Applying Machine Learning Techniques in Liver Disease Detection: A Comprehensive Review

Aman Kumar
Department of Computer Science & Technology
IEC University, Baddi, Solan (H.P.)
Solan, India
Aman11304832@gmail.com

Dr. Randeep Singh
Department of Computer Science & Technology
IEC University, Baddi, Solan (H.P.)
Solan, India
randeeppoonia@gmail.com

Abstract

Machine learning makes use of artificial intelligence to create prediction models more quickly and effectively than traditional approaches by finding hidden patterns in massive amounts of data. With this in mind, these techniques can be used in a number of hepatology-related contexts. In this review, we look at the literature on machine learning in hepatology and the early diagnosis of liver illness. We give a general review of the benefits and drawbacks of machine learning techniques and discuss possible uses for them in liver disease prediction. We predict that the clinical practice of liver disease diagnosis will alter as a result of the application of ML techniques to produce prediction algorithms. Early detection of liver illness allows for timely diagnosis and, in some cases, a full recovery. The information in this review will give readers the chance to become more knowledgeable about the available machine learning (ML) techniques and their possible applications to liver diseases related problems.

Keywords*— *Data Mining, Liver diseases, Machine learning, Classification, Feature Selection

DOINumber:10.48047/nq.2022.20.22.NQ10283 **NeuroQuantology2022;20(22):2897-2906**

I. Introduction

"Medical data mining nowadays provides a wide range of opportunities as a result of the beginning of patient records being digitalized & being kept in digital formats [1]. Scientists can now do research electronically by sifting via this enormous information set gathered from the unplanned trials of a original life, as opposed to earlier times when people had to knock on doors to store information or conduct planned studies on volunteers. Healthcare data mining has the most potential because it allows health networks to systematically utilize information & exploration to detect ineffectiveness, enhance care, and lower costs [2]. Such possibilities are thought to improve healthcare and reduce costs to roughly 30% of total healthcare expenses.

The liver is the most important organ in humans because in addition to cleansing the blood, it also produces bile, excretes bile and bilirubin, regulates protein and carbohydrate metabolism, activates enzymes, stores vitamins, minerals, and glycogen, and produces plasma proteins and clotting factors. Our abdomen's top right corner contains the liver. Alcohol consumption, the use of painkillers, dietary choices, and engaging in several wired behaviors all have the potential to negatively impact the liver [3].

Any abnormality of the liver is referred to as liver disease. Numerous problems, including inflammation from viral and non-infectious sources, malignant tumors, liver scarring, and metabolic disorders are all part of this disease [4]. In addition to being potentially lethal, liver disease is also progressive and frequently asymptotic, necessitating staging. It is difficult but essential to accurately categorize liver diseases in order to stop their progression and prevent disorders such as hepatocellular carcinoma, which is thought to be irreversible.

Steatosis, also known as fatty liver infiltration, is regarded as the early phase of liver disease & might be

brought on by a substantial improvement in the quantity of fat in hepatocytes. It is asymptomatic,& hepatic damage can evolve into other serious disorders including fibrosis. Fibrosis manifests pathologically as a result of tissue or organ damage. The advancement rate of fibrosis, like that caused by chronic hepatitis, is significantly influenced by the source of liver illness. The final phase of practically all chronic liver disorders is cirrhosis. It is characterized by an asymptomatic period, also known as compensated cirrhosis, which is followed through a quickly progressing phase, known as decompensated cirrhosis, in which liver impairment occurs. Primary liver cancer, also named as hepatocellular carcinoma, is the most extreme development of cirrhosis.

If liver illnesses are not identified early enough, they may cause total liver failure, which eventually results in death. Consequently, we may identify liver illness by applying different data mining techniques. Currently, liver function blood test and scan findings are analyzed to identify disorders related to the liver. It is costly & extended [5]. It is feasible to speed up the liver disease diagnosis process by using various data mining methods to make it easier. The prediction will be more accurate if more data are used. In order to build a decision support model that could aid the doctor in predicting liver illness from the dataset, this article examines various data mining strategies.

The use of data science & technique to customize healthcare and enhance patient care delivery has more opportunities as the landscape of healthcare and information technology changes. Fundamentally, machine learning makes use of artificial intelligence to produce prediction models more quickly and effectively than traditional approaches by spotting hidden patterns in huge data sets [6]. In light of this, these techniques can be used in a number of instances to detect liver illness. In this study, we look at the literature on machine learning's tried-and-true uses in detecting liver illness.

The rest of this study is categorize into the following parts Section 2 presents an introduction of data preprocessing techniques used in machine learning, and Section 3 provides more information on classification techniques. The reviews of previously published works are covered in Section 4 while the survey's findings are covered in Section 5.

II. Data Preprocessing in Machine Learning

Prior to mining, a technique called data pretreatment can be used to enhance the quality of the information so that high-quality mining outcomes are generated. It is a data mining technology that transforms unstructured information into a format that could be used. Since raw data (information gleaned from the original world) is never comprehensive, models cannot process it. Many mistakes would be made as a result of this. Data must be preprocessed before being fed via a model because of this. Through the use of data processing techniques, it is possible to dramatically increase both the overall quality of the patterns that are mined and the amount of time required for the actual mining. The four methods for preparing data are data cleaning, data integration, data transformation, data reduction. There are many different types of data preparation challenges. You may consider one of these as feature extraction.

The dataset's properties are reduced using the feature selection technique to produce better results faster. Using a feature selection technique, it is necessary to eliminate redundant and irrelevant properties from the data without lowering accuracy. To prepare the dataset in a standard format, it is also necessary to delete any extraneous fields, missing records, and duplicate records. There are numerous feature selection techniques that can offer us several advantages, such as lessening lifting and training time [7]. Data can then be processed to create the proper format for the application of various classifiers.

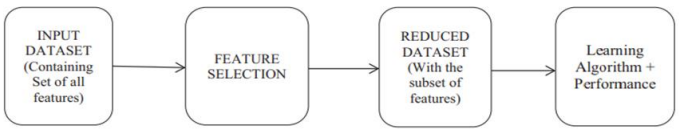

Fig. 1: Feature selection process [8]

Sachin Bagga et al [9] The primary aim of this work is to effectively diagnose liver illnesses by predicting and evaluating them using feature selection techniques & classification algorithms. The Indian Liver Patient Database from the UCI Repository was used to create a total of 583 instances with eleven unique properties. The Greedy Stepwise search strategy and Correlation-based FS Subset Evaluator were employed in this study. For improved outcomes, the following 5 attributes are chosen: TB, DB, Alkphos, sgpt, and Sgot. On a dataset of liver disorders, the classifiers were applied utilizing 10-fold cross-validation testing options. Different classification techniques, such as LR, NB, J48, SMO, IBk, RF, are used to compare the results. The production of every classifier is assessed depend on the cases that were successfully classified and the execution time. On a dataset of liver patient disorders, the outcomes of additional parameters including the kappa statistic, occurrences that were correctly identified, and mean

absolute error were also compared. Utilizing a classifier for logistic regression and feature selection, the best outcome was obtained.

Vijayarani et al. [10] They employ classification algorithms to estimate liver disorders in their study. The suggested approach makes use of well-known methods like NB & SVM. The database was collected from the UCI repository &contains areas for gender, Sgot, ALB, ALP, and DB, among others. According to their most recent research, SVM has the highest accuracy and Nave Bayes has a good execution speed.

III. Classification Approaches Based on Machine Learning

The literature on medical data mining demonstrates that numerous researchers employed various classifier systems to forecast chronic diseases in order to obtain accurate diagnostic outcomes. The machine learning classification methods that are used to identify diseases of the liver are covered in this section. A system for collecting and evaluating data is machine learning. It has the benefit of being able to offer universally applicable solutions. Due to its multidisciplinary approach, it is significant in a number of areas, including technology, healthcare, and engineering. Its advancements have helped to solve a number of problems with liver disease identification. ML enhances computational model performance and cost by automatically processing massive volumes of data [11]. ML approaches could be classified into 3 categories: reinforcement learning, unsupervised learning, both. The following figure shows a few of the classification strategies for detecting liver disease.

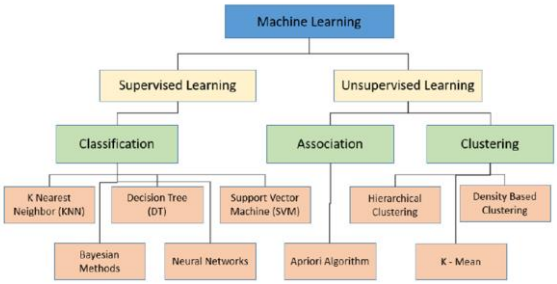

Fig. 2: Classification algorithms in ML [12]

A. *Unsupervised Classification:*

Unsupervised learning describes machine learning methods that extract knowledge from data without labelling. Without any prior training, unlabeled and unknown data are given to AI systems as model input. Unsupervised learning techniques can outperform supervised learning strategies when dealing with extremely complex processes, but the results can be quite unexpected; they have the potential to classify data into unknown types even when they don't exist, which could result in a huge mess. The most popular unsupervised learning method, known as clustering, is utilized for exploratory information evaluation to identify underlying patterns or groupings in information. Even with unsupervised learning, the model that is produced captures relationships, but there is no connection between the output and the inputs. The learning method groups together patterns that are related. The unsupervised clustering techniques Fuzzy-C-Means, K-Means, Hierarchical-based, and K-Medoids are a few examples. The methods for successful disease diagnosis using unsupervised learning are briefly given here.

Abbet et al [13] presented SRMA, which eliminates the need for fully-labeled source datasets by performing domain adaptation through the use of unsupervised learning. Without the need for further tissue annotations, SRMA may successfully transfer the discriminative information gleaned from a small sample of labeled source domain data to a new target domain. Through storing visual similarity with intra-domain & cross-domain self-supervision, suggested approach takes advantage of the structures of both domains. Additionally, we offer a generalized evaluation of suggested strategy that enables the structure to absorb knowledge from many source domains. We demonstrate that, in both single-source and multi-source contexts, our suggested approach superior baselines for domain adaptation of colorectal tissue type categorization, & we further test suggested method using an internal clinical cohort.

Taheri, et al [14] For pathway analysis, we provide a two-stage unsupervised machine learning approach in this paper. Four informative gene sets that may indicate significant COVID-19-related pathways are chosen during the first step. These "representative genes" are linked to the pathophysiology of COVID-19. Then, two separate networks for COVID-19-related illness and signaling pathways were built. The routes of each network are ranked in the second stage based on our defined informative features and some unsupervised scoring approach. We conclude by providing a thorough study of the most crucial pathways in both networks.

B. *Supervised Classification:*

Various inputs and outputs are offered, and during the training phase, relationships between them are discovered. Forecasting anticipated results and building a model that captures the dependencies and relationships among incoming data are the major goals. The output measure could be categorical or quantitative. The model is learned using a set of training

samples. Learning a function that converts inputs to outputs is the goal of the classifier training process. After training, if the algorithm is successful, it may recognize labels for unknown data that are presented during testing. These algorithms use training rules to increase classification task accuracy. Training algorithms use artificial neural networks based on perceptrons, KNN, SVM, & Bayesian approaches.

The many supervised learning-based approaches for detecting liver illness successfully are explored further down.

P. Kuppan et al. [15] With the use of NB, DT, & J48, authors have worked on analyzing data pertaining to Liver Disorder in their research effort. Moreover, properties like the patient's medical history, diabetes, alcohol consumption, smoking were used. Depend on the provided database, it has been determined that males are more likely than females to experience liver disorders. Most affected are those in the age range of 35 to 65, and of these, 26% have the disorder as a result of alcohol usage, 22% as a result of smoking, 4% as a result of obesity, and 5% as a result of diabetes.

A. Gulia et al. [16] Researchers have classified liver patient data in their proposed work utilizing algorithms such the BN, SVM, J48, MLP, & RF. The Center for ML & Intelligent devices has utilized data from the UCI repository. The RF Approach is the best one, with a conclusion reached after the conclusion of their three-phase examination, with an accuracy of 71.87%.

EI-Shafeiy at al. [17] Metabolomics evaluation are some of the digital health data used in their research on electronic health records, and as time has gone on, big data has taken on a different form. In the current study, a dataset of 7000 patients with 23 features, 5295 of them were men and the rest were women, was employed. They use Boosted C5.0, Support Vector Machine, NB, and FS in their suggested study.

C. Reinforcement Learning:

Reinforcement learning techniques employ an algorithm that continuously learns from the way its environment interacts with the data it collects. It chooses the best result, which helps the classification model's performance be optimised.

Salgueiro et al [18] The study suggests using cell phones to monitor patients and automatically classifying them based on the information they acquire. The system is predicated on the premise that Parkinson's patients who do not take their prescribed medications would exhibit various anomalies when walking. While the patient is receiving care, the cell phone is passively gathering the data. Then, with the aid of the data preprocessor, the walking cycles contained in this Parkinson-related biomarker are extracted. Deep Reinforcement Learning is the algorithm that has been suggested for categorization and Medication Adherence controlling. With this work, we show the viability of passive medication adherence monitoring and dynamic treatment regimens, as well as the monitoring of the biomarker walking in Parkinson's patients, utilizing mobile phones.

Dong Zhang et al [19] We suggest a novel pixel-level graph RL (Pix-GRL)approach in this paper. These images are contras able to conventional Gd-enhanced liver tumor pictures since our technology immediately converts standard nonenhanced liver images into AI-enhanced liver tumor images. Each pixel in Pix-GRL contains a pixel-level agent, which investigates the properties of the pixel & produces pixel-level actions to repeatedly change the pixel value, finally producing AI-enhanced photos of liver tumours. The graph convolution that Pix-GRL cleverly embeds to indicate all of the pixel-level agents is most significant. The agent is given a graph convolution to use for feature exploration, which increases efficacy by aggregating long-scale contextual data, & for action output, which increases efficiency by allowing agents to share parameter training. Additionally, our Pix-GRL approach employs a novel reward to quantify pixel-level action in order to dramatically enhance production through taking into account the enhancement in every action in every pixel with respect to both its own future state and that of nearby pixels. Pix-GRL, the first DRL approach for medical image synthesis, considerably improves upon the current medical DRL approaches by switching from a single agent to numerous pixel-level agents. Extensive tests using 325 patients (24,375 photos) and 3 kinds of liver tumour databases (benign, malignant, healthy controls) demonstrate that suggested unique Pix-GRL approach surpasses current medical image synthesis learning techniques. Regarding the size of the tumour, it obtained an SSIM of 0.85 0.06 & a Pearson correlation coefficient of 0.92. These findings demonstrate the possibility of creating an effective clinical replacement for Gd-enhanced liver MR imaging.

TABLE 1: ML CLASSIFIERS FOR DISEASE PREDICTION

Classifier	Merits	Demerits
Naïve Bayes	Low computational costs Effective with both numerical and nominal data	An extremely large number of records are necessary to get good results.

Fuzzy C means	gives best outcome for group of overlapping data	Specification of no. of clusters beforehand.
C4.5	effectively uses both continuous and categorical features executes huge application using less memory.	Issue with Overfitting
XGBoost	Easy to read and interpret algorithm. A durable approach that quickly reduces overfitting.	Because every classifier is required to correct the mistakes made by the predecessors, it is sensitive to outliers.
k-Means	Scales to large data sets. Guarantees convergence.	selecting k manually. scaling across multiple dimensions. Clustering outliers.
Random forest	works effectively for both continuous and categorical variables, Takes care of missing values quickly.	Extended training time, Complexity
Support vector machines	Lesser over fitting Resilient to noise	Computationally expensive
Bayesian networks	Provides extensive support with missing data	Necessitates initial Under- standing of a variety of probabilities
Logistic regression	incredibly quick at categorizing unknown records. Simple to use, understand, and very effective for training	In rare situations, it could result in overfitting. Problems with nonlinearity cannot be solved
k-nearest neighbour	Simple to comprehend and use	Memory restriction Takes long time span to run

IV. LITERATURE REVIEW

Because of its complexity, diversity, and vast volume of data, ML in healthcare is one of the main difficult application areas. It is difficult to forecast disease from such heterogeneous and fragmented data since the health care data sets that are provided are heterogeneous in character. This section provides a thorough literature analysis of numerous machine learning approaches utilize to forecast various diseases.

Durai et al [20] This process is divided into 5 main steps. The actual hepatic patient database that might be retrieved from the UCI repository is first subjected to the min-max algorithm. Significant attributes are distinguished in the second phase via PSO feature selection. This aids in separating the essential data from the liver patients' entire normalized datasets. Following this, the next stage involves categorization and comparison analysis using classification algorithms. The fourth phase is accuracy calculation. It makes use of the RMS & the Root Error values. The evaluation stage is the fifth stage. According to the studies, a straightforward analyzation procedure is carried out to maintain the unity of an accurate outcome reflection. With an accuracy rate of 95.04%, the J48 approach is thought to perform better when it comes to feature selection.

P. Rajeswari et al [21] implemented data mining algorithms to analyze liver diseases. The classification of the information in this research is depend on liver disorders. The training data set, made up of 345 instances with 7 different features, was created by gathering data from the UCI repository. The examples in the database relate to the 2 types of blood tests that are believed to be sensitive to liver disorders that may result from excessive alcohol usage. Methods that control hepatic epithelial cell differentiation have been crucial in developing effective cell culture approaches for programmed differentiation of stem cells to hepatocytes & cell transplantation therapies. Such information also gives the Low (L) and High (H) ions the foundation they need to accurately and quickly create the algorithm by representing the profit as 0 and 1. The data is categorized using the WEKA tool, tested using 10-fold cross validation, and the outcomes are contrasted. Three algorithms FT-Tree, K-Star, NB are contrasted in this study. The outcomes demonstrate that FT-Tree methods achieve maximum accuracy of 97.10%, and K-Star algorithms achieve minimum accuracy of 83.47%.

Dr. S. Vijayarani & Mr. S. Dhayanand [22] The NB & SVM classification methods were utilized in this research to estimate liver illness. These algorithms are contrasted, & performance metrics used are accuracy and execution time. SVM is regarded as best algorithm based on experimental findings since it has highest

classification accuracy (79.6%), followed by NB with 61.2%. NB requires least amount of execution time, at 1670 milliseconds, whereas SVM requires 3210 milliseconds.

Mehdi Teimouri et al. [23] Through classification of outpatient medications, this study seeks to determine prevalence of outpatient diseases. 1412 prescriptions for various conditions made up source of data for this investigation, from which authors isolated ten diseases and identified them. In this study, ailments for which prescriptions are prescribed are identified using data mining technologies. We compare the outcomes with the Nave technique in order to assess how well these strategies perform. The results are then enhanced by merging approaches. Results indicated that Support Vector Machine conducted better as compared to the other approaches, with an accuracy of 95.32%. Accuracy of NB is 67.71%, which is 20% less accurate than that of KNN method, the least accurate of four classification methods. Outcomes show that application of data mining techniques produced good performance in classifying diseases that affect outpatients. These findings can aid in selection of suitable techniques for classifying prescriptions on a broader scale.

M. Banu Priya et al [24] In this research, datasets from liver patients are looked at for purpose of creating classification models to foretell liver illness. In three steps, feature model design and comparative analysis used in this thesis improved prediction accuracy for Indian liver patients. Beginning stage consists of implementing min-max normalization technique to UCI repository's original liver patient datasets. In second stage of liver dataset prediction, PSO FS is utilize to acquire subpart of entire normalized liver patient dataset that only consists important features. In third phase, database is put through to categorization methods. Accuracy would be determined in fourth stage using RMS & RME values. After using PSO feature selection, J48 method is thought to perform better. The evaluation is then completed using accuracy values. As a result, outputs from suggested classification solutions demonstrate that J48 method outperforms all other classification algorithms with accuracy of 95.04% thanks to feature selection.

Mafazalyaqeen, et al [25] In order to aid professionals and their patients in identifying symptoms of illness, shorten time it takes to diagnose it, and prevent passings, this study proposes a new method for liver illness analysis. Proposed method would combine Genetic Algorithm (GA) with principles released by Boosted C5.0 grouping methodology to increase determination time and accuracy. So hereditary calculation is used to strengthen and weaken rules of another measurement rather than utilizing a transformational calculation to create regulations. In contrast to existing work in sector, we demonstrate that our suggested methodology has greater execution & throughput. Their work shows an increase in precision from 81% to 93%.

Parminder Kaur et al [26] used random trees to categorize disorders based on liver like fatty liver, cholestatic disease, hereditary disease, and autoimmune disease. Hospitals provide the data for collection. This study demonstrates how decision trees are used to simulate liver cancer true diagnosis for surgical and non-surgical treatment. For classification of diseases based on liver, random tree method creates rules and decision trees. Weka's random method creates decision trees that enable us to identify disease based on its characteristics and symptoms.

Balakrishnan et al [27] introduced an innovative technique to feature selection for the type II diabetes dataset using SVM ranking and a backward search strategy. Naive Bayes classifier's predictive accuracy greatly increased with the suggested strategy. The procedure employed was quite straightforward yet efficient, and it would undoubtedly aid doctors and other medical professionals in identification of Type 2 diabetes.

Verma et al [28] presented Correlation-based feature selection strategy and used it in hybridised design to diagnose Coronary Artery Disease. Utilizing correlation feature selection strategy PSO and clustering algorithm, most important risk variables for CAD disease were found. C4.5 method, multi-layer perceptrons, multinomial LR, & fuzzy unordered rule induction approach were utilized to build diagnostic models for CAD disease. 10-fold cross validation approach was utilize to validate CAD model. On both clinical data & data related to Cleveland heart disease, MLR algorithm had highest predicted accuracy, whereas MLP method had lowest. The findings of suggested methods were highly encouraging and considerably increased classifier accuracy. As a result, this technique can be useful tool for clinical judgments including diagnosis of CAD condition.

Das et al [29] presented Boosting Based Hybrid Feature Selection, a quick and scalable hybrid technique that combined benefits of both filter & wrapper approaches. Through including forward selection approach & some of advantages of wrapper technique, like natural stopping criterion, authors created more

informed filter method. When applied to DNA dataset using NB & chess database utilizing ID3 approach, this algorithm produced quick and superior results than previous methods. This method considerably improved these classifiers' accuracy. On datasets with large no. of features, suggested hybrid technique was found to be particularly scalable. In future, suggested approach can be used on datasets with several classes to produce more intriguing results. In further development of this study, using k-level DT rather than decision stumps could produce interesting outcomes.

Peng Y. et al [30] created hybrid FS methodology integrating filter & wrapper approach for classification of biomedical information. To address problems of high dimensional biomedical information & to enhance performance of SVM classifier, suggested method added feature pre-selection stage and utilized Receiver Operating Characteristics curves. P roposed strategy greatly enhanced classification performance, according to experimental results with biomedical databases, and its outcomes outperformed those of Sequential Forward Floating Search method by wide margin. Additionally, pre-selection process that was implemented helped to address over-fitting issues. Proposed approach has lot of potential for categorising biomedical data.

Gurbuz et al [31] By giving SVMs the ability to adapt, diseases could be diagnosed. With the aid of adaptive SVM, goal was to suggest quick and adaptive diagnostic method. To obtain better outcomes, bias value in conventional SVM was modified. Proposed classifier presented a set of "if-then" rules as its output. The developed approach was utilize to diagnose breast cancer and diabetes, and it provided 100% accurate categorization rates for each condition. Future research should focus on developing more effective techniques for modifying bias value in traditional SVM.

Patil et al [32] provided hybrid model for predicting type-2 diabetes based on clustering and classification. Suggested model uses k-fold cross-validation for prediction together with C4.5 classification algorithm and K-means clustering. Hybrid strategy enabled model to achieve outcomes, with an accuracy of 92.38%, which can be highly beneficial for doctors in making wise therapeutic decisions regarding diabetes. Future research can focus on creating more accurate disease diagnosis models.

Ubeyli et al [33] Use of multiclass SVM with error-correcting output codes allowed for diagnosis of erythemato-squamous illnesses . For same purpose, multilayer perceptron NNs and recurrent NNs were also employed. The goal was to identify six erythemato-squamous disorders using best categorization strategy. Even with outcomes acquired, it was discovered that performance of "Multiclass SVM" was best with 98.3% accuracy, "RNN was good with 96.6% high accuracy, but Multilayer Perceptron NN did not perform well and accuracy drastically decreased to 85.4%. With these findings, authors came to conclusion that "Multiclass SVM" and "RNN" could be employed for erythemato-squamous illness diagnosis.

Yang et al [34] Adaptive Support Vector Machines was suggested approach that attempted to adjust classifiers to various data distributions. By incorporating adaptability into traditional SVM method, suggested technique was able to improve classifier efficiency. The crucial aspect was that it enabled classifiers to be adjusted to any form of dataset. This was accomplished by adding "delta function" to traditional SVM. Using results, it was discovered that ASVM performed more effectively than ensemble approach & other adaption strategies.

Tsipouras et al. [35] presented DSS based on fuzzy rules for automated detection of coronary artery disease. The process was broken down into four steps. Using C4.5 technique, decision tree was first built from training dataset. In order to provide clear model, rules were taken out of tree. Then crisp rules were converted into fuzzy model. Results showed that classification accuracy increased dramatically from 5% to 35% when the fuzzy model was optimized. Additionally, proposed method outperformed classification outcomes of ANN and ANFIS algorithms. The method can aid clinicians in making conclusions regarding presence or absence of CAD illness.

Table 2 provides a thorough analysis of prediction algorithms for detection of diabetes, heart disease, kidney, and liver disorders.

TABLE 2. COMPREHENSIVE REVIEW OF MINING TECHNIQUES FOR VARIOUS DISEASES

Author Name	**Techniques used**	**Highest Accuracy Model**	**Accuracy**
M. A. Zriqat et al [36]	NB, Discriminant, Decision Tree, Random Forest, and SVM	Decision tree and Random Forest	99.0%
M. E. Emre et al [37]	NB, Decision Trees i.e. C5.0 boosted, CART, C4.5, and C5.0.	CART	87%

	random forest algorithms		
M. Azrar [38]	Decision Tree, NB, KNN,	Decision Tree	75.65%
B. S. Kumar [39]	J48, CART and Naïve bayes	J48 and CART is cost efficient	99 %
M. L. Z. Alkaragole et al [40]	Bayesian, SVM, Decision Tree	Hybrid proposed method of decision tree and SVM	94%
M. S. Gharibdousti et al [41]	SVM, J48, Naive Bayes	SVM	75.75%
S. Zeynu et al [42]	K-Star, SVM, NB and J48	J48	99%
E. Avci et al [43]	multilayer perceptron, naive bayes and J48 decision tree	J48	87.3 %
L. A. Nguyen et al [44]	Naive Bayes and SVM	SVM	76.32 %
Islam et al (45)	RF, SVM, ANN, LR	LR	76.3%
Wu et al (46)	RF, LR, ANN, NB	RF	87.5%
Sowa et al (47)	LR, kNN, SVM, DT, RF	RF	79%

V. CONCLUSION

Many data mining algorithms are effective at predicting different ailments, including heart, liver, kidney, and diabetes. Big data challenges statistical models, which have failed. Dealing with enormous data quantities requires the use of data mining. Machine learning algorithms are most frequently and widely utilized for disease prediction, according to an analysis of the literature currently in print. Biostatistical methodologies may be augmented by machine learning techniques to address issues in all areas of medicine. Though many questions in the detection of liver disease can be answered with normal biostatistics, ML can outperform this, especially when looking at clinical prediction questions. Hepatology has been using ML techniques more and more increasingly to investigate the vast amount of clinical, genetic, radiologic, and pathologic data related to liver illness. A more precise medical approach to the practise of hepatology is ultimately anticipated as a result of applying these techniques to unravel the complexity of liver disease and improve the identification of more ideal biomarkers and therapeutic approaches.

References

[1] Durairaj, M., Ranjani, V. (2013).“Data mining applications in healthcare sector: a study”. International journal of scientific technology research, 2(10), 29-35.

[2] Teimouri, M. et al, “Detecting diseases in medical prescriptions using data mining tools and combining techniques”, Iranian J. Pharm. Res. 15, 113–123 (2016)

[3] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis”, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114

[4] Arshad, Insha, et al. "Liver disease detection due to excessive alcoholism using data mining techniques." 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2018.

[5] E. Barrio et al, “Liver disease in heavy drinkers with and without alcohol withdrawal syndrome”, Alcohol Clinical Exp. Research, 28 : 131 – 136, 2004.

[6] W. Nanyue, Y. Youhua, H. Dawei, X. Bin, L. Jia, L. Tongda X. Liyuan, S, Zengyu, C. Yanping and W. Jia. (2015), “Pulse Diagnosis Signals Analysis of Fatty Liver Disease and Cirrhosis Patients by Using Machine Learning,” The Scientific World Journal, vol. 2015, Article ID 859192, pp. 1-9.

[7] A. M. Hall and A. L. Smith. (1999), “Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper”, In Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, AAAI Press pp. 235- 239.

[8] Jain, Divya, and Vijendra Singh. "Feature selection and classification systems for chronic disease prediction: A review." Egyptian Informatics Journal 19, no. 3 (2018): 179-189.

[9] Sachin Bagga, Singh, Jagdeep, and Ranjodh Kaur. "Software-based prediction of liver disease with feature selection and classification techniques." Procedia Computer Science 167 (2020): 1970-1980.

[10] S. Vijayarani, and S. Dhayanand. (2015) "Liver disease prediction using SVM and Naïve Bayes

algorithms." International Journal of Science, Engineering and Technology Research (IJSETR) vol. 4, no. 4, pp. 816-820.

[11] E. Frank, M. A. Hall, and I. H. Witten, The WEKA Workbench. Online Appendix for. (2016) "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

[12] Nabeel, Muhammad, Shumaila Majeed, Mazhar Javed Awan, Hooria Muslih-ud-Din, Mashal Wasique, and Rabia Nasir. "Review on Effective Disease Prediction through Data Mining Techniques." International Journal on Electrical Engineering & Informatics 13, no. 3 (2021).

[13] Abbet, Christian, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. "Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection." Medical image analysis 79 (2022): 102473.

[14] Taheri, Golnaz, and Mahnaz Habibi. "Comprehensive analysis of pathways in Coronavirus 2019 (COVID-19) using an unsupervised machine learning method." bioRxiv (2022).

[15] P. Kuppan, N. Manoharan. (2017) "A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes", International Journal of Computing Algorithm, vol. 6, no. 1, pp. 2278-239.

[16] A. Gulia, R. Vohra, P. Rani. (2014), "Liver Patient Classification Using Intelligent Techniques," International Journal of Computer Science and Information Technologies (IJCSIT), vol. 5, no. 4, pp. 5110-5115.

[17] A. El-Shafeiy, L. Ali. Engy, El-Desouky and S. M. Elghamrawy. (2018) "Prediction of Liver Diseases Based on Machine Learning Technique for Big Data." In International Conference on Advanced Machine Learning Technologies and Applications, pp. 362-374. Springer, Cham.

[18] Salgueiro, Armando de Jesús Plasencia, Yulia Shichkina, Arlety García García, and Lynnette González Rodríguez. "Parkinson's disease classification and medication adherence monitoring using smartphone-based gait assessment and deep reinforcement learning algorithm." Procedia Computer Science 186 (2021): 546-554.

[19] Dong Zhang, Xu, Chenchu, Jaron Chong, Bo Chen, and Shuo Li. "Synthesis of gadolinium-enhanced liver tumors on nonenhanced liver MR images using pixel-level graph reinforcement learning." Medical Image Analysis 69 (2021): 101976.

[20] Durai, Vasan, Suyan Ramesh, and Dinesh Kalthireddy. "Liver disease prediction using machine learning." Int. J. Adv. Res. Ideas Innov. Technol 5, no. 2 (2019): 1584-1588.

[21] P.Rajeswari et al, "Analysis of Liver Disorder Using Data mining Algorithm", GJ C ST, Volume 10, Issue 14 (Ver. 1.0) , P a g e no. 48, November 2010.

[22] S. Vijayarani et al, "Liver Disease prediction using SVM and Naïve Bayes Algorithms", IJSETR Vol. 4, Issue 4, April-2015.

[23] Teimouri, M. et al, "Detecting diseases in medical prescriptions using data mining tools and combining techniques", Iranian J. Pharm. Res. 15, 113–123 (2016)

[24] M. Priya et al, "Performance Analysis of Liver disease prediction using Machine Learning Algorithms", IRJET e-ISSN: 2395-0056 Vol. 05, Issue: 1, Jan 2018.

[25] Hassoon, Mafazalyaqeen, et al. "Rule optimization of boosted c5. 0 classification using a genetic algorithm for liver disease prediction." 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017.

[26] P. Kaur et al, "Classification of Liver based diseases using Random Tree", IJAET, June, 2015.

[27] Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. "SVM ranking with backward search for feature selection in type II diabetes databases. In: Systems, man and cybernetics", 2008. SMC 2008. IEEE international conference on. IEEE; 2008. p. 2628–33.

[28] Verma L, Srivastava S, Negi PC. "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data". J Med Syst 2016;40 (7):1–7.

[29] Das S. Filters, "Wrappers and a boosting-based hybrid for feature selection". In: ICML, vol. 1; 2001. p. 74–81.

[30] Peng Y, Wu Z, Jiang J. "A novel feature selection approach for biomedical data classification". J Biomed Inform 2010;43(1):15–23.

[31] Gürbüz E, Kılıç E. "A new adaptive support vector machine for diagnosis of diseases". Expert Syst 2014;31(5):389–97.

[32] Patil BM, Joshi RC, Toshniwal D. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37(12):8102–8.

[33] Ubeyli ED. Multiclass support vector machines for diagnosis of erythemato squamous diseases. Expert Syst Appl 2008;35(4):1733–40.

[34] Yang J, Yan R, Hauptmann AG. "Cross-domain video concept detection using adaptive svms". In: Proceedings of the 15th ACM international conference on multimedia. ACM; 2007. p. 188–97.

[35] Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV, Naka KK, Michalis LK. "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling". IEEE Trans Inf Technol Biomed 2008;12 (4):447–58.

[36] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," arXiv preprint arXiv:1704.02799, 2017

[37] I. E. Emre, N. Erol, Y. I. Ayhan, Y. Ozkan, and C. Erol, "The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining," International journal of medical informatics, vol. 123, pp. 68–75, 2019.

[38] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 320–323, 2018.

[39] R. C. S. Mary and B. S. Kumar, "Comparison of various data mining algorithms in the prediction of risk for gestational diabetes," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 8, p. 74, 2018

[40] M. L. Z. Alkaragole and A. P. S. Kurnaz, "Comparison of data mining techniques for predicting diabetes or prediabetes by risk factors," 2019.

[41] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of chronic kidney disease using data mining techniques," in IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE), 2017, pp. 2135–2140.

[42] S. Zeynu and S. Patil, "Survey on prediction of chronic kidney disease using data mining classification tech- niques and feature selection," International Journal of Pure and Applied Mathematics, vol. 118, no. 8, pp. 149–156, 2018.

[43] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "Performance comparison of some classifiers on chronic kidney disease data," in 2018 6th International Symposium on Digital Forensic and Security (ISDFS). IEEE, 2018, pp. 1–4

[44] P. A. A. Nguyen, C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. J. Li, "Prediction of fatty liver disease using machine learning algorithms," Computer methods and programs in biomedicine, vol. 170, pp. 23–29, 2019.

[45] Islam MM, Wu C-C, Poly TN, Yang H-C, Li Y-CJ. "Applications of Machine Learning in Fatty Liver Disease Prediction". Stud. Health Technol. Inform. 2018; 247:166–170.

[46] [46] Wu C-C, Yeh W-C, Hsu W-D, Islam MM, Nguyen PAA, Poly TN, et al. Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. 2019; 170:23–29.

[47] [47] Sowa J-P, Heider D, Bechmann LP, Gerken G, Hoffmann D, Canbay A. Novel algorithm for noninvasive assessment of fibrosis in NAFLD. PLoS One. 2013;8:e62439.