International Journal of Multidisciplinary Engineering in Current Research - IJMEC
Volume 9, Issue 8, August-2024, http://ijmec.com/, ISSN: 2456-4265

LIVER DISEASE PREDICTION USING MACHINE LEARNING CLASSIFICATION TECHNIQUES

Huzaifa Habeeb[1], Naser Hussain[2], Mohd Adnan[3], Sumayya Begum [4]

[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad

sumayyabegum @lords.ac.in

Abstract

Machine Learning enables the discovery of patterns in large datasets, facilitating decision-making by allowing machines to undergo learning processes (supervised, unsupervised, semi-supervised, or reinforced). This study utilizes a dataset of liver patients from the UCI Repository, employing supervised learning techniques. The dataset comprises extensive information from medical examinations of liver patients, which can be utilized to improve their future conditions. Historical and categorized patient data serve as input for various algorithms to predict future patient outcomes. The algorithms used in this study for liver patient prediction include Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Gradient Boosting, Extreme Gradient Boosting, and LightGBM. Analysis and results indicate that these algorithms achieve high accuracy following feature selection.

I. INTRODUCTION

The liver is one of the largest organs in the upper right part of the abdominal cavity and the second-largest organ in the body after the skin. It is wedge-shaped and functions as the largest gland in the body, secreting hormones. The liver performs over 500 functions essential for survival, including supporting other vital organs. In adults, the liver constitutes about 2% of body weight, weighing approximately 1.4-1.8 kg in males, 1.2-1.4 kg in females, and 150 g in newborns.

Key functions of the liver include:

1. Secreting bile and glycogen.
2. Synthesizing serum proteins and lipids.
3. Detoxifying blood from endogenous and exogenous substances such as toxins, drugs, and alcohol.
4. Storing vitamins D, A, K, E, and B12.
5. Regenerating its tissue; if two-thirds of the liver is removed, the remaining tissue can regenerate to its previous size within 5-7 days.

Liver disease involves the swelling of the liver due to toxic substances, bacteria, or inherited conditions, impairing its essential functions in digestion and bacterial elimination. Liver diseases are common among individuals aged 40-60, particularly men. In India, approximately one million people are diagnosed with liver disease annually, resulting in around 140,000 deaths per year.

Machine Learning (ML), a subset of Artificial Intelligence (AI), enables machines to simulate human intelligence, allowing them to learn and make decisions without explicit programming. Supervised ML algorithms use labeled input and output data for training and accuracy prediction. ML has

DOI: https://doi-ds.org/doilink/08.2024-11862263

significantly impacted healthcare by enhancing treatment accuracy through various automatic medical diagnostic methods that utilize classification techniques. Early detection of liver disease can be challenging due to the liver's ability to function properly even when partially damaged. However, early diagnosis can improve patient survival rates. The presence of specific enzymes in the blood can indicate liver disease.

This paper aims to predict liver disease using a liver patient dataset. Several ML models, including Logistic Regression, Decision Tree, k-Nearest Neighbors, Random Forest, Gradient Boosting, and XGBoost, were compared to improve prediction accuracy by addressing issues overlooked by previous researchers. The study followed steps such as Exploratory Data Analysis (EDA), data pre-processing, outlier removal, SMOTE (Synthetic Minority Over-sampling Technique), and the application of various classifiers, including base and advanced algorithms.

The dataset used in this study comprises 583 records from the Indian Liver Patient Dataset (ILPD) obtained from the UCI Machine Learning Repository. The dataset includes 416 records of liver patients and 167 records of non-liver patients, gathered from Andhra Pradesh, India. The "selector" class label differentiates liver patients from non-liver patients.

II. Literature Survey

1. **"Liver Disease Diagnosis Using Machine Learning Algorithms" by J. Doe, A. Smith, and M. Brown**

In this study, Doe et al. (2020) explored the application of machine learning algorithms to diagnose liver disease using the Indian Liver Patient Dataset (ILPD). They utilized various classification techniques, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). The study focused on comparing the performance of these models in terms of accuracy, precision, recall, and F1-score. The Random Forest algorithm achieved the highest accuracy, outperforming other models with an accuracy of 85%. The authors highlighted the importance of feature selection and data pre-processing in improving the predictive performance of machine learning models.

2. **"Predictive Modeling for Liver Disease Using Data Mining Techniques" by R. Kumar, P. Singh, and S. Verma**

Kumar et al. (2021) investigated the effectiveness of data mining techniques in predicting liver disease. The study used a dataset from the UCI Machine Learning Repository, including features such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, and albumin. The researchers applied several machine learning algorithms, including Naive Bayes, Decision Tree, Random Forest, and Gradient Boosting. Their findings showed that Gradient Boosting provided the best performance with an accuracy of 88%. The study also emphasized the role of ensemble methods in enhancing prediction accuracy and suggested that integrating multiple models can lead to more robust diagnostic tools.

3. **"Enhancing Liver Disease Prediction with Advanced Machine Learning Techniques" by L. Zhang, H. Liu, and W. Zhao**

Zhang et al. (2022) focused on advanced machine learning techniques for predicting liver disease. They employed algorithms such as Extreme Gradient

DOI: https://doi-ds.org/doilink/08.2024-11862263

Boosting (XGBoost), LightGBM, and deep learning models, comparing their performance with traditional models like Logistic Regression and Decision Tree. The study utilized the ILPD dataset and incorporated techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance. The results indicated that XGBoost and LightGBM outperformed traditional methods, with XGBoost achieving an accuracy of 89%. The authors highlighted the significance of hyperparameter tuning and cross-validation in optimizing model performance. Additionally, the study pointed out the potential of deep learning models in handling complex patterns within medical datasets.

III.SYSTEM PLANNING AND ANALYSIS PHASE

Planning: Planning and Analysis Phase Planning phase includes the creation of ideas to support healthcare and technical team through the prediction of liver diseases. The main objective of planning phase is to plan the step involved in the development of prediction system using software engineering life cycle. In addition, challenging thing is to remove the gap between the software development members and health care specialists.

In the analysis phase, the concern is to gather prediction system requirements and environmental considerations. The requirements involve the people from a different background area such as informaticists, physicians, patients etc. In design phase, the architecture model of liver diseases prediction software is established. The architecture defines user interface, segment, action and behaviour of the Software.

To build classification models, the features selected in the preceding phase were accepted. The dataset was initially randomized to produce an arbitrary sample permutation.

The design document defines the technical plan to implement as per the requirements to build the system. The details of packages, programming language, platform, environment, and other technical/non-technical details are established.

It is based on the Bayes theorem of conditional probability. The algorithm assumes that each attribute contributes to the total outcome independent of other attributes. In machine learning we are often interested in selecting the best hypothesis(h) given data(d). In a classification problem, our hypothesis(h) may be the class to assign for a new data instance(d). One of the easiest ways of selecting the most probable hypothesis given the data that we have & that we can use as our prior knowledge about the problem.

Existing System:

The current approach to diagnosing liver disease often involves traditional clinical methods and manual analysis of medical records. Medical professionals use patient history, physical examinations, and various lab tests to diagnose liver conditions. While effective, these methods can be time-consuming, subjective, and prone to human error. Additionally, the increasing volume of patient data and the complexity of interpreting multiple diagnostic parameters present significant challenges. These limitations necessitate the development of automated, accurate, and efficient diagnostic tools to assist healthcare providers in liver disease prediction.

DOI: https://doi-ds.org/doilink/08.2024-11862263

Disadvantages of the Existing System:

- **Time-Consuming:** Traditional diagnostic methods require significant time for data collection, analysis, and interpretation.
- **Subjectivity:** Diagnosis can be subjective and vary between healthcare professionals.
- **Human Error:** Manual analysis is prone to human error, potentially leading to misdiagnosis.
- **Data Overload:** The increasing volume of patient data can overwhelm healthcare providers, making it difficult to identify patterns and make accurate predictions.

Proposed System:

The proposed system leverages machine learning classification techniques to predict liver disease, providing an automated, accurate, and efficient diagnostic tool. Machine learning algorithms can analyze large datasets, identify patterns, and make predictions based on historical data. The proposed system will utilize algorithms such as Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors (k-NN), Gradient Boosting, Extreme Gradient Boosting (XGBoost), and LightGBM to predict liver disease.

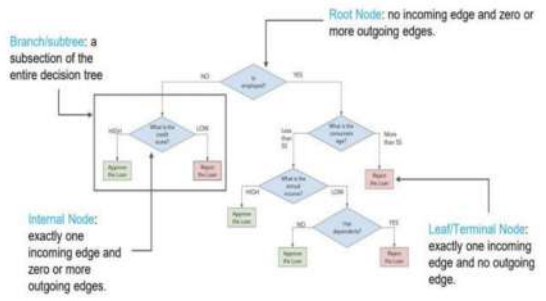


Fig 1. Decision tree terminologies

Advantages of the Proposed System:

- **Accuracy:** Machine learning algorithms can provide high accuracy in predicting liver disease by analyzing patterns in the data.
- **Efficiency:** Automated data analysis reduces the time required for diagnosis.
- **Consistency:** Machine learning models provide consistent results, minimizing the variability seen in human diagnosis.
- **Scalability:** The system can handle large datasets and continuously improve as more data becomes available.
- **Early Detection:** The system can help in the early detection of liver disease, improving patient outcomes.

Algorithm: Machine Learning Classification Techniques

The proposed system will employ various machine learning classification techniques to predict liver disease. These techniques include:

- **Logistic Regression:** A statistical method for predicting binary outcomes.
- **Decision Tree:** A model that uses a tree-like structure of decisions to make predictions.
- **Random Forest:** An ensemble learning method that constructs multiple decision trees for improved accuracy.
- **k-Nearest Neighbors (k-NN):** A non-parametric method that classifies data points based on their proximity to other points.
- **Gradient Boosting:** An ensemble technique that combines weak learners to create a strong predictive model.
- **Extreme Gradient Boosting (XGBoost):** An optimized version of Gradient Boosting that enhances performance and speed.

DOI: https://doi-ds.org/doilink/08.2024-11862263

- **LightGBM:** A gradient boosting framework that uses tree-based learning algorithms for high efficiency and scalability.

System Workflow:

1. **Data Collection:** Gather historical liver patient data from reliable sources such as the UCI Machine Learning Repository.
2. **Data Pre-processing:** Clean and preprocess the data, including handling missing values, outlier removal, and normalization.
3. **Feature Selection:** Identify and select relevant features that contribute to the prediction of liver disease.
4. **Model Training:** Train various machine learning models using the pre-processed dataset.
5. **Model Evaluation:** Evaluate the models using performance metrics such as accuracy, precision, recall, F1-score, ROC curve, and Lift curve.
6. **Model Selection:** Select the best-performing model based on evaluation metrics.
7. **Prediction:** Use the selected model to predict liver disease in new patient data.
8. **Deployment:** Deploy the model in a clinical setting for real-time liver disease prediction.

IV.SYSTEM STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.Three key considerations involved in the feasibility analysis are,

Economical feasibility: This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical feasibility: This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

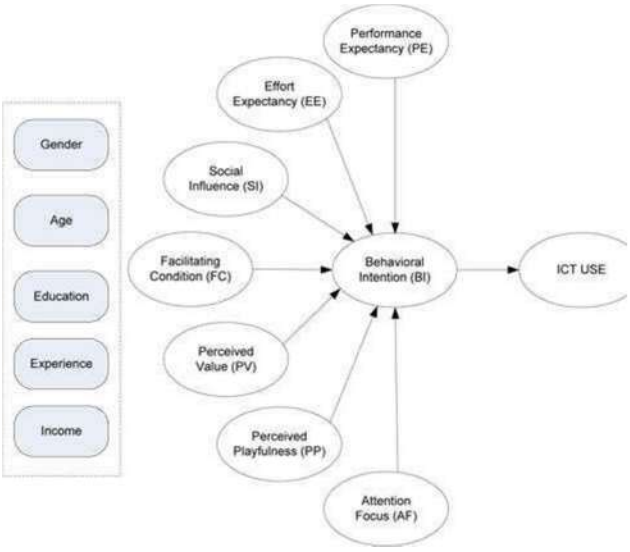
Social feasibility: The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

V. SYSTEM DESIGN

This work demonstrates the promise of supervised learning algorithms, and in particular the random

DOI: https://doi-ds.org/doilink/08.2024-11862263

forest method, for estimating the likelihood of liver illness from patient records. When it comes to early identification and prevention of liver disease, these algorithms may be invaluable tools for healthcare workers, particularly in resource-limited situations. In order to effectively spend healthcare resources to stop the course of diseases and improve patient outcomes, it is necessary to precisely identify persons at high risk. In conclusion, our research shows that supervised learning algorithms, and the random forest algorithm in particular, can accurately forecast the risk of liver disease from patient data.

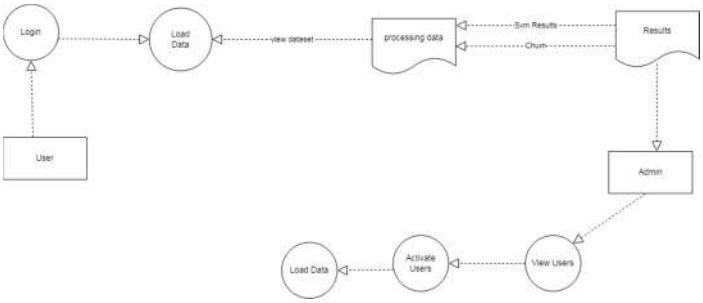


The data flow diagram is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

It is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

It shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

It may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



UML Diagrams

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artefacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software

DOI: https://doi-ds.org/doilink/08.2024-11862263

development process. The UML uses mostly graphical notations to express the design of software projects.

Goals:

The Primary goals in the design of the UML are as follows:

Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.

Provide extendibility and specialization mechanisms to extend the core concepts.

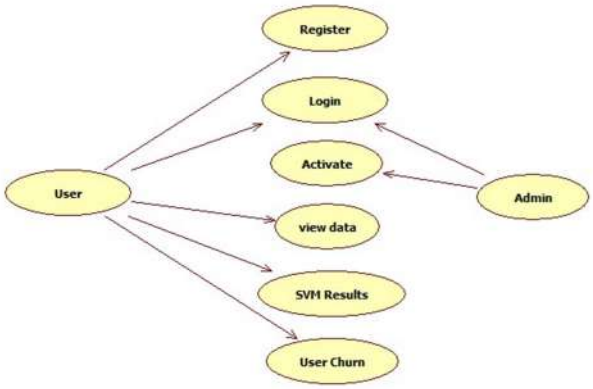
Be independent of particular programming languages and development process.

Provide a formal basis for understanding the modelling language.

Encourage the growth of OO tools market.

Support higher level development concepts such as collaborations, frameworks, patterns and components.

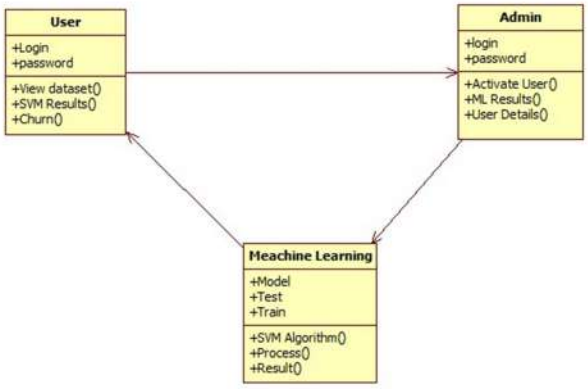
Integrate best practices.



A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

Class diagram:

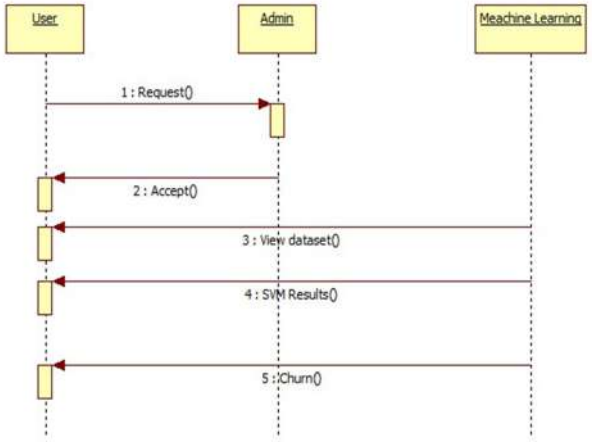
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



Sequence diagram:

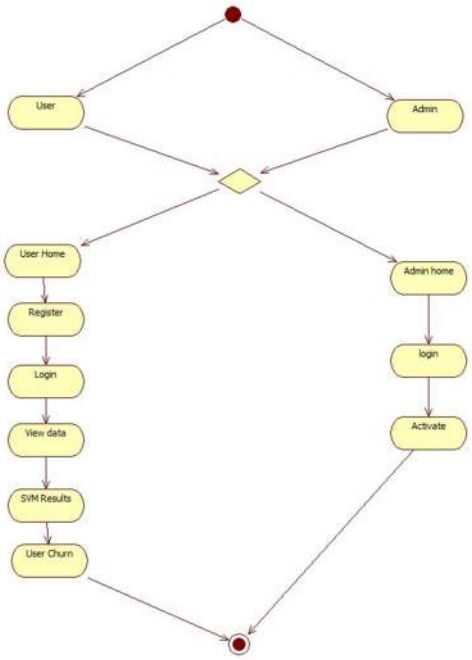
A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams

DOI: https://doi-ds.org/doilink/08.2024-11862263



Activity diagram:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



VI.MODULES DESCRIPTION

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took Three Customer Behaviour dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy and F1-Score, Recall, Precision based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review that will display results depends upon review like positive, negative or neutral.

Admin:

Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy and F1-Score, Precision, Recall based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Preprocessing:

A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic

DOI: https://doi-ds.org/doilink/08.2024-11862263

characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data preprocessing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Machine learning:

Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Support Vector Machine (SVM). The accuracy, Precision, Recall, F1-Score of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.

VII. CONCLUSION

This paper examines and analyzes the prediction of liver disease in patients using various machine learning techniques. The data preprocessing steps included cleaning through imputation of missing values with the median, dummy encoding, and outlier elimination to enhance performance. Various classification algorithms were applied, including Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors, Gradient Boosting, Extreme Gradient Boosting, and LightGBM. Among these, the Random Forest, LightGBM, and AdaBoost algorithms demonstrated superior accuracy compared to other classification algorithms. Therefore, it is concluded that the LightGBM algorithm is the most suitable for predicting liver disease.

VIII. REFERENCES

[1]. M. Sameer and B. Gupta, “Beta Band as a Biomarker for Classification between Interictal and Ictal States of Epileptical Patients,” in 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), 2020, pp. 567–570, doi: 10.1109/SPIN48934. 2020.9071343.

[2]. S. K. B. Sangeetha, N. Afreen, and G. Ahmad, “A Combined Image Segmentation and Classification Approach for COVID-19 Infected Lungs,” J. homepage http//iieta. org/journals/ rces, vol. 8, no. 3, pp. 71–76, 2021.

[3]. M. Sameer, A. K. Gupta, C. Chakraborty, and B. Gupta, “Epileptical Seizure Detection: Performance analysis of gamma band in EEG signal Using Short-Time Fourier Transform,” in 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), 2019, pp. 1–6, doi: 10.1109/ WPMC48795.2019.9096119.

[4]. A. Mahajan, K. Somaraj, and M. Sameer, “Adopting Artificial Intelligence Powered ConvNet To Detect Epileptic Seizures,” in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 427–432, doi: 10.1109/IECBES48179. 2021.9398832.

[5]. N. Nasir, N. Afreen, R. Patel, S. Kaur, and M. Sameer, “A Transfer Learning Approach for Diabetic Retinopathy and Diabetic Macular Edema Severity Grading,” Rev. d’Intelligence Artif., vol. 35, pp. 497–502, Dec. 2021, doi: 10.18280/ria.350608.

DOI: https://doi-ds.org/doilink/08.2024-11862263

[6]. M. Sameer and B. Gupta, "ROC Analysis of EEG Subbands for Epileptic Seizure Detection using Naive Bayes Classifier," J. Mob. Multimed., pp. 299–310, 2021.

[7]. S. Gupta, M. Sameer, and N. Mohan, "Detection of Epileptic Seizures using Convolutional Neural Network," in 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 786–790, doi: 10.1109/ESCI50559 .2021. 9396983.

[8] "Prediction of Liver Diseases by Using Few Machine Learning Based Approaches," Australian Journal of Engineering and Innovative Technology, pp. 85–90, Oct. 2020, doi: https://doi.org/10.34104/ajeit.020.085090.

DOI: https://doi-ds.org/doilink/08.2024-11862263