79914_Auto_Edited-check.docx

Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 79914

Manuscript Type: REVIEW

Machine learning insights concerning inflammatory and liver-related risk comorbidities in non-communicable and viral diseases

Martinez JA *et al*. ML in CLD

Abstract

The liver is a key organ involved in a wide range of functions, whose damage can lead to chronic liver disease (CLD). CLD accounts for more than two million deaths worldwide, becoming a social and economic burden for most countries. Among the different factors that can cause CLD, alcohol abuse, viruses, drug treatments, and unhealthy dietary patterns top the list. These conditions prompt to perpetuate an inflammatory environment and oxidative stress imbalance that favor the development of hepatic fibrogenesis. High stages of fibrosis can eventually lead to cirrhosis or hepatocellular carcinoma (HCC). Despite the advances achieved in this field, new approaches are needed for the prevention, diagnosis, treatment and prognosis of CLD. In this context, the scientific community is using machine learning (ML) algorithms to integrate and process vast amounts of data with unprecedented performance. ML techniques allow the integration of anthropometric, genetic, clinical, biochemical, dietary, lifestyle and omics data, giving new insights to tackle CLD and bringing personalized medicine a step closer. This review summarises the investigations where ML techniques have been applied to study new approaches that could be used in inflammatory-related, hepatitis viruses-induced, and coronavirus disease 2019-induced liver damage and enlighten the factors involved in CLD development.

Key Words: Machine learning; Liver inflammation; Liver disease; Viral diseases

Martinez JA, Alonso-Bernáldez M, Martínez-Urbistondo D, Vargas-Nuñez JA, Ramírez de Molina A, Dávalos A, Ramos-Lopez O. Machine learning insights concerning inflammatory and liver-related risk comorbidities in non-communicable and viral diseases. *World J Gastroenterol* 2022; In press

Core Tip: Chronic liver disease has become a global burden and new approaches need to be explored to tackle this disease. In this context, machine learning (ML) techniques bring a whole new set of opportunities to study novel approaches and biomarkers for prevention, diagnosis, treatment, and prognosis of inflammatory and virus-related liver diseases. The application of ML algorithms constitutes a pivotal piece of personalized medicine, allowing the integration of different phenotypical and genotypical data for a precision outcome concerning inflammatory liver comorbidities in non-communicable and viral diseases.

INTRODUCTION

Liver is a key organ involved in relevant homeostatic metabolic and detoxifying human functions[1]. Thus, the liver is the epicentre of an organ-organ network weaving a series of complex interactions in the organism, which makes liver damage an underlying adverse condition in a whole set of diseases. Chronic liver disease (CLD) can be caused mainly by alcoholic liver-related dysfunctions, hepatitis B virus (HBV), hepatitis C virus (HCV), drug treatments, or non-alcoholic fatty liver disease (NAFLD), as recently updated to the term metabolic-associated FLD or NAFLD (Figure 1)[2,3]. Patients with liver-related diseases need frequent follow-ups and careful monitoring, since CLD can eventually lead to cirrhosis or hepatocellular carcinoma (HCC) if not diagnosed on time for treatment or surgery. These CLD-related conditions have become a global burden, whose mortality associated rates have increased over the years reaching more than 2 million deaths worldwide[4].

CLD is usually accompanied by an unhealthy inflammatory environment[5]. The immune response is a fundamental process to maintain homeostasis within the organism defence machinery and is characterized by the secretion of pro-inflammatory cytokines, like interleukin (IL)-1, tumor necrosis factor-α (TNF-α), and prostaglandin E2, in an acute manner in order to resolve a sudden damage[5]. However, if sustained over time, these abnormal levels of inflammatory cytokines cause low-grade inflammation (LGI). LGI is a silent condition that predisposes to the development of metabolic and infectious diseases that has become a worldwide health issue[6]. Patients with CLD, such as non-alcoholic steatohepatitis (NASH), present impaired immune function, dysbiosis, insulin resistance (IR), and LGI, all of which can aggravate infectious diseases' progression and perpetuate excess of adipose tissue, over stimulating the production of adipose-derived inflammatory molecules [5,7-9].

The liver also secretes important hepatokines that act as signaling proteins modulating functions in other organs and being involved in a wide range of conditions, such as IR and adipogenesis[1]. For instance, fibroblast growth factor-21 (FGF-21) is a mediator participating in glucose metabolism mainly secreted by the liver that modulates adipogenesis, while fetuins, liver-derived plasma proteins, are participating in metabolic impairment and inflammation[1]. A dysregulation in systemic cytokines prompts fat accumulation in hepatocytes, which in turn promotes local secretion of pro-inflammatory hepatokines, leading to liver steatosis and IR. In addition, immune cells also find difficulty in this inflammatory environment to exert its role appropriately. Persistent inflammatory signals over time also abnormally activate immune cells, impairing the body's ability to fight infection, repair tissue damage, or recover from possible poisoning. Inflammation comes hand in hand with increase oxidative stress, a state characterized by an imbalance in favoring the accumulation of higher reactive oxygen (ROS) and nitrogen species. These molecules in unusual concentrations damage the cell and environmental milieu by promoting the expression of pro-inflammatory genes, resulting in a vicious cycle. Thus, CLD presents an oxidative atmosphere, probably linked to the pro-inflammatory state[10,11]. This environment is the perfect setting for the fibrogenic process to unfold, an underlying condition of CLD that is characterized by progressive accumulation of fibrillar extracellular matrix in the liver[12]. The stage of hepatic fibrosis has been associated with the risk of mortality and liver-related morbidity in patients with NAFLD[13], virus-induced hepatitis[14,15] and alcoholic-derived liver disease[16], eventually leading to HCC.

In this context, infection by human hepatitis viruses (HHVs) are the most common cause of hepatitis, leading to the activation of the immune system and the subsequent inflammatory response[17]. HBV and HCV acute infections can be now often resolved with antiviral and immune therapy, however, in a significant percentage they can progress to chronic hepatitis. This persistent infection can lead to comorbidities outside the liver, like arthritis, vasculitis, myalgia, and peripheral neuropathies[18]. Moreover, besides HHVs, another new infectious disease appeared in late 2019 that can cause liver damage: Coronavirus disease 2019 (COVID-19). COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, and it has become a global health issue since its outbreak in 2020 was declared a pandemic. Beyond lung function, COVID-19 can affect a wide variety of tissues, like gastrointestinal track, kidneys, and liver, with an underlying adverse inflammatory environment[19]. This inflammatory-related condition has been strongly associated to metabolic status, worsening diseases like obesity, diabetes, and hypertension[7,20-22]. For instance, COVID-19 can increase hepatic lipid accumulation by mitochondrial and endoplasmic reticulum (ER) dysfunction, or worsen NAFLD if it was already present. A recent systematic review depicted that the parameters normally used for liver impairment screening were significantly increased in COVID-19 patients[23], placing CLD as a risk factor for progressive and severe COVID-19[24,25].

CLD is a global health problem and new methods are needed to tackle this life-threatening condition. In this line, this review aims to explore machine learning (ML)-based approaches to manage CLD and develop biomarkers for diagnosis and prognosis. Its goal is to shed light on the factors involved in CLD to help health professionals in clinical management with the support of ML and identify new targets that can define therapeutic care lines in viral infections and non-communicable diseases (NCD), with an impact on liver functions with an inflammatory component. This includes the new disease COVID-19.

MECHANISMS BY WHICH NCD AND INFLAMMATORY/IR PHENOMENA CAN AFFECT LIVER FUNCTION

The incidence of NCD, such as cardiovascular diseases and diabetes, has skyrocketed in the last decades, pressing authorities to establish developmental goals to achieve in the near future in terms of decreasing NCD-caused mortality[26]. Some of the risk factors that contribute to the development of NCD are excess of adipose tissue and high levels of glycaemia. In this context, adipose tissue plays a key role in the development of FLD by secreting adipokines and other molecules, like free fatty acids (FFA)[8].

An energy excess prompts fat accumulation in the organism and the subsequent dysregulation of this tissue. This is of relevance since an inflamed adipose tissue results in increased levels of FFA and pro-inflammatory cytokines, IR, and infiltration of macrophages in the liver by the activation of Th1 and Th17 cells[8]. FFA enter the liver through the portal vein and trigger a series of reactions. For instance, they serve as ligands to toll-like receptor-4 complex, stimulating the production of TNF-α through the activation of nuclear factor-kappa B, favoring an inflammatory environment. Moreover, the excess of fat drives the polarization state of this increased number of macrophages from anti-inflammatory M2 to pro-inflammatory M1 macrophages and prompts fat accumulation in the liver and IR[8]. This adipose-derived macrophages also secrete inflammatory molecules, like TNF-α and IL-6, and adipokines, such as visfatin [also named nicotinamide phosphoribosyl transferase (NAMPT)]. NAMPT has gained relevance as a pivotal molecule linking adipose tissue and FLD. NAMPT is a pleiotropic molecule that can be found in an extracellular (eNAMPT) or an intracellular (iNAMP) form. Studies indicate that eNAMPT has enzyme and cytokine-like activity, stimulating the release of pro-inflammatory cytokines. Meanwhile, iNAMPT catalyses the rate-limiting step in nicotinamide adenine dinucleotide (NAD+) formation. Because of this NAD+ boosting property, levels of iNAMPT have been proposed as beneficial for the homeostasis of the cell due to influencing the activity of NAD-dependent enzymes, such as sirtuins (SIRT). Remarkably, SIRT1 plays a key role in the liver by modulating the acetylation status of target molecules in lipid metabolism[27].

Furthermore, IR is characterized by hyperglycaemia and the subsequent hyperinsulinemia to counteract high glucose levels, being a risk factor for NCDs, particularly type 2 diabetes, where it has been closely linked to oxidative stress[28]. A normal insulin signaling pathway starts with the activation of insulin receptor so that it can bind to phosphoinositide 3-kinase to ultimately activate protein kinase B (Akt). Activated Akt drives glucose entry into the cell by promoting GLUT4 expression and glycogen synthesis[29]. Oxidative stress impairs this signal transduction through a lot of different mechanisms, like inhibiting the transcription factors insulin promoter factor 1 and peroxisome proliferator-activated receptor gamma, which mediate insulin and GLUT-4 expressions, respectively. Moreover, under hyperglycaemic conditions, fetuin A hepatokine inhibits the insulin receptor and promotes inflammation, while FGF-21 inhibits lipid accumulation and increases insulin sensitivity. Dysregulation of this hormones, together with oxidative stress imbalance, lead to impaired insulin signaling[30].

The metabolic conditions underlying the development of NCD are complex and they often reinforce each other, perpetuating an inflammatory environment and oxidative stress imbalance. As the orchestrating organ, these processes converge in the liver, affecting metabolic functions and setting the bases for the onset of fibrogenic process characteristic of CLD.

MECHANISMS BY WHICH VIRAL INFECTIONS AND INFLAMMATORY/IR PHENOMENA CAN AFFECT LIVER FUNCTION

Persistent virus-associated liver damage can progress to CLD, which pressures health systems with a big social and economic burden. Although lots of resources have been invested to study the molecular mechanisms that mediate this process, results are diverse and still being under investigation by the scientific community. HHVs directly infect hepatocytes and the internalization into the cell is believed to happen by endocytosis, requiring the interaction with several host cell factors[17]. However, viral entry of HBV and HCV within hepatocytes is unclear and further research is needed to elucidate this question. It has recently been identified sodium taurocholate co-transporting polypeptide as an HBV receptor that would mediate HBV cell entry[31]. In the case of HCV, specific intercellular adhesion molecules appear key to cell adhesion and subsequent internalization[32].

Regarding HBV and HCV replication, it has been found that liver X receptor-α (LXR-α) plays a key role. LXR-α is a transcription factor whose activation triggers the expression of different genes that directly or indirectly modulate these viruses' replication, as well as the lipid and inflammatory alterations associated to CLD[33]. This inflammation is also mediated by the nucleotide-binding oligomerization domain-like receptor protein 3, which is activated by the abnormal production of ROS after a viral infection occurs in the liver. This ROS increase is associated to a decreased expression of nuclear factor-e2-related factor-2, a transcription factor that regulates ROS/recepteur d'origine nantais balance by maintaining redox homeostasis. These alterations compromise the normal state of the cell, laying the foundations on which the fibrotic process of CLD begins[11].

In the case of COVID-19, the mechanisms by which liver damage can occur are more unclear, but it is widely accepted that inflammation plays a huge role. This infection can trigger an exaggerated immune response leading to an uncontrolled cytokine release, also known as "cytokine storm". It is characterized by abnormal levels of IL-6, IL-1, C-C motif chemokine ligand (CCL)-5, chemokine (C-X-C motif) ligand (CXCL)-8, CXCL-1, and TNF-α among others[19]. This inflammatory cascade affects bile duct function, since cytokines like TNF-α, IL-1 and IL-6, can induce hepatocellular cholestasis by down-regulating hepatobiliary uptake and excretory systems[34].

Furthermore, the presence of this inflammatory environment can upregulate the expression of angiotensin converting enzyme 2 (ACE2) receptor in different tissues, like the adipose tissue and the liver[35-39]. This is of relevance since ACE2 receptors are the main cell entrance of the SARS-CoV-2 virus and they are present in different tissues. Particularly in the liver, the cholangiocytes (characteristic cells of bile duct)[40], as well as liver vascular endothelial cells[41], express ACE2 receptors. Hepatocytes and cholangiocytes are permissive to SARS-CoV-2 virus, mediating subsequent entrance into the liver[42]. Several studies have found that *ACE2* expression in hepatocytes is increased under hypoxia[43], a frequent condition in COVID patients, and fibrotic conditions[44]. Besides ACE2 receptors, transmembrane serine protease 2 (TMPRSS2) and paired basic amino acid cleaving enzyme (FURIN) have been noted as significant for infection in the liver[45,46]. In this context, *ACE2* expression is increased in patients in

HCV-related cirrhosis[44], whereas *TMPRSS2* and *FURIN* expressions are upregulated in patients with obesity and NAFLD[47]. Moreover, infection by SARS-CoV-2 increases glucose-regulated protein 78 and 94, two biomarkers of ER stress[48,49], and impairs mitochondrial function[50]. This process is of interest since this state has been associated to *de novo* lipogenesis in hepatocytes[51], which could eventually lead to steatosis in these patients.

The use of therapeutic drugs can be another underlying cause of liver damage[3]. Because of detoxifying functions, the liver is subject to drug-induced damage coming from a wide range of approved drugs. Oncology drugs account for most of hepatotoxicity reported cases, followed by those used for infectious diseases[3]. Since the beginning of the COVID-19 pandemic, a wide range of different treatments (antivirals, antibiotics, antimalaria, or corticosteroids) have been used in the absence of an efficient drug to treat severe infections. This pharmacological administration could explain that drug-induced liver injury appears in nearly 25% of COVID-19 patients[23], a consequence to consider when addressing liver damage in this disease.

ML APPROACHES IN INFLAMMATORY AND LIVER-RELATED COMORBIDITIES IN NON-COMMUNICABLE AND VIRAL DISEASES

Despite all the advances in the mechanisms driving the onset of these diseases, new techniques to detect innovative biomarkers for diagnosis, and prognosis, as well as to discover novel drugs are needed, like for example artificial intelligence (AI). AI seeks to mimic human behavior, and within this science, ML is the most common approach[52]. The advances in computational science in the last decades have permitted the development of powerful algorithms based on this science. ML algorithms are particularly relevant for biological research, because they allow the processing and integration of the huge amount of data that the latest advances in this field have brought by applying statistical methods to enable machines to improve with experiences. This methodological approach can be categorized into two big groups: Supervised and unsupervised learning. In supervised algorithms, data is tagged in order to train the algorithm and fit it appropriately, whereas if it is unsupervised, the algorithm learns patterns from unlabeled data[53]. ML algorithms are generally assessed by simple methodologies like sensitivity, specificity, and accuracy. While sensitivity evaluates the proportion of true positives correctly identified, specificity evaluates the proportion of true negatives. Meanwhile, the accuracy value indicates the number of times the model is correct[54].

Supervised algorithms can be divided into two categories depending on the purpose: Prediction, in which the algorithm is fed and trained predictive models to data, or classification, that consists in clustering data within explanatory groups[55,56]. Predictive algorithms are based on regression models and the most used are linear and logistic regression (LR), support vector machine (SVM), support vector regression (SVR), extra tree regression (ETR), artificial neural networks (ANN), and decision trees (DT). Regression models analyse the influence of one or multiple variables on a nominal or ordinal categorical outcome. ANN are more complex mathematical models (deep learning algorithms), that mimic the brain neural network, like the convolutional neural network (CNN), in which an input is fed through a hidden layer of lots of different well connected and structured nodes to produce a final output. In deep neuronal network (DNN) models, a great number of hidden successive layers use the output from the previous layer as input in a more complex algorithm. DT can also classify data, like random forest (RF) or gradient boosting (GB) models. Instead of minimizing error, these models determine thresholds derived from input data, assigning weight values to variables. Other models of classification are the Ada-Boost, Bayesian network (BN), Naïve Bayes (NB), K-Nearest Neighbours (KNN), and Linear Discriminant Analysis (LDA) that group data into clusters[55,56]. All these models can shed light into biological questions and are normally used indistinctively to obtain the best performance with the same dataset. For instance, Mijwil and Aggarwal[57] analysed and compared 7 ML algorithms to predict appendix illness in the same dataset, revealing that certain models performed better than others, allowing for higher accuracy and results.

In FLD, the common techniques used in diagnostics are based on techniques like ultrasonography and magnetic resonance imagining (MRI). These methods are subjective, and the informed outcome mainly relies on the interpretation of the professional carrying out the procedure. Several investigations have studied the implementation of ML in order to classify FLD and other liver diseases by using images from ultrasounds, computed tomography (CT), and MRI[58,59]. However, the downside of this approach is that the quality of the images differs from one another because of several factors, such as equipment precision and interpersonal differences, for instance. Therefore, there is a need for ML approaches to help in image segmentation and some authors have already implemented this technique to improve clinical practice[60,61]. Moreover, ML can help with the integration of more complex information beyond imaging to study and diagnose liver diseases, since patients with CLD in the developmental phase require frequent follow-ups to check the progress of the disease and early detect changes in the diagnosis[58]. For example, patients with HHV-induced CLD are normally on antivirals, however there is no consensus or guidelines about when to stop antiviral therapy or even if quitting these drugs will increase HCC risk. Therefore, new approaches need to be established to classify and prevent the development of more severe illnesses, like cirrhosis or cancer. In this line, ML approaches can be used to measure liver fibrosis, optimize diagnosis, and predict disease progression of CLD[62]. Table 1 summarizes selected studies that have used ML for these purposes, which have been collected for this review, and Table 2 summarizes the most repeated inputs from all compiled ML models along with the most repeated predictive results for the main four inflammation-related liver conditions.

ML in inflammation-related liver disease

In the last years, promising results have been found when applying ML approaches in CLD. Regarding prevention, Fialoke *et al*[63] screened 108139 patients to identify those diagnosed with benign steatosis and NASH, a type of NAFLD, train ML classifiers for NASH and healthy (non-NASH) populations and predict NASH disease status on patients diagnosed with NAFLD according to aspartate transaminase (AST), alanine transaminase (ALT), and platelet levels. In this line, another study detected body mass index (BMI), triglycerides (TG), gamma-glutamyl transpeptidase (GGT), ALT, and uric acid as the top 5 features contributing to NAFLD, being the BN the model that performed best[64]. Accordingly, Yip *et al*[65] selected TG, ALT, white blood cell count, HDL-c, hemoglobin A1c (HbA1c), and the presence of hypertension, as the six variables to build ML models, of which Ada-Boost outperformed the others individually and described the NAFLD status in 922 subjects. More recently, Pei *et al*[66] designed a ML model which integrated medical records as a clinical variable to classify FLD. Concretely, they selected the variables of age, height, BMI, hemoglobin, AST, glucose, uric acid, low-density lipoprotein protein, alpha-fetoprotein, TG, HLD protein, and carcinoembryonic antigen. They tested six different ML models in 3419 participants, of which 845 were diagnosed with FLD: LR, RF, ANN, KNN, extreme gradient boosting (XGBoost) (a type of GB model), and LDA. Results from these authors showed that the XGBoost model had the highest performance, followed by LR and ANN, to predict the risk of FLD. BMI, uric acid, and TG levels were the top three variables associated to FLD risk across the six analysed models.

When it comes to diagnosis and treatment, several ML models have been tested for different purposes obtaining good specificity, sensitivity, and accuracy values[62]. For example, to determine the stage of liver fibrosis, some authors have used CT images processed by segmentation algorithms. Choi *et al*[67] used CNN upon CT images, whereas Chen *et al*[68] employed RF, KNN, SVM and the NB classifiers with real-time tissue elastography imaging, age, and sex as feeding variables. In both cases ML approach outperformed the classical methods. Regarding treatment, different ML models have been used to define the best therapy for liver diseases such as carcinomas and virus-induced hepatitis. Jeong *et al*[69] used DNN to classify intrahepatic cholangiocarcinoma susceptible to adjuvant therapy following resection according to laboratory and clinicopathological markers and found it more accurate than the commonly used staging system. Wübbolding *et al*[70] studied the prediction of early virological relapse analyzing soluble immune markers using supervised ML approaches like KNN, RF and LR. This study showed that IL-2, monokine induced by interferon γ/CCL9, RANTES/CCL5, stem cell factor, and TNF-related apoptosis-inducing ligand in combination were more reliable in predicting virological relapse than viral antigens. In the same way, researchers have used ML classifiers to explore new methods able to better predict prognosis of liver diseases[71-74]. The weighted variables are usually CT images and/or biochemical parameters that involved invasive and costly methods. However, researchers have recently proposed volatile organic compounds as new biomarkers for progression and prognosis of liver disease. These researchers monitored isoprene, limonene, and dimethyl sulphide concentrations from a breath sample in liver patients compared to healthy subjects. They used regression ML models (LR, ETR, SVR, and RF) to demonstrate that these approaches together with breath profile data can predict clinical scores of liver disease[75]. These findings are promising and open the way for new safe and non-invasive approaches to study liver function and for diagnosis purposes.

ML methods have been also employed when studying the comorbidities of liver-related diseases, like obesity, diabetes, and cardiovascular diseases[53,55,76]. For example, ML algorithms have been built to study the risk factors associated to overweight and obesity development, showing that BMI, age, dietary pattern, blood test results, socioeconomic status, and sedentarism were key factors when studying excess of adipose tissue[77]. In this line, further research has revealed by ML techniques that the minutes devoted to physical activity in one week[78], as well as specific species of gut microbiota[79], are also crucial for obesity prediction. ML algorithms have also elucidated the risk factors of childhood obesity, of which parental BMI and the upbringing environment play a huge role[80-82]. Furthermore, researchers have observed by training a multivariate LR model with a dataset of 3634 children and adolescents vitamins' intake, that vitamins A, D, B1, B2, and B12 were associated in a negative manner with obesity in this cohort[83]. These results are of interest, since new insights are needed to discover novel targets to tackle comorbidities that affect liver function.

ML in hepatitis virus-induced liver damage

HBV and HCV infections can dangerously become chronic if not treated early and with the right treatment[84]. While scientists are still relentlessly working on an effective vaccine against HCV, a good and efficient diagnosis is key to prevent chronic HCV infection (CHC), and ML algorithms have been elucidated for this purpose. Thus, Butt *et al*[85] designed an ANN model and trained it with a dataset of 19 variables, among which age, gender, BMI, transaminase, and platelet (PLT) count levels were included. The algorithm was able to better identify the stage of hepatitis C compared to other XGBoost, RF, and SVM models tested by other researchers with a higher precision rate and a decreased missing rate.

ML algorithms have been also applied and compared to traditional methods used to follow HHV-induced advanced liver disease[86-88]. For instance, Wei *et al*[87] used a GB model trained with the same variables that the formula fibrosis-4 (FIB-4) uses, which are age, AST, ALT and PLT levels in a cohort of 490 HVB patients, and two cohorts of HCV patients (n = 240 each). The GB model outperformed FIB-4 score in classifying hepatic fibrosis and the existence of cirrhosis. Barakat *et al*[89] designed a RF model that also outperformed FIB-4 score, as well as the AST/platelet ratio index (APRI), for prediction and staging of fibrosis in children with hepatitis C. In this line, data of 72683 veterans with CHC were used to predict the progression of the disease. GB models were used and compared with cross-sectional or linear models fed with variables like transaminases levels, alkaline phosphatase (ALP), PLT, AST, APRI, albumin, bilirubin, glucose, white blood cells, and BMI were included in the dataset. Results showed that APRI, PLT, AST, albumin, and AST/ALT ratio were the best predictors for featuring CHC progression[88].

Regarding therapy, CHC can be effectively treated with direct-acting antiviral (DAA) therapy, a novel treatment that targets viral non-structural proteins. Although it has null side effects compared to standard treatment, it has some downsides: Treatment failure in a low percentage of the cases, a very high cost, and no treatment duration established[90]. New methods to define this therapy duration are needed to optimize adherence and success. Precisely, Feldman *et al*[91] studied the prediction of DAA treatment duration in hepatitis C patients using XGBoost, RF, and SVM models. They used the dataset of 240 patients with prolonged first course of DAA against another one of 3478 patients on standard duration. Age, gender, comorbidities, and previous hepatitis C treatment record were considered. The predictive model constructed with XGBoost obtained the best performance in predicting prolonged DAA treatment, in which the presence of cirrhosis, type 2 diabetes, age, HCC and previous standard treatment were the most determining variables. Meanwhile, Kamboj *et al*[92] used ML approaches in the search of repurposed drugs that could target non-structural proteins, developing regression-based algorithms able to identify inhibitors of these proteins, and proposing new drugs to test in CHC.

A huge milestone when treating chronic HBV infection (CHB) is seroclearance of HBV surface antigen (HBsAg)[84]. It has been demonstrated that seroclearance of HBsAg is associated to a better prognosis in CHB. Some authors used ML models to predict HBsAg seroclearance in a cohort of 2235 patients, of which 106 achieved it. They used XGBoost, RF, and LR, among other models, and tested a total of 30 categorical and continuous variables, including gender, drinking history, initial diagnosis and treatment, age, BMI, and serum and radiological indicators. Results revealed that XGBoost model showed the best predictive performance, indicating that HBsAg levels were the best predictor for HBsAg seroclearance, followed by age and the level HBV's DNA[93].

Interestingly, ML has also contributed to personalized medicine in this field. HHVs evolve and adapt to different cellular environments in order to scape immune responses and drugs to survive. These adaptations rely on high mutagenetic activity, especially within the target genes of antivirals. Regarding HBV, Chen *et al*[94] used ML to identify patients with HCC or CHB based solely on genetic differences and found that the RF model impressively discriminated both cases based on the *rt* gene sequence of HBV. Moreover, Mueller-Breckenridge *et al*[95] ultra-deep sequenced 400 HBV samples and used a RF model to classify the status of a particular HBsAg according to the novel viral variants encountered. Results showed 5 genotypes that could benefit personalized healthcare. In the case of HCV, Kayvanjoo *et al*[96] built several ML algorithms and trained them with two datasets of responders *vs* non-responders of antiviral therapy in HCV-infection caused by two different strains. These investigations reported novel genetic markers that could predict therapy response with high accuracy. These results are very promising since they contribute to bringing personalized medicine to the public system.

ML in COVID-19-induced liver damage

A recent systematic review depicted that the parameters normally used for liver impairment screening were significantly increased in COVID-19 patients[23]. Particularly, several studies show that levels of AST or/and ALT can increase in these patients up to 20%, bilirubin up to 14%, ALP up to 6%, and GGT levels up to 21%. Prothrombin is a protein synthesized in the liver that results in thrombin, a protein with a critical role in coagulation function. Prolonged prothrombin is a symptom of decreased production of coagulation factors, characteristic of liver disease. For this reason, the prolonged prothrombin time (PT) is another parameter usually checked when screening for liver injury, and it has been described that COVID-19 patients present nearly a 10% increase in PT[23]. Besides biochemical alterations, COVID-19 illness can lead to hypoxemia, impaired cardiac function, and secondary damage due to multiple organ dysfunction, what can result in liver injury in patients with or without a prior liver disease. Therefore, new insights in the relationship between this recent infectious illness and liver disease are expected.

The use of ML approaches has been encouraged by the National COVID Cohort Collaborative Consortium in order to early detect, predict and follow up severe COVID-19 cases since the pandemic started[97]. For instance, some researchers used the XGBoost approach and found that age, CT scan result, body temperature, lymphocyte levels, fever, and coughing, can classify influenza patients from COVID-19 patients[98].

Bhargava *et al*[99] tried different ML approaches to detect novel COVID-19 and discriminate between pneumonia using CT and X-ray scans as inputs. These authors pre-processed by normalization the images and then segmented them by fuzzy c-means clustering. Results showed that SVM model was the one that better classified patients in COVID-19 positive, pneumonia, and healthy groups, obtaining a very high accuracy. In this same line, obesity and liver disease were identified as risk factors for higher clinical severity in a cohort of 174568 adults with severe acute respiratory syndrome associated with SARS-CoV-2 infection by a multivariable LR model[97]. Interestingly, a German study of 8679 patients used a LR model and come to again identify liver disease and BMI as determinant risk factors for 180-d all-cause mortality in hospitalized COVID-19 patients[100]. A case-control study with COVID-19 patients compared to patients with community-acquired pneumonia showed how, by applying a GB model, the category of liver function appeared as one of the top systematic predictors for COVID-19 risk factors, being albumin, total bilirubin, and ALT among the most important input variables[101]. Furthermore, a study with 710 enrolled patients diagnosed with COVID-19 identified AST levels as the top predictor for COVID-19 related hospitalization based on a RF algorithm, followed by age and diabetes mellitus[102]. A stepwise linear regression model identified IL-6 and granzyme B as potential predictors of liver dysfunction, characterized by an elevation in the levels of ALT and/or AST[103]. Other authors designed a model for detecting liver damage testing different ML approaches with laboratory parameters as the input variables. SVM was the model with best accuracy, and AST and ALT levels the variables with best predictive scores[104]. In this context, the newest version of the CURIAL model was developed to identify COVID-19 patients by using vital signs, blood gas, and laboratory blood tests. It showed greater sensitivity, making this model a potential emergency workflow[105,106]. All these ML-based methods would dramatically improve time of diagnosis, free hospitals' laboratories and rooms of potential positive subjects, and reduce costs if implemented in the public health system.

AI has also been employed to discover new drugs potentially efficient to tackle SARS-CoV-2 infection[107]. Baricitinib is a drug initially approved for rheumatoid arthritis that was selected by ML as a potential drug to treat COVID-19. Researchers proved the anti-inflammatory and antiviral properties of this drug in human liver spheroids infected with live SARS-CoV-2 for, among others, check any potential drug-induced liver injury[107]. Due to the good results, researchers moved on to a clinical trial where they tested baricitinib in a few COVID-19 patients. Levels of liver enzymes were not altered, except for a transient increase in liver aminotransferases in all patients that remitted in the following 72 h without interrupting treatment. Authors state that this might be reflective of disease severity rather than a drug-induced injury, showing overall well tolerance and results in this pilot study[108]. In summary, ML approaches support liver biochemistry as a prognostic tool in COVID-19 disease.

PERSONALISED MEDICINE IN LIVER-RELATED DISEASES SUPPORTED BY ML

In the early 21st century, the Human Genome Project started the genomic era in which new disciplines like precision medicine appeared. Precision medicine aims to deliver targeted treatments based on a group of individual factors that greatly influence the onset and progression of a disease, like omics sciences. This approach covers a great number of patients, overcoming potential drug adverse effects and ensuring effectiveness of the treatment. In this context, computational advances have greatly contributed to the escalation of this science by lowering the costs of omics analysis and allowing the processing and integration of enormous amount of data based on ML algorithms (Figure 2).

ML have permitted the development of diagnostics and therapeutics based on the integration of omics data (genomics, epigenomic, transcriptomics, proteomics, metabolomics, and metagenomics) with clinical data. The ultimate goal is to bridge these omics data with the phenotype to bring molecular accuracy to the diagnosis, treatment, prognosis, and recurrence process of a pathological condition. This methodology has been used in a wide range of diseases in the search of more efficient and effective approaches, like heart and liver diseases[109,110]. For example, ML algorithms fed with omics data have been able to predict mortality in patients with alcoholic hepatitis. In this study, routine clinical variables of 210 patients with this disease were used to build 6 different datasets to assess mortality at 30 and 90 d. Five different ML models were tested, obtaining the best performance in predicting 30-d mortality with a GB model using bacteria, MetaCyc pathways and clinical data, as well as LR using viral and clinical data[111].

In hepatitis B, it has been found that ML algorithms can be very useful in assessing HBV associated-HCC progression. Ye *et al*[112] analyzed 67 HBV-positive HCC samples without or with intrahepatic metastases and discovered key genes for metastatic progression and survival training ML models. The majority of them were inflammatory or related with inflammation process, like IL-2 receptor and osteopontin, which encodes an extracellular cytokine ligand whose overexpression favors metastasis. These authors were able for the first time to draw a molecular signature useful to classify metastatic HBV-HCC patients, opening the way for early detection and new treatments to increase patient survival. In hepatitis C, the CC and CT genotypes of rs12979860 polymorphism in the *IL28B* gene have been associated to liver fibrosis progression, being able to predict antiviral treatment effectiveness[113]. Moreover, ML algorithms have allowed to diagnose advanced liver fibrosis according to rs12979860 genotype with higher performance compared to APRI and FIB-4 scores[114]. In this study, patients were divided in two groups according to HCV-related liver fibrosis stage: None to moderate fibrosis (n = 204) or with advanced fibrosis (n = 223). ML algorithms revealed *IL28B* genotype as first predictor, while the second one depended on the mentioned genotype. For instance, in CT patients, PLT, albumin, and age were the determining variables, while for TT patients, white blood cell count was the decisive feature to assess advanced fibrosis probability.

ML approaches have also helped to categorize obesity in different subtypes based on metabolic status[115-117]. For example, Masi *et al*[115] studied a cohort of 2567 subjects suffering from obesity and made clusters of metabolically healthy, or metabolically unhealthy patients based on clinical and biochemicals variables using two ML models. The first model showed that IR, body fat, HbA1c, red blood cells, age, ALT, uric acid, white blood cells, insulin growth factor-1, and γGT were the top predictors of a metabolically healthy obesity, revealing the importance of liver function. Other authors have also used ML models to classify 882 obese patients in subtypes of obesity according to glucose, insulin, and uric acid levels[116]. Results showed four stable metabolic clusters in this cohort, which were characterized by a healthy metabolic status, or by hyperuricemia, hyperinsulinemia, and hyperglycemia, respectively. Furthermore, Lee *et al*[117] explored three-way interactions between genome, epigenome, and dietary/lifestyle factors using GB and RF models in a subset (n = 394) of the exam 8 of the Framingham Offspring Study cohort. Interestingly, GB obtained the best performance, revealing 21 single nucleotide polymorphisms, 230 methylation sites in relevant genes (like *CPT1A*, *ABCG1*, and *SREBF1*), and 26 dietary factors as top predictors for obesity. Intake of processed meat, artificially sweetened beverages, French fries, and alcohol intake, among other dietary factors, was highly associated with overweight/obesity.

Personalized and precision medicine aims to harmonize the greatest number of factors so that diagnosis, prognosis, and treatment are based on the greatest number of decision elements. Much remains to be investigated to establish guidelines in the context of personalized medicine. However, it is safe to say that precision medicine will drive modern medicine, combining the most classic variables with the newest digital ones. Health professionals must be prepared to understand and implement these new technologies in the near future.

CONCLUSION

In summary, ML science can process and integrate a vast amount of different data with unprecedented outstanding performance. The objective of this article was to collect the information derived from ML techniques in liver damage induced by inflammatory conditions, including the new disease COVID-19. The main role of ML in liver pathologies is to help identify high risk patients for referral to specialized centres. Results show that the use of ML models have brought new insights into biology and medicine questions that can be very useful in determining the next directions towards research in diagnosis, prognosis, and treatment of inflammatory and virus-related liver diseases, leading the way to personalized medicine. Also biomarkers concerning inflammation/IR related to liver disease can be boosted by ML strategies. This review clarifies and compiles the importance of the different factors involved in CLD and analysed by ML algorithms, which can be useful information for clinicians, like endocrinologists and gastroenterologists, and other healthcare professionals with a focus on hepatology and bioinformatics.

79914_Auto_Edited-check.docx

ORIGINALITY REPORT

3%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet	49 words — 1%
2	link.springer.com Internet	30 words — < 1%
3	v3r.esp.org Internet	25 words — < 1%
4	www.semanticscholar.org Internet	23 words — < 1%
5	doaj.org Internet	14 words — < 1%
6	res.mdpi.com Internet	14 words — < 1%
7	doctorpenguin.com Internet	13 words — < 1%
8	pubmed.ncbi.nlm.nih.gov Internet	13 words — < 1%
9	www.nature.com Internet	13 words — < 1%
10	stemcellres.biomedcentral.com Internet	

12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS