Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Investigating Critical Care: Innovations in Disease Identification and Prognostication for Cardiovascular, Renal, Pulmonary, and Hepatic Health with Emphasis on MIMIC-III and MIMIC-IV

Shahnazeer C K, Dr. G Sureshkumar

Department of Computer Science, School of Engineering & Technology, Pondicherry University Karaikal Campus, Karaikal-609605, Puducherry (UT), India.
*Email: 1*shahnazeer.ck@pondiuni.ac.in*

This research investigates the application of healthcare datasets, particularly MIMIC-III and MIMIC-IV (Medical Information Mart for Intensive Care III & IV), to address the difficulties related to cardiovascular disease (CVD), renal disease, pulmonary disease, and liver disease. Medical care offers comprehensive data and critical information regarding the etiology and consequences of diseases. This research emphasizes the significance of medical data centers that manage and analyze data to anticipate and identify diseases. Specifically, it identifies the risk prediction methodologies and the obstacles encountered in cardiovascular disease detection, encompassing modifiable and non-modifiable risk factors. This research focuses on using datasets to detect and predict chronic kidney disease (CKD) risk, with a particular emphasis on early detection. Furthermore, it analyzes the techniques for categorizing hepatic and pulmonary disorders within the accessible dataset. This study seeks to enhance critical care utilization comprehension to tackle existing healthcare challenges through research outcomes.
Keywords: MIMIC-III, MIMIC-IV, CVD, CKD, lung disease, liver disease.

1. Introduction

The growing demand for electronic health records (EHRs) to store patient health data within international healthcare systems is propelling the proliferation of machine learning tools that

implement machine learning techniques to tackle the challenges faced in this field. The technical and biomedical sectors still need to work on providing reliable and valuable techniques to enhance an individual's health status efficiently, contrary to the immense scope and anticipation in this domain. Handling an enormous EHR dataset is one of the primary obstacles. Researchers and medical professionals face a significant challenge in accessing EHRs conveniently regarding security, interconnectivity, and privacy.

An effort made by MIMIC, which has been an exemplar in accessing massive EHR data sets explicitly in this domain, as well as the privacy concerns of EHR data, is resolved by de-identifying patients' health updates [1]. [2] Version four (MIMIC-IV) emerged in 2020 and has undergone various successive improvements to become this dataset's recent (prime) version.

MIMIC [3] still faces challenges in being part of this domain while operating in the tasks associated with data cleansing and preparatory processes, even though they are readily accessible. This problem is crucial because of the multidisciplinary technical and medical expertise when working with EHR datasets frequently. The mimic study emphasizes that researchers must follow ideal standards in data analysis, as they did not perform data cleansing procedures to guarantee the accuracy of the real-time clinical records represented in the dataset. Not adhering to the above-mentioned best practice guidelines can result in an inappropriate research layout that is unreliable, invasive, and biased. [4] Their recent study noted some instances of explicitly challenging MIMIC designs. In addition, it is very challenging to compare and recreate various research projects due to the absence of a consistent framework for extracting and preprocessing MIMIC-IV [5]. For MIMIC, cohort extraction and preprocessing protocols are available (examined in the following section). It emphasizes stabilizing particular preprocessing rather than offering an adaptable and customized protocol that abides by a wholly investigated method. A well-defined data preprocessing pipeline is necessary, which is user-friendly, adaptable for batch definition, and customizable by the user.

They focus primarily on this research to address the disparity above by offering a customized protocol designed to make MIMIC-IV data to employ in the forthcoming tasks. The preliminary data is purified by removing abnormalities, allowing the user to extrapolate the missing data. To limit the dimensions, it offers some choices for creating batches of clinical attributes by utilizing the typical coding method. The sequential data partitioning into some periods makes the time series filtration as per the user's desire, which leads to the generation of a seamless time-series dataset. Users can create the patient community as per their preferences since having the customized features. Researchers can replicate this research using this protocol but must record and share the customized procedures. In addition to the other preprocessing equipment in our pipeline, the modeling and evaluation procedure will incorporate two additional components. In the modeling process, the most popular ML tools and deep learning sequential models are appended in the pipeline to predict the data. The evaluation section includes numerous typical methods to evaluate the developed model's efficiency and has some alternatives to inspect and interpret the model's integrity. As a result, this pipeline enhances MIMIC usability. It is highly accessible to researchers, which can limit the dataset's cleaning, preprocessing, and access to processing time and expertise needed for having enormous studies in MIMIC-IV (around 300 in mid-2022) [6].

This research opt for the MIMIC-IV version 2.1 database utilized in BETH Israel Deaconess Medical Center. It comprises the patient records admitted from 2008 to 2019 [2] [1], which includes various parameters such as administrative, lab results, and medication data. The patient is assigned a unique identifier in MIMIC-IV based on the admission record. Once the hospital admits the patient, they migrate to various departments like the ambulatory surgery unit and emergency care. Finally, the patient is shifted to the ICU and indicated with an ICU identifier, which lasts until the patient is discharged from the ICU and transferred to another department in MIMIC-IV.

2. Significance/Challenges of MIMIC-III

Beth Israel Deaconess Medical Center gathered the explicitly accessible MIMIC-III database, including the patient's numerous health records from 2001-2012 [7]. It consists of the data regarding treatment procedures followed, primary health indicator readings of hospitalized patients, demographics of the specific individual, medication details, laboratory data, radiology results, medical recommendations, and fatality-related data. Developing electronic types of equipment and epidemiology for analyzing and treating critical illness has recently required the MIMIC-III database to be in great demand for optimizing decision strategy. It also contains the chronological information of the patients during the entire hospitalization. The data in tabular format stores the patient's identifier in columns like traditional spreadsheets. In contrast, the row indicates the specific information of the patient's identifier. Identifiers denote tables linked by the suffix 'id,' and they also use the prefix operator 'd,' which assists with logically interpreting the target identifier (e.g., dictionary).

2.1 The patient stays

This table [8] is so peculiar it contains the hospitalized information of the patient within the hospital premises. It includes various data related to admission details, period of ICU stays [8], patient's discharge endorsement, specific data about the patient, medical treatment followed, and specialized department migration within the hospital. Someone notices that this table has a particular connection. For example, an ICU identifier links to a single admission and patient identifier. Meanwhile, this patient's identifier connects to a hospitalized and ICU identifier of various hospitals.

2.2 Critical care

This table specifies the patient's information about who is hospitalized in the ICU[8]. It represents the information about the caretaker's identifier and all procedures undertaken in the medical chart. It also contains temporal data about time, date, events of clinical services, monitoring, and introduction of electronic tools for patient care. Additionally, it includes de-identifying information regarding the ECG report, radiological result, medical records, and outcome of the patient during their ICU hospitalization, discharge summary, and the overall treatment details given to the patient during hospitalization.

2.3 Hospital record system

This table [9] contains the stored data records within the medical care management. Also, it comprises the data regarding diagnosis as per the ICD standard and data about treatment procedures followed by indicating specific code details. It also contains the underlying diagnosis details, which help evaluate an invoice. As per ICD, the data related to varied lab results, microbiological analysis data, procedures provided to the patient, and organized prescription is noted in specific codes and stored in this table. Figure 1 and Figure 2 are the graphical representations of the patient distribution, admission number, admission distribution, and ICD-9 codes.

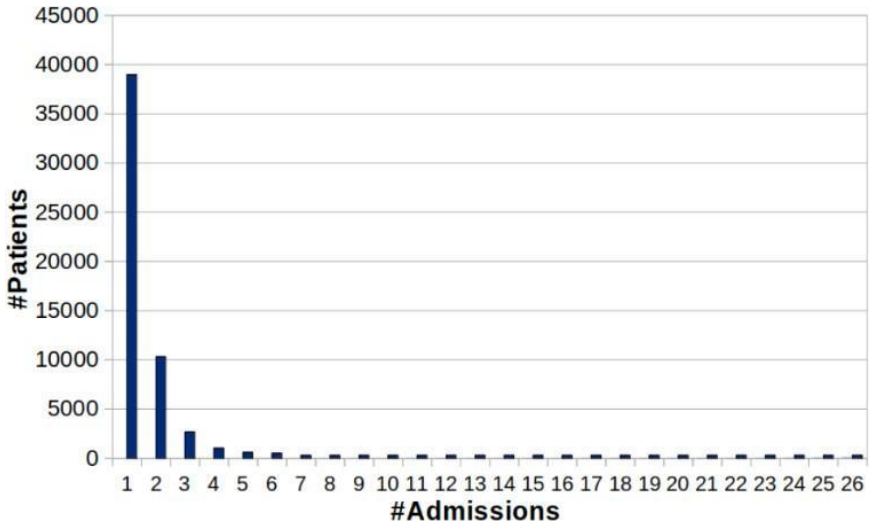

Fig.1. distribution vs. admission number [9]

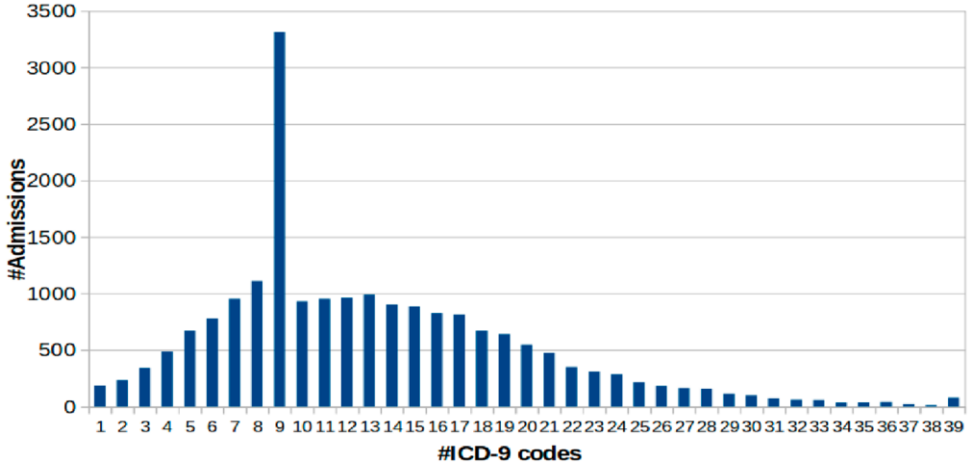

Fig. 2. Admission distribution vs ICD-9 codes

Consequently, ICD-9 faces a higher level of granularity which is a major challenge and it has a major impact on many researchers. An encoded method called Clinical Classification

Software (CCS) is employed to rectify this issue [9]. By using ICD-9, the specialized tabular format is defined in this classification policy to achieve the lowest possible granularity standard description. The main objective is to facilitate fast medical data analysis and an effective reporting system. As a result, after utilizing and deploying CCS, the subsequent admission score drops from 9.3 with 1049 feasibilities to 3.2 with 1031 feasibilities. Finally, this shows that the CCS encoding system utilization limits the 18 orders of metric feasibilities. Because of its simplicity, this aspect helps in prediction efficacy. Even, though granularity is reduced by this approach, there are some demerits of affecting the results of prediction and description by introducing less data. Also, it might affect the decision-making and classification performance. One more challenge is that 19,911 samples which is inadequate to deploy neural network training algorithm referred to 13 codes. As a result, this research shows that the MIMIC-III data set needs further interpretation to improve its applicability in medical analysis.

3. Multi-diseases

This paper mainly focuses on the four organ-related illnesses. The organs are the heart, lungs, liver, and kidney. See below for more details.

3.1 Cardiovascular disease (CVD)

CVD is a disease associated with blockage of blood vessels, distress in the chest leads to heart attack and other heart failure, and diseases may result in various acute diseases and mortality [10]. In the past 15 years, it has become number one in the top ten list for causing death. There were 15 million deaths in 2015 [11]. CVD plays a significant cause of mortality across the world, as per research conducted in January 2017. The World Health Organization in 2020 declared that 17.9 million people die every year due to this disease, which will become the primary cause of death globally.

Furthermore, the mortality rate is increasing every year due to coronary ailments. Experts expect the population to exceed 23.6 million by 2030. CVD is the prime cause of death globally, including cerebrovascular disease, coronary heart disease, strokes, peripheral arterial disease, transient ischemic attacks (TIA), vascular illness, and chronic heart illness.

3.1.1 Challenges of CVD detection

Even though this model looks so effective, it still faces some difficulties in 21 geographical regions; the WHO_CVD Risk Chart Working Group introduced a new chart that aims to predict the risk in hospitals and nationwide campaigns for public health [12]. CVD is a primary cause of death and other ailments globally, and it is crucial to prevent its growth by early detection and proper treatment.ML algorithm demonstrates that it can detect CVD. It is achieving more accurate and reliable outcomes by solving various issues. To address the current risk prediction model issues, the population-centered model needs further improvement, and it requires more money, time, and effort to gather higher observational data with diligent updates. The risk prediction method is at the pinnacle of the advancement.

Lack of standardized data: Training and validation processes in ML algorithms require a high quantity of data. However, CVD detection has a smaller amount of standardized data. Precise, stable, and fully developed data can impact the performance of the ML technique.

Imbalanced data: The training data may be imbalanced by the unequal proportion of CVD and non-CVD cases, and most of the requests are non-CVD: the ML algorithm's worst performance and inaccurate findings in CVD detection result from this.

The complexity of CVD: In CVD detection, detecting the most crucial factors becomes very challenging since it is a most dangerous disease associated with numerous health indicators and signs. This challenge regarding risk factors and complicating variables affects the outcome's precision in the ML technique. By considering ethical issues, the ML algorithm in health care affects some ethical values like privacy, bias, and adequately informed decisions. In the ML algorithm, ensure that the data is unbiased, explicit, and follows ethical standards.

Interpretability: Black boxes are frequently considered in the ML algorithm, making it problematic to analyze the outcome and comprehend how the algorithm proposed this determination. Figure 3 below displays the number of people with and without CVD.

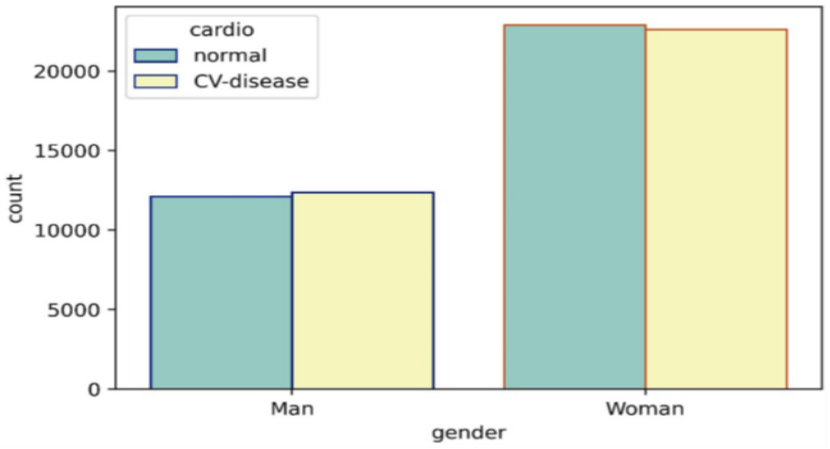

Fig. 3. The number of diseased and non-diseased people

3.1.2 Modifiable Risk Factors

One of the major concerns in the health sector is cardiovascular disease. Luckily, it is mainly associated with "behavioral risk factors." It indicates that a person's unhealthy lifestyle choices, such as not exercising, eating poorly, drinking alcohol, being obese, abusing tobacco, and having high blood pressure or hypertension, may contribute to CVD. Therefore, guiding a person's behavior in specific ways can reduce the risk of CVD. Low salt intake, a healthy, balanced diet, quitting alcohol and tobacco, and continuing regular exercise can all help lower the risk of CVD. As per experienced medical professional recommendations, consuming the prescribed medicines regularly for diabetes, high blood pressure, and high blood cholesterol can prevent heart attack and stroke.

3.1.3 Non-Modifiable Risk Factors

CVD is associated with many modifiable risk factors in lifestyle or behavior, but still, it has some other risk factors that are not modified. The patient's past lifestyle, age, and race are the three prominent non-modifiable risk factors. Researchers have discovered that a genetic predisposition mainly causes CVD. For instance, Hereditary tendencies like blood clotting inclinations, which can comprise genetic traits, can cause atherosclerosis, high blood lipids inflammation, and so on. Likewise, there is a direct correlation between CVD and genetic tendencies.

One of these genetic disorders is hypercholesterolemia. An insufficient cholesterol level is usually a modifiable ailment, but due to genetic disorders, it cannot be lessened or eradicated. Physicians should take into their family history and race. Some specific regions, such as African areas, the Caribbean, and South Asia, have a higher impact on this disease. Still, scientific studies have not evaluated an appropriate way to increase this risk. Eventually, healthcare professionals consider age as a non-modifiable risk factor.

Due to aging, physical conditions change naturally, leading to heart and blood vessel variations. It can cause the risk of CVD directly. Some physiological changes in our body during aging are the myocardium becoming rigid and blood vessels losing elasticity. This variance may lead to poor blood pumping and oxygen transportation to the organs and tissues. This non-modifiable characteristic enhances the risk of CVD.

.2 Kidney disease

The kidney is the primary internal organ in our body that controls blood pressure, balances the blood, and is essential for specific hormone production. Kidney Disease Improving Global Outcomes (KDIGO) describes chronic kidney disease (CKD) as a physical structure or physiological anomaly that lasts for more than 3 months [13]. Nephrolithiasis, kidney stones, hemolytic uremic syndrome, kidney cyst formation, a rare blood disorder, blood clots, breakdown of muscular tissue, glomerulonephritis, and other conditions are among the conditions that can cause kidney ailments [14],[15]. The symptoms of some CKD cases do not exist until the last stage of this disease, which makes it more challenging to evaluate the exact level of risk [16]. The medical statistics declared report in 2005 that 57 million cases were affected by CKD, of which 38 million people died. In COVID-19, the mortality rate of CKD patients with non-COVID-19 is 4.5%, whereas the CKD patients with COVID-19 is 44.5%. By 2050, the type 2 diabetes rate will be more than 150 million people, which will be the primary cause of various kidney ailments, according to World Health Organization (WHO) analysis [16]. The most common kidney disorders are cysts, hydronephrosis, and stone formation, which can be treated and prevented easily in the starting phase [17].

On the other hand, these conditions (i.e., "cardiorenal syndrome and uremia") may result in acute CKD and kidney malignancies. Although CKD is associated with significant negative consequences, CVD remains a major cause of mortality worldwide [18]. The healthcare provider screens and detects CKD in the patient early and prescribes medicines to consume, which can alter the disease advancement and lower the progress of the final stage of CKD and acute CVD disease [19].

Diagnosing this type of CKD is essential as soon as possible to protect the patient's life. Using various basic methods, medical professionals collect accurate details medical professionals collect to identify renal disease. These techniques include laboratory test results, such as blood and urine testing and physician observation. The blood test determines the "glomerular filtration rate" (GFR), which indicates renal function. The urine test shows whether the kidney works correctly by displaying the albumin level. Medical professionals must diagnose at the proper time with the help of a potent and standardized model created by utilizing the potential data sources. Machine Learning (ML) plays a significant role in developing an effective model in the medical diagnosis field, enabling rapid and accurate decision-making.

ML has a subfield called deep learning (DL), which utilizes a series of actions performed during training to look for essential linkages within the dataset. DL, a multilayer DL model, highly impacts medical devices and has the potential to handle non-linear data. For instance, James et al. (2010) developed a predictive model utilizing DL, which predicts CKD based on medical test results. A similar statement is employed in Ma et al. (2018). DL encounters significant obstacles in its many applications owing to the variation in medical information, which improves the stability and accuracy of the created replica and results in repeating diagnostic models and erroneous rules. As a result, the training procedure can produce a model with a high variance rather than one that guarantees reaching the maximum values. The team utilizes a diverse and varied DL model to address this issue. This process is known as ensemble learning. Merging the conventional and ensemble learning strengths overcomes the limitations of a single model and generates a more adaptable and standardized model [22]. There are two types of Ensemble learning base: learner and diversity [23]. The primary type is homogeneous learning, accomplished using various sample data. Next, achieved heterogeneous learning by utilizing other models.

The authors developed ensemble classifiers by using multiple combinations such as bagging [24], boosting [25], and stacking [26]. Our research states that the stacking ensemble model generates a stable, adaptable, standard model. Many researchers have declared ensemble learning an accurate and practical model.

Selecting an appropriate feature list constructs a practical model, which is very crucial. The feature selection is deeply evaluated in the ML algorithm to attain the precise outcome in medical applications. The feature selection comprises three categories: wrapper, filter, and embedded [27]. Our research selects the optimal feature list using four feature selection methods. Based on the data mentioned above, the primary goal of this study is to generate the ensemble DL, which uses the optimal feature subset and enhances the prediction performance. Compared with the existing techniques, our proposed feature list performs well in the early detection of CKD in the medical aspects.

CKD is very complicated and crucial to threaten the survival of the patient. It is so unfortunate that early-stage symptoms of CKD might be ambiguous, and some other signs may be confused with different diseases. For instance, some of the kidney disease symptoms include high blood pressure, lower albumin level, and rapid fall in white blood cells, which may cross over with hypertension, liver disease, anemia, and heart diseases. Hence, to

achieve data interpretation and accurate prediction results, the most efficient and accurate model is required to assist medical professionals in diagnosis.

3.2.1 Risk detection and prediction for chronic kidney disease

The researchers attempted to identify early or preventive measures after discovering kidney disease risk factors in human life. Detecting disease from the already affected patient is called disease detection. However, disease prediction states that it will be affected later. Consequently, the research is under two categories: detection and prediction. As per the first category, most researchers started CDK detection with the same dataset [28]. For early detection of CVD, the researchers [29] utilized SVM and ANN. Begin with the data preprocessing and replace the missing variables. Next, apply the tenfold cross-validation. The accuracy of ANN is up to 99.75%. Hence, our approach concluded that ANN performs better than SVM in accuracy. The drawback of this study is the limited sample size, which leads to dimensionality challenges. Utilizing the SVM approach addresses this issue. This study proposed the dl approach for CKD identification.

Within the identical year, the authors [30] introduced an intelligent classification model for CKD, which is Density-based Feature Selection (DFS) with ant colony-based Optimization (D-ACO). This method has to eradicate the replicated data to solve the issue regarding the growing feature count, which will eventually provide the solution for problems like expensive computation, weak interoperability, and overfitting. The author attained detection accuracy of 95% with only 14 of 24 features by using this method. Simultaneously, the originators [31] suggested a DNN model to identify earlier whether CKD is present or not. In this method, the author employed cross-validation to prevent the overfitting issue. It results in 97% accuracy, which performs better than the other models like Logistic, Naïve Bayes, Random Forest, SVM, and Adaboost. Later, the writers [32] developed an ensemble approach with random subspace and bagging to create an ideal model for CKD detection that achieved 100% accuracy on the previous dataset. Preprocessing the data is the first step, followed by managing missing values and standardizing the data. Three base learners, KNN, Naïve Bayes, and Decision Tree, were voted more often before choosing this algorithm.

Merging the base classifiers enhances classification performance in this research. As a result of the experiments, the efficiency matrices showed that the proposed model performs better than the other classifiers. The random subspace method surpasses the bagging algorithm in many scenarios.

3.3 Lung Disease

Lung disease encompasses a wide range of disorders affecting the lungs, which are crucial for respiration. These diseases are among the most prevalent medical conditions globally, with primary causes including smoking, infections, and genetic factors. One common lung disease is asthma, a chronic condition characterized by inflammation and narrowing of the airways. This results in wheezing, shortness of breath, chest tightness, and coughing. Another significant lung disease is Chronic Obstructive Pulmonary Disease (COPD), which includes conditions like emphysema and chronic bronchitis. COPD impedes airflow, complicates respiration, and causes symptoms including persistent dyspnea, chronic cough, mucus secretion, and wheezing.

Additionally, lung cancer, which originates in the lungs, is a significant health concern. Its symptoms often include a persistent cough, coughing blood, shortness of breath, and chest pain. The impact of these diseases underscores the importance of early detection, preventive measures, and effective treatment to improve the quality of life for those affected.

3.3.1 Risk detection and prediction for lung diseases

This research [33] uses artificial intelligence (AI) to predict ventilator-associated pneumonia (VAP) as a guide for the early identification of high-risk populations in clinical practice. Most data were from public databases, with machine learning being the main algorithm. The random forest (RF) model was the most commonly used. However, none of the studies predicted that AI models will be vital tools for VAP risk prediction in the future.

This research focuses [34]on ICUs, where routine data, both structured and unstructured, must be analyzed to forecast interventions for high-risk patients. While machine learning models have demonstrated superior performance, researchers have employed traditional statistical approaches. An ensemble model that predicts pneumonia patient outcomes aggregates many data sources based on the Medical Information Mart for critical care dataset. With an accuracy of 0.98 of the F1-score, the ensemble model fared better than the other two models based on caregiver narratives and structured data. Analyzing forecasts lets one pinpoint the primary influences on individual and collective results.

The authors [35] explored the interaction between in-hospital mortality and the red cell distribution width (RDW) to platelet ratio (RPR) in patients experiencing acute exacerbations of chronic obstructive pulmonary disease (AECOPD). The cohort comprised 1738 AECOPD patients from the eICU Collaborative Research Database and 1922 AECOPD patients who were at least 18 years old from MIMIC-III and MIMIC-IV. Multivariable logistic regression was employed in the study to assess the relationship between RPR and in-hospital death. The model increased the probability of in-hospital mortality in patients with second when accounting for confounders. Despite its good predictive performance, the prediction tactics allowed clinicians to identify patients at high risk of in-hospital mortality swiftly. The factors included in the prediction approach included age, ventilation, temperature, WBC, creatinine, hemoglobin, infectious illnesses, etc. Researchers observed no statistically significant fluctuation in the baseline features of record patients when they divided the data into training and testing sets at a 7:3 ratio (Table 1[35]). Figure 4 shows the visual representation of the model in the experimental set.

The authors found a negative correlation between serum albumin levels and deaths in the hospital in patients receiving critical care for chronic obstructive pulmonary disease (COPD) [36]. Using multivariate Cox regression analysis, the work used a retrospective observational cohort from the US Medical Information in Intensive Care database (MIMIC-IV). Serum albumin levels adversely link with in-hospital mortality, resulting in a 12.4% total in-hospital mortality rate.

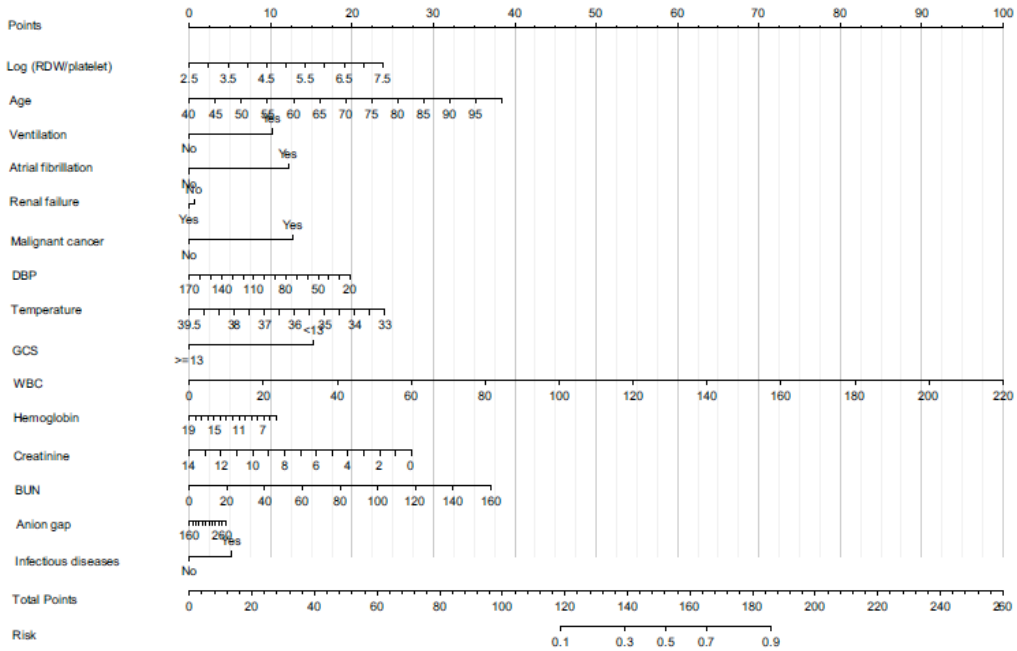

Fig. 4. The nomogram of the prognosis model.[35]

The study [37] focused on the relationship between critically ill patients' serum sodium levels and mortality from concomitant chronic obstructive pulmonary disease (COPD) and also utilized multivariable Cox regression analysis and Kaplan-Meier curves with data from the MIMIC-IV database. A constrained cubic spline explored the non-parametric relationship between serum sodium levels and death. The results showed a significant relationship between 1- and 3-year mortality in critically ill patients with concurrent COPD and hypo- and hypernatremia. The study showed that higher death rates in these patients were linked to both hyponatremia and hypernatremia, providing a new context for the therapeutic strategy of varying serum sodium levels.

The research conducted by the writers [38] explored the correlation between blood eosinophil concentrations and death in critically ill patients undergoing an acute exacerbation of chronic obstructive pulmonary disease (AECOPD). Applied extracted data from the MIMIC-III V1.4 database, and then the logistic regression model investigated the association between eosinophils and outcomes. Adjusted odds ratios of the preliminary blood eosinophil concentrations were linked to 0.792, 0.812, 0.847, and 0.914 for in-hospital mortality, in-ICU mortality, hospital length of stay, and ICU length of stay, respectively, following the creation of two multivariate regression models. Elevated blood eosinophil levels are associated with a lower in-hospital death rate and a shorter duration of stay in critically sick persons suffering an abrupt exacerbation of lung illness. Table 1 displays the dataset's medical data and demographic information.

Table 1: Diagnostic characteristics of 1019 AECOPD patients from the MIMIC-III Dataset

Variable	Survivors (n=887)	Survivors (n=887)
Age, years	71.2 (63.7,79.7)	79.0 (72.3,83.4)
Gender, male, N (%)	443 (49.9)	67 (50.8)
SAPS II	35 (29,43)	47 (39,59)
SOFA score	3 (2,5)	6 (4,8)
PH	7.34 (7.27,7.39)	7.32 (7.24,7.39)
PaO_2	87 (69,127)	87 (69,138)
$PaCO_2$	56 (47,72)	55 (41,73)
Length of hospital stay, days	7.1 (4.8,11.8)	7.5 (3.3,13.9)
Length of ICU stay, days	2.9 (1.5,5.7)	4.1 (1.8,9.3)
Emergency	873 (98.4)	131 (99.2)
Elective	4 (0.5)	0
Urgent	10 (1.1)	1 (0.8)

SOFA is ‘sequential organ failure assessment’ and SAPS is ‘simplified acute physiology score’.

The goal of the research [39] was to create and verify a machine-learning model for the early detection of moderate-to-severe cases of inhalation-induced acute respiratory distress syndrome (ARDS). The model utilized the RF method during a 90-hour timeframe that ended six hours before the beginning of moderate-to-severe respiratory failure, using data from the electronic ICU and the three most available vital signs. The researchers used two separate validation cohorts to verify the learned RF classifier and extracted rules for clinicians using a stable and interpretable rule set. The model identified several predictive indicators that might utilized to predict ARDS six hours before it starts in critical care units, including resp_96h_6h_min < 9 and resp_96h_6h_mean ≥ 16.1. With its predictive solid power for moderate-to-severe ARDS, this model may help physicians make better decisions and make it easier for patients to join preventative programs, leading to better outcomes.

Utilizing the Medical Information Mart for Intensive Care (MIMIC-III) and Telehealth Intensive Care Unit (eICU) Collaborative Research Database (eICU-CRD) databases, this study [40] created a machine learning-based mortality prediction approach for patients with acute respiratory distress syndrome (ARDS). The random forest technique was used to construct the model and evaluate it against other scoring schemes. Regarding forecasting in-hospital mortality, 30-day mortality, and 1-year mortality, the model outperformed 'SAPS-II, APPS, OSI, and OI.' Lactate level and platelet count were the most powerful predictors. Machine learning outperformed current grading systems by a large margin when predicting ARDS mortality. Table 2 displays the dataset's medical data and demographic information. The MIMIC III Dataset showed a 19.6% in-hospital death rate.

Table 2: Diagnostic characteristics of 2,235 ARDS patients from the MIMIC-III Dataset

Variable	Expired at hospital	Alive at hospital
Patients with ARDS, N (%)	437 (19.6)	1,798 (80.4)
Age, years	70.0 (22.9)	62.5 (25.4)
Gender, male, N (%)	242 (55.4)	996 (55.4)
BMI, kg/m2	27.3 (7.6)	27.9 (7.6)
PH	7.40 (0.10)	7.40 (0.10)
FiO2	59.0 (21.1)	59.0 (10.0)
PaO2	114.8 (45.2)	126.9 (44.3)
Length of hospital stay, days	18.2 (22.6)	21.6 (18.7)
Length of ICU stay, days	9.7 (11.4)	10.0 (10.8)
Emergency	363 (83.1)	1,354 (75.3)

Elective	54 (12.4)	398 (22.1)
Urgent	20 (4.6)	46 (2.6)
Mean arterial pressure, mmHg	75.3 (11.7)	78.4 (10.8)
Heart rate, bpm	94.1 (17.4)	89.8 (15.9)

BMI is body mass index, BPM is beats per minute, and the categorical variables are presented as N (%).

3.4 Liver disease

The liver is one of the major internal organs in our human body. Finding any malfunction in the liver can be life-threatening. The sole therapy for this is transplantation, which replaces with another liver within the deadline. Various factors that help detect the ailment early on are albumin, age, gender, total bilirubin, SGPT, ALP, etc. Researchers use these features in the studies mentioned above to identify liver disease. Even if this research uses various ML techniques based on classification, it faces more challenges. Rather than an ensemble model, the majority of the recent study utilized a simple ML model. Several data preprocessing methods are available to optimize the outcome. Researchers have yet to do any research on this technique.

Moreover, some types of research successfully employ effective feature selection and transformation methods. Several ensemble algorithms used in this research are boosting, bagging, stacking, and so on to address this issue and achieve a more significant outcome. Also, the model's performance is optimized using advanced data p reprocessing techniques with proper feature scaling and selection procedures.

3.4.1 Dataset for Classifying liver disorders

Researchers classified the liver disease using the Indian Liver Patient Dataset (ILPD) from the UCI Machine Learning Repository [41]. This dataset assigns the target variable and ten features in the 11 columns. Gender, age, 'total proteins (TP), albumin (ALB), albumin and globulin ratio (A/G), alanine aminotransferase (SGPT),' alkaline phosphatase (Alkphos), and aspartate aminotransferase (SGOT) are among these characteristics. Table [8] formulates the attributes of the affected person's full features. Researchers divide the results into two categories: patients with liver sickness and those without it. The results classify patients as either having a liver disease or a non-liver illness. Figure 5 displays the data of 583 patients, both with and without liver disease, from the northeastern Indian state of Andhra Pradesh.

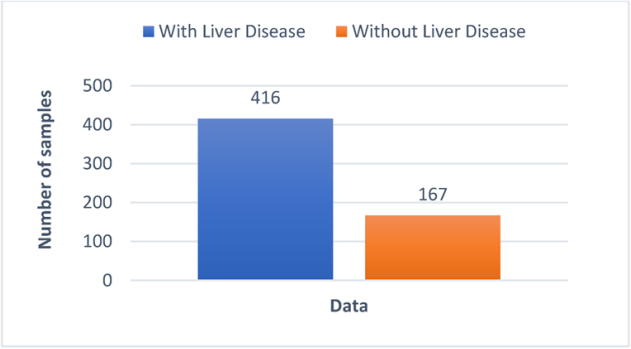

Fig. 5: The split between liver disease patients and non-patients.

The researchers [42] suggested a rule-based approach for data classification of various liver illnesses using machine learning algorithms. 'Rule induction (RI), SVM, decision trees (DT), naive Bayes, and ANN' using the 'k-fold cross-validation' methodology are some of the machine learning techniques used in this model. Across all models, the DT linked to a 'rule-based classification' algorithm achieves greater exactness. It can produce a dataset with 583 records and 12 characteristics. The originators utilize A decision tree method [43] for detecting liver fibrosis levels. The decision tree model proves that it should end with better classification accuracy. Researchers utilize classification techniques like c4.5, SVM, naive Bayes, logistic regression, and neural networks to examine liver disease disorders [44]. The AP data set with the C4.5 model outperformed the UCLA dataset with other models. The writers [45] predicted progressive fibrosis in patients with chronic hepatitis C using mathematical models and medical diagnostics. During the training phase, the data was categorized into two sets: (1) minor to moderate fibrosis (f0–f2) and (2) severe fibrosis (f3–f4), depending on the metavir score. Creating decision trees, genetic algorithms, multilinear regression models, and particle swarm optimization achieves an advanced fibrosis prediction. Researchers found that age, platelet count, albumin, and AST are related to advanced fibrosis. The authors [46] utilized backpropagation models and SVM for liver disease classification.

The model uses the UCI repository dataset as a training dataset. A backpropagation model performed better than SVM. Han Ma et al. determine the best prediction model for Non-Alcoholic Fatty Liver Disease (NAFLD) detection [47]. The people's health records collected from the medical checkups held at Zhejiang University's first affiliated hospital generated the model. The Bayesian network model outperformed 11 other models. It is beneficial for the medical community to diagnose liver disease patients using the live graphical user interface [48].

[49] used the dataset of Indian Liver Patients, 583 patient records to train the model on ten features. A comparison analysis of many classification algorithms, including 'SVM, KNN, ANN,' and 'logistic regression,' was done to determine the best prediction model. This comparison led to ANN obtaining better accuracy. Using the UCI repository dataset and 15 life quality parameters, the author [49] suggested a C4.5 decision tree model. This study made a performance comparison between the C4.5 model and the k-means clustering algorithm. The result stated that C4.5 yielded more accurate data.

4. Conclusion

The survey determined that the MIMIC-III and MIMIC-IV datasets are essential for advancing critical care in predicting, detecting, and managing cardiovascular, renal, and hepatic diseases. The researchers can identify modifiable and non-modifiable risk factors for cardiovascular disease and develop predictive models to enhance patient treatment by analyzing the hospital record system. Employing this information facilitates predicting and identifying chronic kidney disease while mitigating disease development and improving patient health. Moreover, it enables the prognosis of liver illness and the formulation of therapy strategies by classifying liver disease based on comprehensive patient data. These databases enhance the prediction and detection of pneumonia and COPD. It primarily

concentrates on the patients' death rate. Addressing challenges related to data integrity, model uniformity, and security threats is necessary to utilize the critical care dataset properly. Future studies should address these issues and comprehensively analyze the information to improve the development of the healthcare system and clinical outcomes.

Declaration of Competing Interest

The authors declare that none of the work reported in this study could have been influenced by any known competing financial interests or personal relationships.

References

1. A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals," 2000. [Online]. Available: http://www.physionet.org
2. A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci Data, vol. 10, no. 1, Dec. 2023, doi: 10.1038/s41597-022-01899-x.
3. A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci Data, vol. 10, no. 1, Dec. 2023, doi: 10.1038/s41597-022-01899-x.
4. W. Boag and P. Szolovits, "EHR Safari: Data is Contextual," 2022. [Online]. Available: https://blog.neurips.cc/2021/08/23/neurips-2021-ethics-guidelines
5. M. B. A. McDermott, S. Wang, N. Marinsek, R. Ranganath, M. Ghassemi, and L. Foschini, "Reproducibility in Machine Learning for Health," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.01463
6. M. Gupta, R. Poulain, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti, "Flexible-Window Predictions on Electronic Health Records," 2022. [Online]. Available: www.aaai.org
7. Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo, "Natural Language Processing for EHR-Based Computational Phenotyping," IEEE/ACM Trans Comput Biol Bioinform, vol. 16, no. 1, pp. 139–153, Jan. 2019, doi: 10.1109/TCBB.2018.2849968.
8. A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, "Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease," Biomedicines, vol. 11, no. 2, Feb. 2023, doi: 10.3390/biomedicines11020581.
9. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
10. V. Kedia, S. R. Regmi, K. Jha, A. Bhatia, S. Dugar, and B. K. Shah, "Time Efficient IOS Application for CardioVascular Disease Prediction Using Machine Learning," in Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, Institute of Electrical and Electronics Engineers Inc., Apr. 2021, pp. 869–874. doi: 10.1109/ICCMC51019.2021.9418453.
11. S. Omar and N. Mohamed, "A Cardiovascular Disease Prediction Using Machine Learning Algorithms." [Online]. Available: https://www.semanticscholar.org/paper/Effective-Heart-
12. F. Farzadfar, "Cardiovascular disease risk prediction models: challenges and perspectives," 2019, doi: 10.1016/S2214¬109X(19)30365¬1.
13. "Official JOurnal Of the internatiOnal SOciety Of nephrOlOgy KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease." [Online]. Available: www.publicationethics.org
14. C. P. Kovesdy, "Epidemiology of chronic kidney disease: an update 2022," Apr. 01, 2022, Elsevier B.V. doi: 10.1016/j.kisu.2021.11.003.
15. Y. Zhou and J. Yang, "Chronic Kidney Disease: Overview," in Chronic Kidney Disease: Diagnosis and Treatment, 2019. doi: 10.1007/978-981-32-9131-7_1.

16. V. Jha et al., “Chronic kidney disease: Global dimension and perspectives,” 2013, Elsevier B.V. doi: 10.1016/S0140-6736(13)60687-X.
17. Y. Wu and Z. Yi, “Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks,” Knowl Based Syst, vol. 200, Jul. 2020, doi: 10.1016/j.knosys.2020.105873.
18. K. Swathi and G. Vamsi Krishna, “Prediction of Chronic Kidney Disease with Various Machine Learning Techniques: A Comparative Study,” in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-981-19-6880-8_27.
19. K. Matsushita, S. H. Ballew, A. Y. M. Wang, R. Kalyesubula, E. Schaeffner, and R. Agarwal, “Epidemiology and risk of cardiovascular disease in populations with chronic kidney disease,” 2022. doi: 10.1038/s41581-022-00616-6.
20. M. T. James, B. R. Hemmelgarn, and M. Tonelli, “Early recognition and prevention of chronic kidney disease,” 2010, Elsevier B.V. doi: 10.1016/S0140-6736(09)62004-3.
21. F. Ma, Q. You, J. Gao, J. Zhou, Q. Suo, and A. Zhang, “Risk prediction on electronic health records with prior medical knowledge,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Jul. 2018, pp. 1910–1919. doi: 10.1145/3219819.3220020.
22. B. Navaneeth and M. Suchetha, “A dynamic pooling based convolutional neural network approach to detect chronic kidney disease,” Biomed Signal Process Control, vol. 62, Sep. 2020, doi: 10.1016/j.bspc.2020.102068.
23. N. El-Rashidy et al., “Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning,” Neural Comput Appl, vol. 34, no. 5, pp. 3603–3632, Mar. 2022, doi: 10.1007/s00521-021-06631-1.
24. P. Jayanthi, “Machine learning and deep learning algorithms in disease prediction: Future trends for the healthcare system,” in Deep Learning for Medical Applications with Unique Data, 2022. doi: 10.1016/B978-0-12-824145-5.00009-5.
25. Q. Sun and B. Pfahringer, “LNAI 7691 - Bagging Ensemble Selection for Regression,” 2012.
26. R. O. Odegua, “An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking) Rising Odegua Nossa Data An Empirical Study of Ensemble Techniques (Bagging, Boosting and Stacking).” [Online]. Available: https://www.researchgate.net/publication/338681864
27. A. Sharafati, S. B. Haji Seyed Asadollah, and N. Al-Ansari, “Application of bagging ensemble model for predicting compressive strength of hollow concrete masonry prism,” Ain Shams Engineering Journal, vol. 12, no. 4, pp. 3521–3530, Dec. 2021, doi: 10.1016/j.asej.2021.03.028.
28. C. Kaur, M. S. Kumar, A. Anjum, M. B. Binda, M. R. Mallu, and M. S. Al Ansari, “Chronic Kidney Disease Prediction Using Machine Learning,” Journal of Advances in Information Technology, vol. 14, no. 2, 2023, doi: 10.12720/jait.14.2.384-391.
29. N. A. Almansour et al., “Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study,” Comput Biol Med, vol. 109, pp. 101–111, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.017.
30. M. Elhoseny, K. Shankar, and J. Uthayakumar, “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease,” Sci Rep, vol. 9, no. 1, Dec. 2019, doi: 10.1038/s41598-019-46074-2.
31. H. Kriplani, B. Patel, and S. Roy, “Prediction of chronic kidney diseases using deep artificial neural network technique,” in Lecture Notes in Computational Vision and Biomechanics, vol. 31, Springer Netherlands, 2019, pp. 179–187. doi: 10.1007/978-3-030-04061-1_18.
32. O. A. Jongbo, A. O. Adetunmbi, R. B. Ogunrinde, and B. Badeji-Ajisafe, “Development of an ensemble approach to chronic kidney disease diagnosis,” Sci Afr, vol. 8, Jul. 2020, doi: 10.1016/j.sciaf.2020.e00456.

33. J. Zhang, P. Yang, L. Zeng, S. Li, and J. Zhou, "Ventilator-Associated Pneumonia Prediction Models Based on AI: Scoping Review," 2024, JMIR Publications Inc. doi: 10.2196/57026.
34. C. Mugisha and I. Paik, "Pneumonia Outcome Prediction Using Structured and Unstructured Data from EHR," in Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 2640–2646. doi: 10.1109/BIBM49941.2020.9312987.
35. S. Chen, Y. Shi, B. Hu, and J. Huang, "A Prediction Model for In-Hospital Mortality of Acute Exacerbations of Chronic Obstructive Pulmonary Disease Patients Based on Red Cell Distribution Width-to-Platelet Ratio," International Journal of COPD, vol. 18, pp. 2079–2091, 2023, doi: 10.2147/COPD.S418162.
36. M. Ling et al., "Relationship between human serum albumin and in-hospital mortality in critical care patients with chronic obstructive pulmonary disease," Front Med (Lausanne), vol. 10, 2023, doi: 10.3389/fmed.2023.1109910.
37. L. Fan et al., "Association Between Serum Sodium and Long-Term Mortality in Critically Ill Patients with Comorbid Chronic Obstructive Pulmonary Disease: Analysis from the MIMIC-IV Database," International Journal of COPD, vol. 17, pp. 1143–1155, 2022, doi: 10.2147/COPD.S353741.
38. J. Yang and J. Yang, "Association between blood eosinophils and mortality in critically ill patients with acute exacerbation of chronic obstructive pulmonary disease: A retrospective cohort study," International Journal of COPD, vol. 16, pp. 281–288, 2021, doi: 10.2147/COPD.S289920.
39. J. Wu et al., "Early prediction of moderate-to-severe condition of inhalation-induced acute respiratory distress syndrome via interpretable machine learning," BMC Pulm Med, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12890-022-01963-7.
40. B. Huang et al., "Mortality prediction for patients with acute respiratory distress syndrome based on machine learning: a population-based study," Ann Transl Med, vol. 9, no. 9, pp. 794–794, May 2021, doi: 10.21037/atm-20-6624.
41. M. Mayer, "Fast Imputation of Missing Values," 2018.
42. Y. Kumar and G. Sahoo, "Prediction of different types of liver diseases using rule based classification model," Technology and Health Care, vol. 21, no. 5, pp. 417–432, 2013, doi: 10.3233/THC-130742.
43. M. Essaaidi, M. Nemiche, Institute of Electrical and Electronics Engineers. Morocco Section, and Institute of Electrical and Electronics Engineers, Proceedings of 2015 IEEE World Conference on Complex Systems.
44. S. Hashem et al., "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," IEEE/ACM Trans Comput Biol Bioinform, vol. 15, no. 3, pp. 861–868, May 2018, doi: 10.1109/TCBB.2017.2690848.
45. IEEE Staff, 2017 International Conference on Emerging Trends and Innovation in ICT (ICEI). IEEE, 2017.
46. H. Ma, C. F. Xu, Z. Shen, C. H. Yu, and Y. M. Li, "Application of Machine Learning Techniques for Clinical Predictive Modeling: A Cross-Sectional Study on Nonalcoholic Fatty Liver Disease in China," Biomed Res Int, vol. 2018, 2018, doi: 10.1155/2018/4304376.
47. J. Jacob, J. Chakkalakal Mathew, J. Mathew, and E. Issac, "Diagnosis of Liver Disease Using Machine Learning Techniques," International Research Journal of Engineering and Technology, 2018, [Online]. Available: www.irjet.net
48. [A. L. G Sri Jagadguru, "Chronic Liver Disease Prediction Analysis Based on the Impact of Life Quality Attributes," 2019. [Online]. Available: https://www.researchgate.net/publication/370340538
49. V. Durai, S. Ramesh, and D. Kalthireddy, "Liver disease prediction using machine learning," 2019. [Online]. Available: www.IJARIIT.com