-learning techniques with user-friendly web development. Leveraging Flask, a web application with minimal overhead framework, and fundamental HTML and CSS, the intuitive interface enables users to input relevant health parameters. The machine-learning model examines the information to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

provide real-time predictions regarding liver health. By harnessing supervised learning, accurate results aid in early diagnosis and personalized treatment recommendations. The simplicity of the interface ensures accessibility for both medical professionals and individuals seeking self-assessment. The "Liver Disease Predictor" project contributes to advancing healthcare analytics and improving patient outcomes.

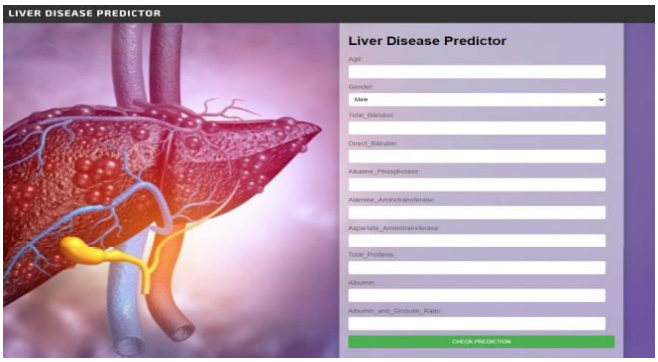

Figure 12: Prediction Webpage

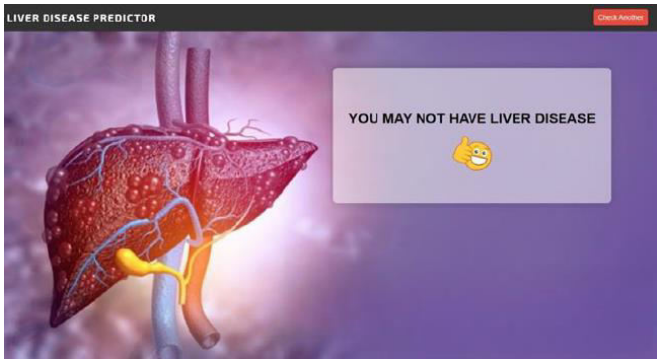

Figure 13:Result Webpage

Table 3 illustrates the comparison among various studies done on the topic in recent years with the proposed model of the paper. It can the observed from the comparison that the proposed model shows much higher accuracy compared to the studies done using other machine learning models to predict LD patients.

Table 3. Comparison result with existing system

Author	Year	Dataset	Models	Accuracy
Thaiparnit et al. [22]	2018	ILPD	RF	75.76 %
Poonguzharselvi et al. [21]	2021	UCI	RF	84%
Karasu et al.[14]	2018	ILPD	LG	73.97 %
Singh et al.[13]	2019	ILPD	LG	72.50 %
Saima et al. [11]	2021	ILPD	DT	94.28%
Rabbi et al.[4]	2020	ILPD	AdaBoost	92.19%

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205173 |

VIII. CONCLUSION

In conclusion, the application of supervised machine learning techniques for predicting and classifying liver disease offers a promising avenue within the realm of healthcare. By employing algorithms trained on labeled datasets, this methodology exhibits considerable potential in improving diagnostic accuracy and enabling timely interventions. Through the analysis of various patient parameters, including clinical history, laboratory results, and imaging data, these models can discern patterns indicative of different liver conditions with a high level of precision.

Moreover, the integration of such systems into clinical practice carries significant implications, empowering healthcare professionals to make well-informed decisions regarding patient management and treatment plans. By accurately identifying individuals at risk of liver disease or stratifying the severity of existing conditions, these models contribute to early intervention efforts and ultimately enhance patient outcomes.

REFERENCES

[1] K. Gupta, N. Jiwani, N. Afreen, and D. D, "Liver Disease Prediction using Machine-learning Classification Techniques," 2022 IEEE 11th International Conference on Communication Systems and Network Techkumnologies (CSNT), Indore, India, 2022, pp. 221-226, doi: 10.1109/CSNT54456.2022.9787574.
[2] A. Sivasangari, B. J. Krishna Reddy, A. Kiran and P. Ajitha, "Diagnosis of Liver Disease using Machine-learning Models," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 627-630, doi: 10.1109/I-SMAC49090.2020.9243375.
[3] Chieh-Chen, Wu., Wen-Chun, Yeh., Wen-Ding, Hsu., Mohaimenul, Islam., Phung, Anh, Nguyen., Tahmina, Nasrin, Poly., Yao-Chin, Wang., Hsuan, Chia, Yang., Yu-Chuan, Jack, Li. (2019). Prediction of fatty liver disease using machine-learning algorithms. Computer Methods and Programs in Biomedicine, doi: 10.1016/J.CMPB.2018.12.032.
[4] Md., Fazle, Rabbi., S., M., Mahedy, Hasan., Arifa, Islam, Champa., Md., AsifZaman., Md., Kamrul, Hasan. (2020). Prediction of Liver Disorders using Machine-learning Algorithms: A Comparative Study. doi: 10.1109/ICAICT51780.2020.9333528.
[5] S. N. N. Alfisahrin and T. Mantoro, "Data Mining Techniques for Optimization of Liver Disease Classification," 2013 International Conference on Advanced Computer Science Applications and Technologies, Kuching, Malaysia, 2013, pp. 379-384, doi: 10.1109/ACSAT.2013.81.
[6] S. Ambesange, R. Nadagoudar, R. Uppin, V. Patil, S. Patil, and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques," *2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC)*, Vijiyapur, India, 2020, pp. 1-6, doi: 10.1109/B-HTC50970.2020.9297949.
[7] V. Singh, M. K. Gourisaria and H. Das, "Performance Analysis of Machine-learning Algorithms for Prediction of Liver Disease," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-7, doi: 10.1109/GUCON50781.2021.9573803.
[8] Elias, Dritsas., Maria, Trigka. (2023). Supervised Machine-learning Models for Liver Disease Risk Prediction. Computers, doi 10.3390/computers12010019.
[9] S.Venkata Balaji, N.Aneel, .Jagadeesh, M.Devendran. (2023). Liver Disease Prediction Using Machine-learning. https://www.ijfmr.com/papers/2023/3/2955.pdf
[10] Fahad, B., Mostafa., Easin, Hasan. (2021). Machine-learning Approaches for Binary Classification to Discover Liver Diseases Using Clinical Data. arXiv: Machine-learning,

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details