Liver International

WILEY

Liver INTERNATIONAL

REVIEW

Artificial Intelligence: An Emerging Tool for Studying Drug-Induced Liver Injury

Hao Niu[1,2,3] | Ismael Alvarez-Alvarez[1,2,3] | Minjun Chen[4]

[1]Servicios de Aparato Digestivo y Farmacología Clínica, Hospital Universitario Virgen de la Victoria, Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND, Universidad de Málaga, Málaga, Spain | [2]Centro de Investigación Biomédica en Red Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto de Salud Carlos III, Madrid, Spain | [3]Plataforma de Investigación Clínica y Ensayos Clínicos IBIMA, Plataforma ISCIII de Investigación Clínica, SCReN, Madrid, Spain | [4]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, USA

Correspondence: Ismael Alvarez-Alvarez (iaalvarez@uma.es) | Minjun Chen (minjun.chen@fda.hhs.gov)

Received: 27 November 2024 | **Revised:** 29 January 2025 | **Accepted:** 8 February 2025

Handling Editor: Raúl Andrade and Luca Valenti

Funding: This study was supported by grants from Instituto de Salud Carlos III, cofunded by Fondo Europeo de Desarrollo Regional—FEDER, cofunded by the European Union (grant number: PI21/01248; PID2022-140169OB-C21, PT23/00137) and by the Agencia Española de Medicamentos y Productos Sanitarios. CIBERehd and Plataforma de Investigación Clinica are funded by ISCIII. HN holds a postdoctoral research contract funded by Junta de Andalucía (POSTDOC_21_00780). Funding for open access charge: Universidad de Málaga/CBUA. The funding sources had no involvement in the writing of the report or in the decision to submit the manuscript for publication.

Keywords: artificial intelligence | drug-induced liver injury | hepatotoxicity | large language model | machine learning | natural language processing

ABSTRACT

Drug-induced liver injury (DILI) is a complex and potentially severe adverse reaction to drugs, herbal products or dietary supplements. DILI can mimic other liver diseases clinical presentation, and currently lacks specific diagnostic biomarkers, which hinders its diagnosis. In some cases, DILI may progress to acute liver failure. Given its public health risk, novel methodologies to enhance the understanding of DILI are crucial. Recently, the increasing availability of larger datasets has highlighted artificial intelligence (AI) as a powerful tool to construct complex models. In this review, we summarise the evidence about the use of AI in DILI research, explaining fundamental AI concepts and its subfields. We present findings from AI-based approaches in DILI investigations for risk stratification, prognostic evaluation and causality assessment and discuss the adoption of natural language processing (NLP) and large language models (LLM) in the clinical setting. Finally, we explore future perspectives and challenges in utilising AI for DILI research.

Abbreviations: AI, artificial intelligence; ALF, acute liver failure; AlogP, Ghose-Crippen-Viswanadhan octanol–water partition coefficient; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUROC, area under the receiver operating characteristic curve; BERT, Bidirectional Encoder Representations from Transformers; Bolasso, bootstrap-enhanced least absolute shrinkage operator; CIOMS/RUCAM, Council for International Organisations of Medical Sciences/Roussel Uclaf Causality Assessment Method; CNN, convolutional neural network; DILI, drug-induced liver injury; DILIN, Drug-Induced Liver Injury Network; DILIst, Drug-Induced Liver Injury Severity and Toxicity; DPA, docosapentaenoic acid; EMA, European Medicines Agency; FDA, Food and Drug Administration; ICD, International Classification of Diseases; Lasso, least absolute shrinkage and selection operator; LLM, large language model; LTKB, Liver Toxicity Knowledge Base; MCC, Matthews correlation coefficient; NLP, natural language processing; PUFA, polyunsaturated fatty acid; QSAR, quantitative structure–activity relationship; RECAM, Revised Electronic Causality Assessment Method; SIDER, Side Effect Resource; ULN, upper limit of normal.

Disclaimer: This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement or recommendation.

Summary

- Drug-induced liver injury (DILI) is a complex disease with a challenging diagnosis, posing a significant public health risk.
- Artificial intelligence (AI)-driven models have advanced DILI research for improving risk stratification, prognostic evaluation and causality assessment.
- In DILI, natural language processing or large language models are mainly used in literature classification, and their utilities as diagnostic tools remain underexplored.
- Overall, AI methodologies offer promising opportunities to advance DILI research.

1 | Introduction

Drug-induced liver injury (DILI) is an uncommon, complex and potentially severe adverse reaction to xenobiotics, including synthetic drugs, herbal products or dietary supplements. Unlike paracetamol overdose (intrinsic drug-induced liver injury) idiosyncratic DILI is usually not dose-related, has a variable onset (days to weeks), and is a common safety cause of postmarketing drug withdrawal [1]. The diagnosis of DILI is challenging due to its unspecific clinical presentation and lack of biomarkers, relying on excluding other liver damage causes. Although DILI often resolves on its own, some cases may progress to acute liver failure (ALF), requiring liver transplantation or resulting in death [2]. A multicentric cohort study estimated that idiosyncratic DILI accounts for 11% of ALF cases in the United States, ranking it the third most frequent cause of ALF after acetaminophen overdose (46%) and ALF of indeterminate cause (12%) [3]. Given its public health risk, new methodologies to better understand DILI are a compelling research opportunity.

In recent years, artificial intelligence (AI) has emerged as a powerful tool for scientific research. Chen et al. [4] reviewed the state-of-the-art predictive models for DILI, identifying limitations such as a lack of understanding of DILI mechanisms, scarcity of human data and insufficient bioinformatic capabilities. However, these limitations have been partially overcome. Rapid advances in the field of omics, that is, the term that encompasses genomics, transcriptomics, proteomics and metabolomics, from the study of genetic variants or changes in genetic expression to the detection of specific signatures in protein or metabolite abundance, have fostered the investigations into DILI mechanisms [5]. Additionally, improved computational capabilities have led to the development of complex in silico models, integrating pharmacokinetic models, drug properties and population variability to estimate DILI risk [6]. Furthermore, the emergence of real-world data from electronic health records, patient-reported outcomes and digital health devices has provided a robust data source beyond traditional epidemiological data [7], though its adoption for studying DILI remains limited [8, 9].

The use of AI methods provides a powerful and innovative research tool capable of handling massive amounts of data, thereby opening new research avenues for gaining deeper insight into DILI. In this review, we summarised the current evidence of AI strategies for the study of idiosyncratic DILI. This review is divided into the following sections: (1) Basic concepts of AI; (2) AI as an emerging tool in DILI; (3) The application of natural language processing (NLP) and large language models (LLMs) and (4) Future perspectives.

2 | Basic Concepts of Artificial Intelligence

The concept of AI was defined as a discipline focused on the science and engineering of constructing intelligent machines [10]. AI encompasses a wide range of technologies that enable computers to replicate human cognitive functions, including learning, pattern recognition, prediction and NLP, thereby performing tasks traditionally requiring human intelligence (Figure 1). In Table S1, a glossary of AI-related terms that will be used in this manuscript is provided.

2.1 | Machine Learning

Machine learning, a subset of AI, is centered on the development of algorithms or statistical models capable of learning from data and making predictions, with the ability to enhance performance over time without extensive reprogramming [11]. Machine learning is generally divided into several categories: (1) supervised learning, where models are trained on labelled datasets to predict specific outcomes; (2) unsupervised learning, which seeks to uncover hidden structures in unlabeled data through techniques such as dimensionality reduction and clustering; (3) semi-supervised learning, which falls between supervised and unsupervised learning with some data not labelled; and (4) reinforcement learning, where the algorithm learns to make decisions by interacting with an environment to maximise cumulative rewards based on feedback from its actions. The machine learning process typically begins with the collection of datasets, sourced from various inputs such as text, images, or

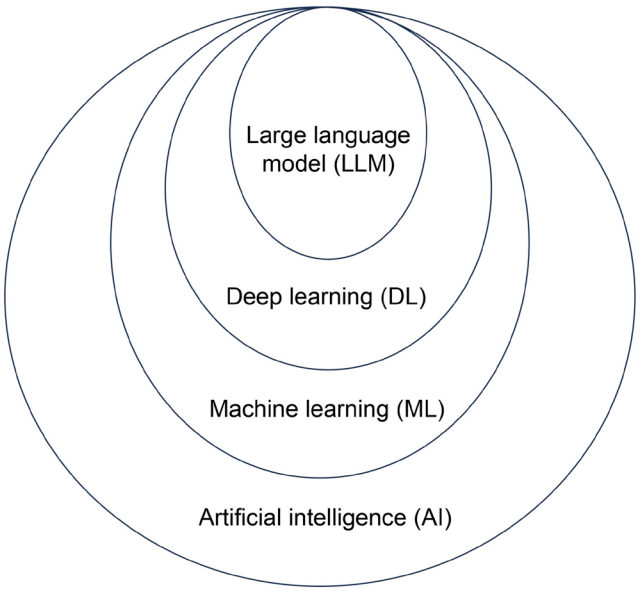

FIGURE 1 | Relationship between artificial intelligence (AI), machine learning (ML), deep learning (DL) and large language model (LLM).

sensor readings, which form the basis for model training. The selection of an appropriate algorithm, such as regression analysis, neural networks or clustering techniques, directs the learning process. Following training, the model undergoes rigorous evaluation, with its performance assessed using statistical metrics, including accuracy, the Matthews correlation coefficient (MCC) and recall. When the model's performance meets predefined thresholds, it is deemed suitable for deployment in real-world applications, such as DILI diagnostics, risk stratification and causality assessment.

2.2 | Deep Learning

Deep learning represents an advanced subset of machine learning that employs multiple layers of artificial neural networks to detect complex patterns within large datasets [12]. These networks are designed to emulate the architecture and functionality of biological neural systems, particularly the human brain, enabling them to process vast quantities of data. A typical deep learning network comprises numerous interconnected nodes or neurons, organised into layers: an input layer that receives data, one or more hidden layers where data processing occurs, and an output layer that generates predictions or classifications. Compared to traditional machine learning, deep learning excels at managing larger datasets and tackling more complex problems, often achieving superior accuracy and performance in fields such as image and speech recognition. One of its primary advantages is the ability to automatically extract features from raw data, reducing the need for manual feature engineering [13]. This capability is particularly valuable when working with unstructured data, such as images, audio or text, where deep learning algorithms can autonomously identify intricate patterns without human intervention. Additionally, deep learning models are highly scalable, improving in performance as they are exposed to more data. The success of deep learning has been fuelled by advancements in computational power and the availability of large-scale datasets. Deep learning also showed significant promise in the study of DILI. By analysing vast amounts of biomedical data, such as genomic information, histopathological images and clinical records, deep learning models can identify biomarkers and predict the likelihood of DILI in patients. For instance, convolutional neural networks (CNNs) can process liver biopsy images to detect subtle signs of injury [14]. However, deep learning faces challenges similar to those in traditional machine learning, including issues of overfitting, limited interpretability and the requirement for large amounts of labelled data, all of which can complicate its practical implementation.

2.3 | Natural Language Processing and Large Language Models

NLP is a specialised subfield of AI focused on enabling effective interaction between computers and humans using natural language [15]. LLMs, such as OpenAI's GPT series [16], represent advanced NLP systems designed to both comprehend and generate human language. LLMs are deep learning algorithms that can recognise, summarise, translate and generate text and other forms of content, utilising knowledge acquired during their pre-training phase. These models are trained on vast corpora of text data, allowing them to capture nuanced meanings, contextual cues and even the subtleties of tone and style. The fundamental principle involves training these models on diverse datasets, equipping them to grasp the complexities of language, context and subtle meanings. Structurally, LLMs are composed of multiple layers of neurons that process input data through intricate mathematical operations, iteratively refining their understanding to produce coherent and contextually appropriate text. In the medical domain, LLMs offer substantial potential to transform various applications. For example, they can assist healthcare professionals by automating clinical documentation, generating patient summaries and facilitating data retrieval from extensive medical literature [17]. Furthermore, by translating complex medical terminology into accessible language, LLMs can improve patient management for DILI, ensuring that individuals better understand their diagnoses and treatment plans [18].

3 | Artificial Intelligence as an Emerging Tool in DILI

3.1 | Risk Stratification

As abovementioned, idiosyncratic DILI is an unexpected, dose-independent reaction that can mimic other liver diseases due to its non-specific clinical presentation. The absence of specific biomarkers complicates the DILI diagnosis and patient stratification for the management of hepatotoxicity risk. The emergence of AI computational methods has drawn the attention of researchers for developing and comparing predictive models for DILI risk using innovative approaches.

3.1.1 | Clinical Studies

For instance, Ma et al. [19], conducted a single-centre retrospective study, creating an automated machine learning model to predict valproic acid-associated transaminase elevations in 1,995 epilepsy patients. They identified white blood cell count, age, height, total bilirubin, valproic acid plasma concentration and oral dose as predictive factors using a univariate generalised linear model and trained the model using several machine learning methods, including gradient boosting, generalised linear models, random forest, deep neural networks and extreme random trees. Among these, the gradient boosting machine-based model showed the best performance.

In another retrospective study, Yu et al. [20] identified diagnostic biomarkers in the salivary metabolites of DILI patients. They used principal component analysis and partial least squares discriminant analysis to evaluate metabolic alterations between DILI patients and healthy controls, followed by a weighted metabolite co-expression network analysis to cluster sets of metabolites related to DILI. Eight candidate metabolites were identified and were used as the inputs for modelling through least absolute shrinkage and selection operator (Lasso) and random forest analysis. Both models showed excellent performance, with five metabolites, i.e., (1) 3-hydroxydecanoic acid; (2) 12-hydroxydecanoic acid; (3) tetradecanedioic acid; (4) hypoxanthine; and (5) inosine, emerging as potential diagnostic biomarkers for DILI.

In a retrospective study with over 7,000 patients with tuberculosis, Xiao et al. [21] developed machine learning models using Lasso-logistic regression, random forest and extreme gradient boosting to predict DILI from anti-tuberculosis drugs. The extreme gradient boosting model demonstrated slightly better discriminatory capacity compared to the others.

In addition, Gonzalez-Jimenez et al. [22] conducted a cross-sectional study within the Spanish DILI Registry to identify host factors and drug properties predicting the biochemical type of liver injury in DILI cases. Using random forest analysis, they identified predictors such as age, therapy duration, daily dose, lipoaffinity index, AlogP, serum half-life, vascular diseases and hybridisation ratio. The resulting decision tree model classified DILI cases based on the predicted probabilities of hepatocellular or cholestatic liver injury. Furthermore, they developed a predictive model for practical use that included age, lipoaffinity and hybridisation ratio, showing a fair discriminant power.

3.1.2 | Genomics

Notably, genomics has emerged as a valuable source of information for AI-driven construction of DILI predictive models. Feng et al. [23] extracted expression data of over 15,000 genes from in vivo DILI positive and negative samples. A total of 375 and 1,574 feature genes were selected based on weight values of vectors. Comparing with support vector machine, a larger gene set based on the deep learning approach achieved superior performance.

Likewise, Wang et al. [24] identified potential genetic biomarkers from 109 DILI samples and 29 healthy controls from the Gene Expression Omnibus database [25]. They analysed the differential gene expression patterns and built six machine learning-based prediction models, identifying genes with a weight of over 1, such as *DDIT3*, *GADD45A*, *SLC3A2* and *RBM24*, as potential biomarkers.

Also, Lai et al. [26] compared artificial neural network, support vector machine and random forest for predicting DILI due to anti-tuberculosis drugs in 127 patients (21 cases with DILI and 106 healthy controls). The artificial neural network model, combining clinical and genetic data, showed the best discriminative power.

3.1.3 | Quantitative Structure–Activity Relationship (QSAR) Models

In addition to modelling with patients' clinical information or omics-derived data, other studies have proposed alternative strategies, such as quantitative structure–activity relationship (QSAR) modelling. QSAR uses physicochemical properties of drugs to build in silico DILI predictive models.

For example, Hong et al. [27] used the DILIrank database [28] and the Mold^2 software [29] to generate molecular descriptors for 721 drugs, developing QSAR predictive models with a decision forest algorithm. The 3-class DILI risk prediction model provides improved performance to differentiate most-DILI drugs from no-DILI drugs in cross-validation and bootstrapping in comparison with the 2-class model.

Furthermore, Li et al. [30] developed the DeepDILI model, a deep learning-powered prediction model. They retrieved information on 1,002 drugs from the FDA DILIst dataset [31] and generated molecular descriptors from the Mold^2 software [29]. By integrating five machine learning algorithms as base classifiers to build the DeepDILI model, the deep learning meta-classifier outperformed individual base classifiers and effectively predicted DILI for newly approved drugs.

More recently, Yang et al. [32] developed several machine learning algorithms based on the information retrieved from the Liver Toxicity Knowledge Base (LTKB) and the LiverTox database [33, 34]. They calculated molecular representation with six physicochemical descriptors and molecular fingerprints. Deep neural network models using ECFP_6, ECFP_4 and MACCS fingerprints showed the best performance. Furthermore, drugs containing certain molecular fragments, like nitrogen and oxygen atoms, were identified as more likely to cause DILI.

In another study, Adeluwa et al. [35] combined molecular structural information, toxicity data, FDA adverse event reports, and gene expression patterns in six cell lines to develop machine learning models for predicting a drug's hepatotoxic potential. They focused on endpoints, including severe DILI associated with liver necrosis, withdrawal or box warning due to liver toxicity, and positive and negative control groups. The individual models showed low sensitivity, leading the researchers to construct an ensembled model to improve performance. However, this ensembled model did not enhance prediction accuracy, indicating that available data on microRNA quantification, molecular descriptors, toxicology profiles and reported events may be insufficient for accurately classifying DILI in real-world scenarios.

Detailed characteristics of the abovementioned studies are summarised in Table 1.

3.2 | Prognostic Evaluation

DILI usually manifests as mild or moderate self-limited damage that resolves spontaneously. However, it can sometimes progress to a fatal outcome. The absence of validated prognostic biomarkers complicates the determination of individual risk for worsening conditions, highlighting the potential of AI computational methodologies to identify prognostic risk factors. Some studies have used AI to establish predictive factors for the outcome in patients with DILI.

Recently, Niu et al. [36] analysed the data from over 900 DILI patients in the Spanish DILI Registry, developing a machine learning model using the bootstrap-enhanced least absolute shrinkage operator (Bolasso) procedure to identify prognostic factors for fatal outcomes (liver transplantation or liver-related death). This model included six factors, i.e., 1) prior drug allergies; 2) hepatocellular damage; 3) female sex; 4) total bilirubin; 5) aspartate aminotransferase (AST); and 6) platelet count, showing excellent discrimination capability. The model was

TABLE 1 | Characteristics of studies that used artificial intelligence to establish a predictive model for risk stratification for drug-induced liver injury (sorted by date).

Study	Study population	Dataset	DILI criteria	Methodology	Predictive model	Model performance	Validation/testing cohort/set	Model performance (validation/testing)
Hong et al. (2017)[a] [27]	—	DILIrank, Mold[2]	Classification of DILI potential based on FDA drug labeling and causality assessment in case reports	Decision forest	Mean atomic van der Waals volume, Moran autocorrelation—lag 2/weighted by atomic polarizabilities, ratio of multiple path counts to path counts, mean atomic polarizability (scaled on Carbon-SP3 atom), Geary autocorrelation—lag 2,3 and 4/weighted by atomic masses, average weight of molecular, Geary autocorrelation—lag 2 and 3/weighted by atomic Sanderson electronegativities	Accuracy: 0.729 Sensitivity: 0.628 Specificity: 0.798 MCC: 0.432 BA: 0.713	Yes	Paritaprevir (most-DILI) Confidence: 0.232 Ombitasvir (no-DILI) Confidence: 0.272 Dasabuvir (most-DILI) Confidence: 0.994 Ritonavir (most-DILI) Confidence: 0.504 Solithromycin (most-DILI) Confidence: 0.098
Feng et al. (2019)[b] [23]	—	ArrayExpress	Defined by severity of liver lesions caused by drugs	Deep learning, SVM	Gene differential expression: 375 genes Deep learning algorithm: 1574 genes	*375 genes* Deep learning Accuracy: 0.934 Sensitivity: 0.920 Specificity: 0.949 AUROC: 0.971 MCC: 0.868 SVM Accuracy: 0.879 Sensitivity: 0.778 Specificity: 0.980 AUROC: 0.892 MCC: 0.774 *1574 genes* Deep learning Accuracy: 0.971 Sensitivity: 0.974 Specificity: 0.968 AUROC: 0.989 MCC: 0.942 SVM Accuracy: 0.889 Sensitivity: 0.788 Specificity: 0.990 AUROC: 0.901 MCC: 0.794	Yes	*Treatment time 4 days* Accuracy: 0.958 Sensitivity: 0.960 Specificity: 0.957 AUROC: 0.988 MCC: 0.917 *Treatment time 8 days* Accuracy: 0.976 Sensitivity: 0.955 Specificity: 1.000 AUROC: 0.991 MCC: 0.953 *Treatment time 15 days* Accuracy: 0.983 Sensitivity: 0.966 Specificity: 1.000 AUROC: 0.996 MCC: 0.967 *Treatment time 29 days* Accuracy: 0.938 Sensitivity: 1.000 Specificity: 0.875 AUROC: 0.976 MCC: 0.882

(Continues)

TABLE 1 | (Continued)

Study	Study population	Dataset	DILI criteria	Methodology	Predictive model	Model performance	Validation/testing cohort/set	Model performance (validation/testing)
Lai et al. (2020)[c] [26]	127 patients	—	ALT > 2 ULN; DBL > 2 ULN; AST, ALP or TBL > 2 ULN	ANN, SVM, RF	Age, sex, weight, ALT, AST, smoking, drinking habit, NAT2*7, OATP1B1*1a/1a, OATP1B1*1a/15, UGT1A1*27/28	Accuracy: 0.880 Sensitivity: 0.750 Specificity: 0.905 AUROC: 0.894	Yes	Accuracy: 0.887 Sensitivity: 0.800 Specificity: 0.904 AUROC: 0.898
Gonzalez-Jimenez et al. (2021)[d] [22]	610 DILI cases	—	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	RF	Age, lipoaffinity, hybridization ratio	AUROC: 0.74	Yes	AUROC: 0.68
Li et al. (2021) [30]	—	DILIst, Mold^2	Consensus of DILI classifications from multiple datasets	Optimised neural network	203 molecular descriptors	Accuracy: 0.781 Sensitivity: 0.846 Specificity: 0.683 AUROC: 0.857 MCC: 0.538 BA: 0.765	Yes	Accuracy: 0.687 Sensitivity: 0.805 Specificity: 0.510 AUROC: 0.659 MCC: 0.331 BA: 0.658
Adeluwa et al. (2021)[e] [35]	PHH, HepG2, HA1E, A-375, MCF7, PC-3	DILIrank, Mold^2, Tox21, FAERS	Classification of DILI potential based on FDA drug labeling and causality assessment in case reports	LR, linear discriminant analysis, division trees, SVM, naïve Bayes, (one-layer) neural network, RF	PHH (SVM): 60 predictors HepG2 (SVM): 72 predictors HA1E (LR): 40 predictors A-375 (LR): 178 predictors MCF7 (LR): 65 predictors PC-3 (RF): 315 predictors	*PHH* Sensitivity: 0.912 Specificity: 0.945 AUROC: 0.969 *HepG2* Sensitivity: 0.924 Specificity: 0.693 AUROC: 0.922 *HA1E* Sensitivity: 0.903 Specificity: 0.389 AUROC: 0.781 *A-375* Sensitivity: 0.826 Specificity: 0.170 AUROC: 0.627 *MCF7* Sensitivity: 0.898 Specificity: 0.222 AUROC: 0.722 *PC-3* Sensitivity: 1.000 Specificity: 0 AUROC: 0.589	Yes	NA

(Continues)

TABLE 1 | (Continued)

Study	Study population	Dataset	DILI criteria	Methodology	Predictive model	Model performance	Validation/testing cohort/set	Model performance (validation/testing)
Ma et al. (2024) [19]	1995 patients	—	Abnormal increase of transaminase levels	Gradient boosting, generalised linear models, random forest, deep neural networks, extreme random trees	Age, height, total bilirubin, white blood cell count, VPA plasma concentration and oral dose	Training set AUROC: 0.855 Testing set AUROC: 0.789	Yes	AUROC: 0.742
Yu et al. (2024) [20]	31 DILI cases 35 healthy controls	—	AST or ALT > 5 ULN or ALP > 2 ULN; TBL > 2.5 mg/dL and elevated AST, ALT or ALP; INR > 1.5 and elevated AST, ALT or ALP	Lasso, RF	Inosine, hypoxanthine, 3-hydroxydecanoic acid, 12-hydroxydecanoic acid, tetradecanedioic acid	Lasso AUROC: 0.998 RF AUROC: 0.969	No	—
Xiao et al. (2024)[f] [21]	1151 DILI cases 5920 non-DILI cases	—	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	LR, RF, extreme gradient boosting	Outpatient DILI (once occurring), outpatient drug-induced hepatitis, (once occurring), outpatient drug-induced hepatitis (sporadically occurring), inpatient drug-induced hepatitis (once occurring), outpatient DILI (sporadically occurring), inpatient drug-induced hepatitis, (frequently occurring), inpatient DILI (once occurring), outpatient DILI (frequently occurring); TBL, ALP, inpatient DILI (sporadically occurring), ALT, fatty liver disease, education level, age	NA	Yes	*Logistic regression* Accuracy: 0.753 Sensitivity: 0.771 Specificity: 0.750 AUROC: 0.848 *Random forest* Accuracy: 0.799 Sensitivity: 0.782 Specificity: 0.802 AUROC: 0.877 *Extreme gradient boosting* Accuracy: 0.812 Sensitivity: 0.760 Specificity: 0.823 AUROC: 0.887

(Continues)

TABLE 1 | (Continued)

Study	Study population	Dataset	DILI criteria	Methodology	Predictive model	Model performance	Validation/ testing cohort/set	Model performance (validation/testing)
Yang et al. (2024)[g] [32]	—	LTKB, LiverTox	Combined DILI classifications from LTKB and LiverTox	RF, LR, naïve Bayes, SVM, k-nearest neighbour, extreme gradient boosting, adaptive boosting, deep neural network	ECFP_6: FP1019, FP378, FP695, FP726, FP222, FP208, FP935, FP887, FP537, FP656 ECP_4: FP650, FP378, FP913, FP486, FP80, FP15, FP41, FP231, FP64, FP356 MACCS: FP146, FP93, FP97, FP151, FP148, FP85, FP141, FP83, FP80, FP54	*E-state* Accuracy: 0.635 Sensitivity: 0.682 Specificity: 0.538 AUROC: 0.656 MCC: 0.213 BA: 0.610 *RDKFP* Accuracy: 0.689 Sensitivity: 0.773 Specificity: 0.519 AUROC: 0.671 MCC: 0.293 BA: 0.646 *MACCS* Accuracy: 0.734 Sensitivity: 0.821 Specificity: 0.558 AUROC: 0.693 MCC: 0.387 BA: 0.689 *Pubchem* Accuracy: 0.698 Sensitivity: 0.758 Specificity: 0.577 AUROC: 0.659 MCC: 0.330 BA: 0.668 *ECFP_4* Accuracy: 0.711 Sensitivity: 0.777 Specificity: 0.577 AUROC: 0.712 MCC: 0.352 BA: 0.677 *ECFP_6* Accuracy: 0.689 Sensitivity: 0.706 Specificity: 0.654 AUROC: 0.713 MCC: 0.344 BA: 0.680	Yes	NA

(Continues)

externally validated in the LATINDILI Network, with over 450 DILI patients, yielding a similar performance.

Some patients may not achieve biochemical resolution within 6 months to a year after DILI recognition, leading to chronic DILI, with a prevalence estimated at 8%–18% [37–39]. Ashby et al. [40] developed a machine learning-based score model to predict resolution trajectories for DILI. They analysed a panel of clinical factors and drug properties from 294 DILI cases collected by the International Drug-Induced Liver Injury Network Consortium (iDILIC). The model, which was built upon four factors significantly associated with prolonged resolution, including serum bilirubin, alkaline phosphatase (ALP) at DILI onset, time to onset and extent of drug metabolism, demonstrated robust differentiation between risk groups for their possibility to resolution within 6 months. This model successfully underwent external validation using 257 cases from the Spanish DILI Registry and 191 cases from the LiverTox database.

In a retrospective study with 168 DILI patients, Fu et al. [41] extracted over 1,600 radiomic features from liver segments and used the Lasso procedure to build a radiomics score. They developed a predictive model of chronicity using clinical characteristics and the radiomics score, which performed well in predicting chronic DILI.

In another single centre study, Zhao et al. [42] analysed the metabolomic profile of 90 hospitalised DILI patients, identifying higher levels of polyunsaturated fatty acid (PUFA) metabolites in chronic DILI. An enrichment analysis linked the upregulated PUFA metabolism-associated pathways to DILI chronicity. Using a random forest analysis, they created a predictive model with adrenic acid and aspartic acid, outperforming other chronicity markers, such as total and direct bilirubin, ALP or taurocholic acid (Table 2).

3.3 | Causality Assessment

In DILI, causality assessment systematically attributes liver injury to a drug after excluding other causes of liver injury and objectively weighting evidence to reach clinical judgement [2]. Several systems, including expert judgment, probabilistic methods and algorithms or scales, have been proposed to determine the relationship between drug intake and liver injury onset. This review will be focused on the latter two methods.

3.3.1 | Probabilistic Methods

The probabilistic approach typically uses Bayesian inference to estimate posterior probabilities from a prior distribution. The main component of Bayesian inference is Bayes' theorem, defined by the inversion of a conditional probability; that is, knowing the probability of the occurrence of an event (X) given an event (Y), the probability of the occurrence of event Y can be calculated given that event X has already occurred [43]. Interestingly, Bayesian inference is a learning technique, and it can be updated with the prior knowledge when new evidence is available, making it an essential tool in machine learning methodologies. Although less popular, some studies have proposed

TABLE 1 | (Continued)

Abbreviations: A-375, human skin melanoma; ALP, alkaline phosphatase; ALT, alanine aminotransferase; ANN, artificial neural network; AST, aspartate aminotransferase; AUROC, area under the Receiver Operating Characteristic (ROC) curve; BA, balance accuracy; DBL, direct bilirubin; DILIst, Drug-induced liver injury severity and toxicity; ECFP, Extended Connectivity Fingerprint; FAERS, Food and Drug Administration (FDA) Adverse Event Reporting System; GEO, Gene Expression Omnibus; HA1E, immortalised kidney cells; HepG2, liver carcinoma; INR, International Normalised Ratio; Lasso, least absolute shrinkage and selection operator; LR, logistic regression; LTKB, Liver Toxicity Knowledge Base; MACCS, Molecular ACCess System; MCC, Matthews correlation coefficient; MCF7, breast cancer; Mold^2, Molecular descriptors from 2D structures; NA, information not available; PC-3, adenocarcinoma; PHH, primary human hepatocytes; RDKFP, RDKit-specific fingerprint; RF, random forest; SVM, support vector machine; TBL, total bilirubin; Tox21, Toxicology in the 21st century; VPA, valproic acid.

[a]Top 10 molecular predictors used in the 2-class DILI prediction models. Model performance of 5-fold cross-validation on the 2-class DILI prediction model.

[b]Validation of deep learning model trained with 1574 genes.

[c]Area under the Receiver Operating Characteristic (ROC) curve of the best performing model: artificial neural network combining clinical and genetic data.

[d]Predictive model for practical use.

[e]Predictive model for severe DILI associated with liver necrosis. Area under the Receiver Operating Characteristic (ROC) curve of the training set.

[f]Top important features selected by extreme gradient boosting.

[g]Top 10 predictors for models built with the deep neural network algorithm using three molecular fingerprints are shown. The area under the Receiver Operating Characteristic (ROC) curve corresponds to the deep neural network algorithm.

TABLE 2 | Characteristics of studies that used artificial intelligence to establish a predictive prognostic model for drug-induced liver injury (sorted by date).

Study	Study population	DILI criteria	Outcome	Methodology	Predictive factors	Model performance	Validation cohort/set	Model performance (validation)
Ashby et al. (2021) [40]	294 DILI cases	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	Time course for recovery	Logistic regression	TBL, ALP, time to onset, drug metabolism	Probability of recovery 46% vs. 93% for the high and low risk groups, respectively	Yes	Significantly different time course for recovery
Zhao et al. (2022) [42]	90 DILI cases 70 healthy controls	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	Chronicity[a]	RF	Adrenic acid, aspartic acid	Sensitivity: 0.737 Specificity: 0.927 AUROC: 0.889	Yes	Sensitivity: 0.714 Specificity: 0.913 AUROC: 0.888
Fu et al. (2023) [41]	157 DILI cases	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	Chronicity[b]	Lasso	Cholestatic/ mixed liver injury, increased radiomics score	Accuracy: 0.872 Sensitivity: 0.697 Specificity: 0.947 AUROC: 0.89	Yes	Accuracy: 0.854 Sensitivity: 0.733 Specificity: 0.909 AUROC: 0.88
Niu et al. (2024) [36]	912 DILI cases	ALT ≥ 5 ULN; ALP ≥ 2 ULN; ALT ≥ 3 ULN and TBL > 2 ULN	Fatal outcome[c]	Bolasso	Prior drug allergies, hepatocellular damage, female sex, TBL, AST, platelet count	AUROC: 0.887	Yes	AUROC: 0.932

Abbreviations: ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUROC, area under the Receiver Operating Characteristic (ROC) curve; Bolasso, bootstrap-enhanced least absolute shrinkage operator; DILI, drug-induced liver injury; Lasso, least absolute shrinkage and selection operator; RUCAM, Roussel Uclaf Causality Assessment Method; RF, random forest; TBL, total bilirubin; ULN, upper limit of normal.

[a]Chronic DILI was defined as serum levels of aspartate aminotransferase (AST), alanine aminotransferase (ALT), or alkaline phosphatase (ALP) above baseline 6 months after DILI onset.

[b]Chronic DILI was defined as persistent abnormalities on liver tests, histological or radiological evidence of chronic liver injury over 6 months.

[c]Fatal outcome was defined as liver-related death or liver transplantation.

probabilistic models for assessing DILI causality. Zapater et al. [44] used data from the Canadian American Ticlopidine Study to estimate the odds of ticlopidine-induced liver injury and designed a Bayesian model calculating the likelihood of DILI due to ticlopidine. The model indicated that ticlopidine could be responsible for abnormal liver function tests in 61% of moderate–severe stroke patients treated with ticlopidine. Likewise, Llanos et al. [45] developed a Bayesian model for early causality assessment of amiodarone, finding that in 48% of cases with elevated liver function tests, amiodarone was responsible for liver damage. However, despite the advantage of learning from new evidence, Bayesian models may lose the detecting power when data is scarce.

3.3.2 | Algorithms and Scales

On the other hand, the use of scales and algorithms in causality assessment promotes objectivity in the evaluation of suspected cases, facilitates homogeneity of the diagnostic approach, and provides a probability category based on a numerical score. Among these, the Council for International Organisations of Medical Sciences/Roussel Uclaf Causality Assessment Method (CIOMS/RUCAM) is the most widely used and accepted tool [46]. However, it has drawbacks such as debatable domain weighting, lack of evidence on factors such as alcohol, pregnancy and age, and the inadequate scoring of atypical DILI phenotypes [47]. To address these, a computational tool named Revised Electronic Causality Assessment Method (RECAM, http://dilirecam.com) was developed using cases from the Drug-Induced Liver Injury Network (DILIN) and the Spanish DILI Registry [48]. Compared to RUCAM, RECAM excluded risk factors and concomitant drugs criteria and scored previous information on hepatotoxicity positively. RECAM, which utilises the information in LiverTox [49], showed a greater sensitivity for classifying higher likelihood categories and better overall agreement with expert opinion. However, its performance in less severe DILI cases is unclear [48]. Recently, RECAM was validated in China, showing a better performance than RUCAM for conventional drugs and herbs [50]. Moreover, in Japan, a modified version, RECAM-J 2023, showed fair performance for highly probable and probable cases and improved for possible cases [51].

DILI-CAT is another computer-assisted tool for causality assessment, particularly in drug development, by recognition of drug-specific phenotypes and using a data-driven algorithm [52]. It evaluates three criteria: latency (days between drug start and liver enzyme elevation), biochemical pattern of liver injury (*R*-value) and AST/ALT ratio at injury onset. Its scoring system assigns points based on parameter value within specific percentile ranges. Points are deducted for values outside the phenotype range or as an outlier. The causality assessment involves identification of a preliminary phenotype, followed by an assessment of the phenotype validity using additional cases. These steps are repeated with new cases throughout clinical development to refine and re-validate the phenotypes.

However, most of the abovementioned causality assessment scales and algorithms are derived from statistical methodologies based on clinical epidemiological information. Notably, the development of novel tools such as DeepCausality is of great interest by using emerging real-world data. This causal inference framework for free text-based documents, such as electronic health records and clinical reports, has proven its utility in ascertaining causality inference in DILI [53].

4 | The Application of Nature Language Processing and Large Language Models

The implementation of LLMs in various societal fields has garnered significant attention due to their extensive capabilities, as well as the challenges associated with their use. In the clinical setting, LLMs play crucial roles in data analysis and disease pattern recognition, assistance in data interpretation and enhancing patients' education for improved self-management. However, these tools present both advantages and disadvantages in the healthcare field.

LLMs can assist in interpreting laboratory results within the context of a patient's medical history, helping clinicians identify diagnostic patterns, analyse medical images, extract relevant diagnostic features, develop predictive models for risk stratification and identify new biomarkers. Conversely, limitations include reliance on training data, which can introduce bias and inaccuracies, lack of contextual understanding and interpretability of model outputs in specific medical scenarios, and ethical concerns related to patient privacy and data security [54, 55].

The performance of NLP and LLMs in clinical workflows remains contentious. One retrospective study compared the accuracy of identifying metabolic dysfunction-associated steatosis liver disease in electronic health records of over 38,000 patients using three approaches: NLP, International Classification of Diseases (ICD) codes and text search. The NLP tool demonstrated significantly better overall performance than ICD codes and text-based search approaches, although ICD showed better precision [56]. Similarly, Sherman et al. [57] developed an NLP algorithm to score liver biopsy findings from free text histopathology reports. The algorithm exhibited excellent agreement with manual validators, accurately identifying histological features such as ballooning, steatosis, lobular inflammation and fibrosis stages. Conversely, a study assessing LLM performance in a clinical decision-making for 2,400 real-world cases with four common abdominal pathologies (appendicitis, pancreatitis, cholecystitis and diverticulitis) found that LLMs performed significantly worse than clinicians on diagnostic accuracy, failing to follow diagnostic guidelines and treatment adherence [58].

Interestingly, the AI-based NLP models exhibit substantial performance in comprehending complex labeling texts, particularly in the context of drug safety evaluation. Wu et al. [59] created an AI-based model leveraging Bidirectional Encoder Representations from Transformers (BERT, a deep learning language model created by Google researchers) to classify DILI from FDA drug labeling text. Evaluated on 750 FDA labeling documents using cross-validation, this NLP-DILI model demonstrated robust performance with a MCC of 0.84 in DILI classification. External validation using European Medicines Agency (EMA) cross-agency data yielded a MCC of 0.79. In another study by Chen et al. [60], an automatic text classification approach based on a document-term matrix from text mining and

the XGBoost classifier also demonstrated robust performance in classifying DILI using drug labeling text.

However, the use of NLP for diagnostic purposes in DILI is currently limited, although some studies have utilised NLP to construct models for searching literature related to DILI. Rathee et al. [61] developed the $DILI_c$ model using a dataset containing DILI-positive and negative literature, along with external datasets from the FDA and the Side Effect Resource (SIDER). This AI-driven model uses NLP to extract relevant words, which are processed by a pattern mining algorithm. The mined patterns are scored, and the resulting scoring matrix is used by a gradient boosting machine to classify the literature.

5 | Future Perspectives

The interest in AI and its applications across various areas in medicine has been steadily growing in recent years. Idiosyncratic DILI is not an exception, as reflected by the increasing body of literature on AI subfields. This review summarised the findings on AI-based algorithms for the diagnosis or prognosis of DILI through clinical, in vivo or in silico studies.

The increasing use of AI-based models is driven, to some extent, by the advent of big data. This surge in the availability of large datasets, such as omics-derived or real-world data, facilitates the training and development of more complex models. Nonetheless, the relatively low incident condition of DILI and the lack of information may compromise the accuracy and validity of the models. Recently, a roadmap for DILI research in Europe advocated the creation of a prospective database integrating clinical information, linked omics data and liver histology data from DILI patients [62]. Within the operational framework, analysing data through AI methodologies presents a significant opportunity to advance DILI research.

Also, advances in NLP and LLM facilitate their progressive adoption in a clinical setting as supportive tools for clinicians. However, barriers such as a lack of knowledge and trust in AI systems may hinder this implementation. Initiatives like LiverAI, an AI-powered chatbot developed by the Spanish Association for the Study of the Liver to provide personalised support and education to its users [63], could support the integration of these tools, highlighting their advantages and limitations in aiding the diagnosis and risk stratification of DILI patients. Nonetheless, it is worth noting that clinicians should be cautious about AI-based models due to possible inaccuracies in the collection and processing of data, which may cause (inadvertent) bias and lead to erroneous conclusions. Therefore, while the adoption of AI tools could be deemed as an appealing opportunity, it is important to remark that AI models should not override clinical expertise.

Overall, the widespread use of AI in research represents a promising advancement in DILI research. Collaboration among stakeholders somehow involved in DILI (academia, clinicians, industry, regulatory authorities) is crucial to construct and validate complex models based on extensive data from diverse sources. These collaborative efforts will deepen our understanding of DILI and, ultimately, improve patient outcomes.

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

*References (64–72) are cited in Table S1.

1. R. J. Andrade, N. Chalasani, E. S. Björnsson, et al., "Drug-Induced Liver Injury," *Nature Reviews. Disease Primers* 5 (2019): 58.

2. European Association for the Study of the Liver, "EASL Clinical Practice Guidelines: Drug-Induced Liver Injury," *Journal of Hepatology* 70 (2019): 1222–1261.

3. A. Reuben, H. Tillman, R. J. Fontana, et al., "Outcomes in Adults With Acute Liver Failure Between 1998 and 2013: An Observational Cohort Study," *Annals of Internal Medicine* 164 (2016): 724–732.

4. M. Chen, H. Bisgin, L. Tong, et al., "Toward Predictive Models for Drug-Induced Liver Injury in Humans: Are We There Yet?," *Biomarkers in Medicine* 8 (2014): 201–213.

5. T. Kralj, K. L. R. Brouwer, and D. J. Creek, "Analytical and Omics-Based Advances in the Study of Drug-Induced Liver Injury," *Toxicological Sciences* 183 (2021): 1–13.

6. J. Lin, M. Li, W. Mak, et al., "Applications of In Silico Models to Predict Drug-Induced Liver Injury," *Toxics* 10 (2022): 788.

7. F. Liu and D. Panagiotakos, "Real-World Data: A Brief Review of the Methods, Applications, Challenges and Opportunities," *BMC Medical Research Methodology* 22 (2022): 287.

8. X. Shi, C. Zuo, L. Yu, et al., "Real-World Data of Tigecycline-Associated Drug-Induced Liver Injury Among Patients in China: A 3-Year Retrospective Study as Assessed by the Updated RUCAM," *Frontiers in Pharmacology* 12 (2021): 761167.

9. M. G. Kim, Y. J. Im, J. H. Lee, E. Y. Kim, S. W. Yeom, and J. S. Kim, "Comparison of Hepatotoxicity of Tegoprazan, a Novel Potassium-Competitive Acid Blocker, With Proton Pump Inhibitors Using Real-World Data: A Nationwide Cohort Study," *Frontiers in Medicine* 9 (2023): 1076356.

10. J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine* 27 (2006): 12–14.

11. J. M. Helm, A. M. Swiergosz, H. S. Haeberle, et al., "Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions," *Current Reviews in Musculoskeletal Medicine* 13 (2020): 69–76.

12. C. Janiesch, P. Zschech, and K. Heinrich, "Machine Learning and Deep Learning," *Electronic Markets* 31 (2021): 685–695.

13. F. Mostafa and M. Chen, "Computational Models for Predicting Liver Toxicity in the Deep Learning Era," *Frontiers in Toxicology* 5 (2024): 1340860.

14. A. Arjmand, C. T. Angelis, A. T. Tzallas, et al., "Deep Learning in Liver Biopsies Using Convolutional Neural Networks," 2019 42nd

International Conference on Telecommunications and Signal Processing (TSP): 496–499.

15. Y. Juhn and H. Liu, "Artificial Intelligence Approaches Using Natural Language Processing to Advance EHR-Based Clinical Research," *Journal of Allergy and Clinical Immunology* 145 (2020): 463–469.

16. K. S. Kalyan, "A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4," *Natural Language Processing Journal* 6 (2023): 100048.

17. J. Habicht, S. Viswanathan, B. Carrington, T. U. Hauser, R. Harper, and M. Rollwage, "Closing the Accessibility Gap to Mental Health Treatment With a Personalized Self-Referral Chatbot," *Nature Medicine* 30 (2024): 595–602.

18. S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, et al., "Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice," *BMC Medical Education* 23 (2023): 689.

19. H. Ma, S. Huang, F. Li, et al., "Development and Validation of an Automatic Machine Learning Model to Predict Abnormal Increase of Transaminase in Valproic Acid-Treated Epilepsy," *Archives of Toxicology* 98 (2024): 3049–3061.

20. S. M. Yu, H. C. Zheng, S. C. Wang, et al., "Salivary Metabolites Are Promising Noninvasive Biomarkers of Drug-Induced Liver Injury," *World Journal of Gastroenterology* 30 (2024): 2454–2466.

21. Y. Xiao, Y. Chen, R. Huang, F. Jiang, J. Zhou, and T. Yang, "Interpretable Machine Learning in Predicting Drug-Induced Liver Injury Among Tuberculosis Patients: Model Development and Validation Study," *BMC Medical Research Methodology* 24 (2024): 92.

22. A. Gonzalez-Jimenez, A. Suzuki, M. Chen, et al., "Drug Properties and Host Factors Contribute to Biochemical Presentation of Drug-Induced Liver Injury: A Prediction Model From a Machine Learning Approach," *Archives of Toxicology* 95 (2021): 1793–1803.

23. C. Feng, H. Chen, X. Yuan, et al., "Gene Expression Data Based Deep Learning Model for Accurate Prediction of Drug-Induced Liver Injury in Advance," *Journal of Chemical Information and Modeling* 59 (2019): 3240–3250.

24. K. Wang, L. Zhang, L. Li, et al., "Identification of Drug-Induced Liver Injury Biomarkers From Multiple Microarrays Based on Machine Learning and Bioinformatics Analysis," *International Journal of Molecular Sciences* 23 (2022): 11945.

25. E. Clough and T. Barrett, "The Gene Expression Omnibus Database," *Methods in Molecular Biology* 1418 (2016): 93–110.

26. N. H. Lai, W. C. Shen, C. N. Lee, et al., "Comparison of the Predictive Outcomes for Anti-Tuberculosis Drug-Induced Hepatotoxicity by Different Machine Learning Techniques," *Computer Methods and Programs in Biomedicine* 188 (2020): 105307.

27. H. Hong, S. Thakkar, M. Chen, and W. Tong, "Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using a Large Set of FDA-Approved Drugs," *Scientific Reports* 7 (2017): 17311.

28. M. Chen, A. Suzuki, S. Thakkar, K. Yu, C. Hu, and W. Tong, "DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans," *Drug Discovery Today* 21 (2016): 648–653.

29. H. Hong, Q. Xie, W. Ge, et al., "Mold(2), Molecular Descriptors From 2D Structures for Chemoinformatics and Toxicoinformatics," *Journal of Chemical Information and Modeling* 48 (2008): 1337–1344.

30. T. Li, W. Tong, R. Roberts, Z. Liu, and S. Thakkar, "DeepDILI: Deep Learning-Powered Drug-Induced Liver Injury Prediction Using Model-Level Representation," *Chemical Research in Toxicology* 34 (2021): 550–565.

31. S. Thakkar, T. Li, Z. Liu, L. Wu, R. Roberts, and W. Tong, "Drug-Induced Liver Injury Severity and Toxicity (DILIst): Binary Classification of 1279 Drugs by Human Hepatotoxicity," *Drug Discovery Today* 25 (2020): 201–208.

32. Q. Yang, S. Zhang, and Y. Li, "Deep Learning Algorithm Based on Molecular Fingerprint for Prediction of Drug-Induced Liver Injury," *Toxicology* 502 (2024): 153736.

33. M. Chen, J. Zhang, Y. Wang, et al., "The Liver Toxicity Knowledge Base: A Systems Approach to a Complex End Point," *Clinical Pharmacology and Therapeutics* 93 (2013): 409–412.

34. *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury* (National Institute of Diabetes and Digestive and Kidney Diseases, 2012).

35. T. Adeluwa, B. A. McGregor, K. Guo, and J. Hur, "Predicting Drug-Induced Liver Injury Using Machine Learning on a Diverse Set of Predictors," *Frontiers in Pharmacology* 12 (2021): 648805.

36. H. Niu, P. Solis-Muñoz, M. García-Cortés, et al., "Prior Drug Allergies Are Associated With Worse Outcome in Patients With Idiosyncratic Drug-Induced Liver Injury: A Machine Learning Approach for Risk Stratification," *Pharmacological Research* 199 (2024): 107030.

37. N. Chalasani, H. L. Bonkovsky, R. Fontana, et al., "Features and Outcomes of 899 Patients With Drug-Induced Liver Injury: The DILIN Prospective Study," *Gastroenterology* 148 (2015): 1340–1352.

38. I. Medina-Caliz, M. Robles-Diaz, B. Garcia-Muñoz, et al., "Definition and Risk Factors for Chronicity Following Acute Idiosyncratic Drug-Induced Liver Injury," *Journal of Hepatology* 65 (2016): 532–542.

39. F. Bessone, N. Hernandez, I. Medina-Caliz, et al., "Drug-Induced Liver Injury in Latin America: 10-Year Experience of the Latin American DILI (LATINDILI) Network," *Clinical Gastroenterology and Hepatology* 23 (2025): 89–102.

40. K. Ashby, W. Zhuang, A. González-Jimenez, et al., "Elevated Bilirubin, Alkaline Phosphatase at Onset, and Drug Metabolism Are Associated With Prolonged Recovery From DILI," *Journal of Hepatology* 75 (2021): 333–341.

41. H. Fu, Z. Shen, R. Lai, et al., "Clinic-Radiomics Model Using Liver Magnetic Resonance Imaging Helps Predict Chronicity of Drug-Induced Liver Injury," *Hepatology International* 17 (2023): 1626–1636.

42. S. Zhao, H. Fu, T. Zhou, et al., "Alteration of Bile Acids and Omega-6 PUFAs Are Correlated With the Progression and Prognosis of Drug-Induced Liver Injury," *Frontiers in Immunology* 13 (2022): 772368.

43. M. A. Martínez-González, A. Sánchez-Villegas, E. A. Toledo Atucha, and J. Faulin Fajardo, *Bioestadistica amigable*, 3rd ed. (Elsevier España, 2014), 72.

44. P. Zapater, J. Such, M. Pérez-Mateo, and J. F. Horga, "A New Poisson and Bayesian-Based Method to Assign Risk and Causality in Patients With Suspected Hepatic Adverse Drug Reactions: A Report of Two New Cases of Ticlopidine-Induced Hepatotoxicity," *Drug Safety* 25 (2002): 735–750.

45. L. Llanos, R. Moreu, A. M. Peiró, et al., "Causality Assessment of Liver Injury After Chronic Oral Amiodarone Intake," *Pharmacoepidemiology and Drug Safety* 18 (2009): 291–300.

46. G. Danan and C. Benichou, "Causality Assessment of Adverse Reactions to Drugs—I. A Novel Method Based on the Conclusions of International Consensus Meetings: Application to Drug-Induced Liver Injuries," *Journal of Clinical Epidemiology* 46 (1993): 1323–1330.

47. M. García-Cortés, C. Stephens, M. I. Lucena, A. Fernández-Castañer, and R. J. Andrade, "Causality Assessment Methods in Drug Induced Liver Injury: Strengths and Weaknesses," *Journal of Hepatology* 55 (2011): 683–691.

48. P. H. Hayashi, M. I. Lucena, R. J. Fontana, et al., "A Revised Electronic Version of RUCAM for the Diagnosis of DILI," *Hepatology* 76 (2022): 18–31.

49. J. H. Hoofnagle, J. Serrano, J. E. Knoben, and V. J. Navarro, "LiverTox: A Website on Drug-Induced Liver Injury," *Hepatology* 57 (2013): 873–874.

50. X. Zhao, Y. Wang, R. Lai, et al., "Validation of the Revised Electronic Version of RUCAM for Diagnosis of DILI in Chinese Patients," *Hepatol Commun* 8 (2024): e0235.

51. A. Tanaka, K. Tsuji, Y. Komiyama, et al., "RECAM-J 2023-Validation and Development of the Japanese Version of RECAM for the Diagnosis of Drug-Induced Liver Injury," *Hepatology Research* 54 (2024): 503–512.

52. R. P. Hermann, D. C. Rockey, A. Suzuki, M. Merz, and H. L. Tillmann, "A Novel Phenotype-Based Drug-Induced Liver Injury Causality Assessment Tool (DILI-CAT) Allows for Signal Confirmation in Early Drug Development," *Alimentary Pharmacology & Therapeutics* 55 (2022): 1028–1037.

53. X. Wang, X. Xu, W. Tong, Q. Liu, and Z. Liu, "DeepCausality: A General AI-Powered Causal Inference Framework for Free Text: A Case Study of LiverTox," *Frontiers in Artificial Intelligence* 5 (2022): 999289.

54. R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large Language Models in Health Care: Development, Applications, and Challenges," *Health Care Science* 2 (2023): 255–263.

55. E. Ullah, A. Parwani, M. M. Baig, and R. Singh, "Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine With a Focus on Digital Pathology - a Recent Scoping Review," *Diagnostic Pathology* 19 (2024): 43.

56. T. T. Van Vleck, L. Chan, S. G. Coca, et al., "Augmented Intelligence With Natural Language Processing Applied to Electronic Health Records for Identifying Patients With Non-alcoholic Fatty Liver Disease at Risk for Disease Progression," *International Journal of Medical Informatics* 129 (2019): 334–341.

57. M. S. Sherman, P. K. Challa, E. M. Przybyszewski, et al., "A Natural Language Processing Algorithm Accurately Classifies Steatotic Liver Disease Pathology to Estimate the Risk of Cirrhosis," *Hepatology Communications* 8 (2024): e0403.

58. P. Hager, F. Jungmann, R. Holland, et al., "Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making," *Nature Medicine* 30 (2024): 2613–2622.

59. Y. Wu, Z. Liu, L. Wu, M. Chen, and W. Tong, "BERT-Based Natural Language Processing of Drug Labeling Documents: A Case Study for Classifying Drug-Induced Liver Injury Risk," *Frontiers in Artificial Intelligence* 4 (2021): 729834.

60. M. Chen, Y. Wu, B. Wingerd, et al., "Automatic Text Classification of Drug-Induced Liver Injury Using Document-Term Matrix and XGBoost," *Frontiers in Artificial Intelligence* 7 (2024): 1401810.

61. S. Rathee, M. MacMahon, A. Liu, et al., "DILIc: An AI-Based Classifier to Search for Drug-Induced Liver Injury Literature," *Frontiers in Genetics* 13 (2022): 867946.

62. M. I. Lucena, M. Villanueva-Paz, I. Alvarez-Alvarez, et al., "Roadmap to DILI Research in Europe. A Proposal From COST Action ProEuroDILINet," *Pharmacological Research* 200 (2024): 107046.

63. D. Marti-Aguado, J. Pazó, A. Diaz-Gonzalez, et al., "LiverAI: New Tool in the Landscape for Liver Health," *Gastroenterología y Hepatología* 47 (2024): 646–648.

64. J. Neuhaus and C. McCulloch, "Generalized Linear Models," *Wiley Interdisciplinary Reviews: Computational Statistics* 3 (2011): 407–413.

65. L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, "A Critical Review of LASSO and Its Derivatives for Variable Selection Under Dependence Among Covariates," *International Statistical Review* 90 (2022): 118–145.

66. L. Breiman, "Random forests," *Machine Learning* 45 (2001): 5–32.

67. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Machine Learning* 63 (2006): 3–42.

68. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," 2016 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794.

69. S. Suthaharan, "Machine Learning Models and Algorithms for Big Data Classification," in *Support Vector Machine* (Boston: Springer, 2016), 207.

70. E. Guresen and G. Kayakutlu, "Definition of Artificial Neural Networks With Comparison to Other Networks," *Procedia Computer Science* 3 (2011): 426–433.

71. T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep Neural Network Concepts for Background Subtraction: A Systematic Review and Comparative Evaluation," *Neural Networks* 117 (2019): 8–66.

72. C. Pittinger and A. Mohapatra, "Chapter 69–Software Tools for Toxicology and Risk Assessment," in *Information Resources in Toxicology (Fourth Edition)* (Academic Press, 2009), 631–638.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.