January 2021

Using Machine Learning On Diverse Datasets To Predict Drug-Induced Liver Injury

Temidayo Adeluwa

USING MACHINE LEARNING ON DIVERSE DATASETS TO PREDICT DRUG-INDUCED LIVER INJURY

by

Temidayo Peter Adeluwa

Bachelor of Science, University of Lagos, Nigeria (2016)

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillments of the requirements

for the degree of

Master of Science

Grand Forks, North Dakota

August 2021

Name: Temidayo Adeluwa

Degree: Master of Science

This document, submitted in partial fulfillment of the requirements for the degree from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

DocuSigned by:
Junguk Hur
874412A5CDE74BB...
Junguk Hur

DocuSigned by:
Marina E. Kim
ECEA0B2E147147E...
Marina Kim

DocuSigned by:
Motoki Takaku
EEA02946E2E4415...
Motoki Takaku

This document is being submitted by the appointed advisory committee as having met all the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

DocuSigned by:
Chris Nelson
2E0AF088C733403...
Chris Nelson
Dean of the School of Graduate Studies

7/26/2021

Date

PERMISSION

Title:	Using Machine Learning on Diverse Datasets to Predict Drug-Induced Liver Injury
Department:	Biomedical Sciences
Degree:	Master of Science

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

Temidayo Peter Adeluwa

August, 2021

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF EQUATIONS

Equation Page

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the members of my advisory Committee for their guidance and support during my time in the master's program at the University of North Dakota. I would also like to thank my advisor, Dr. Junguk Hur, for his many advice and invaluable mentorship.

DEDICATION

To my family, for being my beacon of light and my motivation

To Eniola Ajide and his family, for being my pillar of support and encouragement

To my friends at home and abroad, Lanre, Wale, and Chinyere, and everyone who has supported me on this journey

To Dr. Junguk Hur, for his mentorship, giving me opportunities to learn, and encouraging me to take bold steps

To myself, for taking bold steps

ABSTRACT

A major challenge in drug development is safety and toxicity concerns due to drug side effects. One such side effect, drug-induced liver injury (DILI), is considered a primary factor in regulatory clearance. To develop prediction models of DILI, the Critical Assessment of Massive Data Analysis (CAMDA) 2020 CMap Drug Safety Challenge goal was established with an ultimate goal to develop prediction models based on gene perturbation of six preselected cell-lines (CMap L1000), extended structural information (MOLD2), toxicity data (TOX21), and FDA reporting of adverse events (FAERS). Four types of DILI classes were targeted, including two clinically relevant scores and two control classifications, designed by the CAMDA organizers. The L1000 gene expression data had variable drug coverage across cell lines with only 247 out of 617 drugs in the study measured in all six cell types. We addressed this coverage issue by using Kru-Bor ranked merging to generate a singular drug expression signature across all six cell lines. These merged signatures were then narrowed down to the top and bottom 100, 250, 500, or 1,000 genes most perturbed by drug treatment. These signatures were subject to feature selection using Fisher’s exact test to identify genes predictive of DILI status. Models based solely on expression signatures had varying results for clinical DILI subtypes with an accuracy ranging from 0.49 to 0.67 and Matthews Correlation Coefficient (MCC) values ranging from -0.03 to 0.1. Models built using FAERS, MOLD2, and TOX21 also had similar results in predicting clinical DILI scores with accuracy ranging from 0.56 to 0.67 with MCC scores ranging from 0.12 to 0.36. To incorporate these various data types with expression-based models, we utilized soft, hard, and weighted ensemble voting methods using the top three performing models for each DILI classification. These voting models achieved a balanced accuracy up to 0.54 and 0.60 for the clinically relevant DILI subtypes. Overall, from our experiment, traditional machine learning approaches may not be optimal as a classification method for the current data.

CHAPTER 1

INTRODUCTION

Adverse drug reactions (ADRs) are a common concern of novel drugs and therapeutics. One of the more common targets of ADRs is the liver due to its role in the metabolism of compounds and resulting liver damage is termed as Drug-Induced Liver Injury (DILI)[1–3]. DILI is a unique challenge in drug development due to the inability of animal models to translate to human clinical trials in treatment populations. Assessing DILI risk has been approached in multiple ways during drug development; however, officials often rely on post-marketing surveillance to detect possible long-term side effects such as DILI[4]. The U.S. Food and Drug Administration (FDA) has established the DILIrank dataset, the largest reference drug list ranked for DILI risk in humans, to facilitate the development of predictive models by enhancing drug label DILI annotation with weighted causal evidence[5]. This dataset contains four classifications, including most, less, ambiguous, and no-DILI concern, regarding 1,036 FDA-approved drugs. Additionally, predicting DILI is difficult due to the absence of specific and reliable biomarkers. Traditional biomarkers, including alanine aminotransferase, total bilirubin levels, aspartate aminotransferase, and gamma-glutamyl transferase (among others) are not specific enough to separate DILI from other forms of liver injury[6]. Due to this reason, FDA in 2016 approved investigations into glutamate dehydrogenase and microRNA-122 as potential biomarkers[7]. Messner and colleagues characterized exosomal microRNA-122 in methotrexate and acetaminophen-induced toxicity in hepatic stem cells, HepaRG. They confirmed that microRNA-122 can be used as a sensitive biomarker for DILI[8].

Predictive markers of DILI, determined by compound properties and known variables rather than preclinical studies, would facilitate drug development in a wide variety of ways[9,10]. Multiple groups have attempted to predict DILI using drug compounds or proposed drug properties. Chemical structures[11], gene expression response[12], and patient genetic data have been previously used for DILI prediction using traditional machine learning algorithms. Xu et al. proposed a deep learning model built on a “combined data set” gathered from a variety of sources and used a molecular structural encoding approach for the chemical structures of the drugs in their data[13]. Kohonen et al. proposed a ’big data compacting and data fusion’ concept[14]. In their approach, the authors utilized data from the Connectivity Map (CMap; Broad Institute) database, the Open Toxicogenomics Project-Genomics Assisted Toxicity Evaluation Systems (TG-GATEs; National Institutes of Biomedical Innovation, Japan), the US National Cancer Institute 60 tumor cell line screening (NCI-60), and the US FDA Liver Toxicity Knowledge Base (LTKB). Using these databases, they modeled a predictive toxicogenomics space that captured all possible well-known hepato-pathological changes[14].

Building upon these previous efforts to accurately predict DILI, the Critical Assessment of Massive Data Analysis (CAMDA) in collaboration with the Intelligent Systems for Molecular Biology (ISMB) has proposed the CMap Drug Safety Challenge for their annual conferences in 2018, 2019, and 2020 (Table 1). The previous challenges in 2018 and 2019, while sharing a similar goal to predict potential liver toxicity, also had distinct parameters. The prediction DILI classification in 2018 was a binary positive or negative

Table 1. Previous CAMDA Drug Safety Challenge Summary.
The CMap Drug Safety Challenge has been a repeated effort by CAMDA to develop predictive models for DILI. Previous studies are cited by their year of publication and leading author while also describing the the year in which the challenge was administered by CAMDA and relevant data sources and DILI classifications for prediction.

Authors	CAMDA Drug Safety Challenge	Data Sources	DILI conditions
Current: Adeluwa et al.	2020	CMap L1000, MOLD2, FAERS, TOX21	DILI1, DILI3, DILI5, DILI6
2021: Liu et al.	2019	CMap L1000, SMILES strings, SIDER 4.1,	Most-DILI concern, Less-DILI concern, Ambiguous DILI concern, No-DILI concern
2021: Aguirre-Plans et al.	2019	CMap L1000, DisGeNET, GUILDify, SMILES, DGldb, HitPick, SEA	
2021: Lesinski et al.	2019	CMap L1000, SMILES, annotated Images	
2020: Chierici et al.	2018	Affymetrix GeneChip (MCF7, PC3)	DILI-1, DILI-0
2020: Sumsion et al.	2018	Affymetrix GeneChip (MCF7, PC3)	

DILI status, while in 2019 the challenge was more focused on the potential DILI risk ranging from no concern to most concern with four classifications reflecting the DILIrank dataset[5]. The data, used for predicting the DILI classification of drugs in the 2018 challenge, were limited to microarray data from MCF7 and PC3 cell lines. Chierici et al. in 2018 employed deep learning techniques for the microarray data from 276 compounds but only achieved Matthews Correlation Coefficient (MCC) values of <0.2[15]. Sumsion et al. in the same challenge year utilized more traditional classification algorithms along with soft voting but reached a maximum MCC of 0.2 and maximum accuracy of 70%, while the voting model never performed the best when compared to individual models[16]. Both studies cite struggles with the small sample size and imbalanced datasets; however, resampling, in this case, led to overfitting rather than improved testing accuracy.

The CMap Drug Safety Challenge expanded in 2019 by including not only expression data from L1000 CMap but also by allowing a wide variety of external data sources that were incorporated into each study. Lesinski et al. achieved their best predictive results by incorporating molecular drug properties along with the most informative variables from 5 of 13 cell line expression models via a super learner method[17]. Including molecular property information improved their cell line models’ accuracy up to 73% utilizing a random forest algorithm, which originally ranged from 55% to 61%. Liu et al. built support vector machine and random forest models using chemical descriptions from DILIrank annotation along with expression values from predicted protein targets[18]. This approach produced models with an accuracy of 75.9% that were also able to correctly identify targets associated with the mechanism of action and toxicity of nonsteroidal anti-inflammatory drugs, a class of drugs commonly associated with DILI. Aguirre et al.

utilized the widest array of predictive data, including L1000 CMap expression, drug-target associations, structural data, phenotype-associated gene signatures, protein-protein interactions, and drug targets data[19]. Their models' accuracy remained comparable to other study results at 70%, but they also identified structural dissimilarities within the DILI risk labels used. All three published studies from the 2019 CMap drug safety challenge cited data limitations within their study, including complex dosage-related toxicity, a small sample size, and a small number of compounds with hepatoxicity annotation.

The current CAMDA 2020 challenge was structured in a way to address the previous limitations, while also redefining the relevant DILI classifications. The challenge aimed to predict or classify positive and negative classes within each of four DILI designations, namely DILI1, DILI3, DILI5, and DILI6. DILI1 and DILI3 were clinical classifications based on specific severity scores or established FDA warnings and precautions, while DILI5 and DILI6 served as a negative and positive control class, respectively (Table 2). Drug class labels were assigned by the CAMDA 2020 challenge organizers. DILI1 was described as a severity score ≥ 6 which is associated with high risk based on the DILIrank dataset and LTKB[20]. DILI3 was described as drugs withdrawn, given boxed warnings, or warnings and precautions from the FDA due to either known risk factors or adverse event reporting. DILI5 served as a randomly assigned negative control, while DILI6 was constructed as a positive control based on molecular weight with positive compounds weighing >320 g/mol. The drug list for the study was expanded to 617 drug compounds to improve on the sample size limitations of previous studies; however, these datasets remained highly imbalanced.

Table 2. Drug-Induced Liver Injury Classifications.

Four binary classes of DILI were provided by the CAMDA organizers. DILI1 positive compounds were based on the clinical severity score associated with liver necrosis. DILI3 positive compounds were based on drug already associated with warnings and precautions or that have been withdrawn due to liver toxicity. DILI5 was a random assignment from the organizers as a negative control group while the DILI6 classification was based on molecular weight (>320 g/mol) to serve as a positive control.

Targets	**Positive group**	**Negative group**
DILI1	DILI Severity score ≥6 (N=141)	DILI Severity score <6 (N=476)
DILI3	Withdrawn, box warning, warning & precaution (N=227)	Adverse events and no match (N=390)
DILI5	Assigned DILI endpoint 1 (N=308 positive)	(N=309 negative)
DILI6	Assigned DILI endpoint 2 (N=318 positive)	(N=299 negative)
Note1: *DILI5/DILI6 are controls; DILI5 is randomly split; DILI6 is the positive control, dividing compounds based on their molecular weight >320 g/mol*		

The imbalance within the clinically relevant DILI data is expected considering that many approved drugs do not have a significant hepatoxicity risk; however, the control classes of DILI5 and DILI6 were structured in a balanced manner (Table 3). For this challenge, L1000 drug expression signatures from primary human hepatocytes (PHH), liver carcinoma (HepG2), immortalized kidney cells (HA1E), human skin melanoma (A-375), breast cancer (MCF7), and adenocarcinoma (PC-3) were used as inferred from landmark genes defined by Connectivity Map[21]. These expression responses were simplified to one specific dose at one specific treatment time in order to yield the largest available dataset for training and testing while also addressing previous dosage toxicity concerns. Other non-gene expression data provided included molecular descriptors encoding two-dimensional chemical structure information from MOLD2[22], post-marketing drug adverse event information from FAERS[23], and high-throughput liver toxicity screening results from TOX21[24]. While previous studies also utilized external data sources to improve model performance, the current study focuses on the various types of data processed and provided from the CMap drug safety challenge.

We constructed models to predict each drug's DILI class (positive or negative) within the four DILI classifications (DILI1, DILI3, DILI5, and DILI6) by first evaluating the performance of each dataset in predicting DILI and also by employing ensemble voting with the top three performing models across data types. The gene expression data presented a unique challenge in that not all drugs were tested in each cell line or even in liver-relevant cell lines. To address this, we utilized a Kru-Bor merging method to merge the expression signatures across cell lines into one representative drug signature[25,26].

Table 3. Training Data Imbalance.
The data used for the clinical DILI classes of DILI1 and DILI3 were imbalanced which negatively influenced the models built to predict these classes.

DILI Class	**Negative**	**Positive**
DILI1	326	96
DILI3	262	160
DILI5	218	204
DILI6	197	225

These expression signatures were narrowed down to the top and bottom 100, 250, 500, and 1,000 ranked genes and subjected to feature selection via a Fisher's exact test based on their involvement in DILI positive/negative assigned drugs for each DILI class. FAERS, MOLD2, and TOX21 datasets were also used to construct DILI predictive models, and to address the imbalance of these data we tested resampling techniques. Various traditional classifier algorithms were used to build models on these datasets, and the models were evaluated on a blinded test set by the CAMDA committee. Based on the training area under the curve (AUC) values of these models, the top three algorithms for each datatype (cell expression, FAERS, MOLD2, and TOX21) for each DILI class were included in our ensemble voting model. We tested hard, soft, and weighted voting across these datasets to see if the varying dimensions of data can improve predictive performance.

CHAPTER 2

LITERATURE REVIEW

The purpose of this section is to describe the pathophysiology of DILI and to review the available literature on the status quo of DILI prediction.

Pathogenesis and Mechanisms of DILI

DILI occurs in the liver because the liver is an important site for the metabolism of compounds. Metabolism aims to transform lipid-soluble compounds – which are biologically active – into lipid-insoluble compounds that are easily excreted from the body. Lipophilic drugs are bioactive because they can easily cross the membrane barriers of cells. Usually, active drug compounds are metabolized into inactive forms at which stage they do not interfere with biological processes in the body[27]. In other cases, metabolism converts inactive drugs (prodrug) into active metabolites that can interfere with biological processes in the body[27]. It is also possible for an active metabolite to be converted into many other active metabolites[27].

These conversions are mediated by chemical reactions that take place in the liver. For nomenclature, these reactions are split into Phase I and II reactions but they don't have to take place in that order. Phase I reactions are mediated by the cytochrome P450 (CYP) superfamily of enzymes[27]. These enzymes modify drugs into lipophobic drugs i.e. water-soluble drugs. The reactions they catalyze include oxidation, reduction, hydroxylation, deamination, sulphoxidation, and various forms of dealkylation. Reactive metabolites that are potentially – and directly – toxic to the cells are generated in this phase.

In phase II reactions, reduced or oxidized forms of drug compounds are conjugated through various methods including acetylation by way of N-acetyltransferases (NATs), glucuronide conjugation by way of UDP-glucuronosyltransferases (UGTs), methylation through thiopurine S-methyltransferases (TPMTs) and/or catechol O-methyltransferases (COMTs), addition of glutathione substrates through glutathione S-transferases (GSTs), and sulfation by sulfotransferases (SULTs). These conjugation processes make it possible for metabolites to be effluxed through transporters[27].

The amount of metabolites and reactive oxygen species (ROS) that the liver is exposed to make it a potential site of damage. This is one proposed mechanism of the DILI formation. These metabolites can interfere with the structure of proteins – and consequently, their functions and localization – by covalent bonding [to these proteins][28]. This direct mechanism of DILI formation can result in hepatocellular damage and death through endoplasmic reticulum (ER) stress, mitochondrial dysfunction, and interference with signaling pathways[29]. Intracellular calcium signaling and composition can be interfered with resulting in lysis of the cells[30]. While this direct mechanism of DILI formation is easily understood, it is insufficient in explaining the involvement of the immune system. Thus, the recruitment of other cells through the immune system is proposed as another important mechanism of DILI formation.

The stress and damage caused by drugs can trigger inflammatory reactions of the innate immune system, through the release of damage-associated molecular patterns (DAMPs) like adenosine triphosphate (ATP), heparin sulfate, DNA, heat-shock proteins, and high mobility group box 1 protein (HMGB1)[31]. These DAMPs can lead to the production of reactive oxygen/nitrogen species, neutrophil inflammation, and an increase in cytokines and inflammatory chemokines through the activation of Toll-like receptors[32]. Also, DAMPs can be recognized by pattern recognition receptors (PRRs) that are present on antigen-presenting cells (APCs), leading to the activation of the adaptive immune system.

A more popular hypothesis for DILI formation that involves the immune system is the hapten hypothesis[29,33]. Here, it is proposed that DILI is caused by haptens which are small molecules that become immunogenic when they are bonded with carrier molecules like proteins[9]. The formed hapten-protein adducts activate the innate immune system, which leads to the production of inflammatory chemokines and cytokines. Other inflammatory mediators like Fas and IFN-gamma (interferon-gamma) can cause direct liver damage[34,35]. In turn, the innate immune system activates the cells of the adaptive immune system through T cell responses. The manner through which these haptens are presented to T cells is dependent on the HLA (human leukocyte antigens) haplotype of the individual, in turn determining the immune response, further explaining a genetic basis for idiosyncratic DILI[34,35].

As the principal system for exporting bile salts outside the liver, the blockade of the bile salt export pump (BSEP) has been hypothesized as another DILI mechanism. One study found that a genetic loss-of-function deficiency of the BSEP system led to liver failure and cholestasis[36]. By blocking the BSEP, there is an increased concentration of bile acids within the liver, which can lead to hepatocellular stress, mitochondrial dysfunction, ER and organellar stress[37,38]. Medications like troglitazone, sunitinib, bosentan, and cyclosporine A have been implicated in inhibiting BSEPs and causing DILI[37,39–43].

Types of DILI

Classically, the pathogenesis of DILI has determined its classifications such that DILI is divided into two types, namely intrinsic (or direct) DILI and idiosyncratic DILI[44]. Some drugs can cause direct liver toxicity when used at high doses beyond their therapeutic indices. In this case, the DILI type is said to be intrinsic, and it is predictable and dose-related[44]. In many cases, intrinsic DILI occurs after a short period of exposure to the medication (at doses beyond the recommended dose). Acetaminophen (Tylenol) is a commonly used non-steroidal anti-inflammatory drug (NSAID) for treating fever and pain that is well-characterized for causing acute liver failure slightly beyond the maximum recommended dose of 4g per day[45–47]. The acute liver failure in these patients was marked by elevated alanine aminotransferase (ALT)[48] and aspartate aminotransferase (AST)[46]. Interestingly, acetaminophen is responsible for most cases of acute liver failure[48].

Nearly all presentations of DILI in the clinic are idiosyncratic (or unpredictable) DILI, and patient-dependent. This type of DILI is defined by having no direct liver toxicity, dose-independent (thus, occurring even at minimum/recommended doses), unpredictable, severe (or fatal), and rarely-occurring[49–52]. One of the commonest histological phenotypes of idiosyncratic DILI is acute hepatitis[53], marked by increased alanine aminotransferase concentrations. Responsible for close to 15% of acute liver failure due to idiosyncratic DILI[54,55], acute hepatitis is caused by medications such as diclofenac, nitrofurantoin, isoniazid, sulfonamides, and floroquinolones[56,57]. Another phenotypic representation of DILI is cholestatic hepatitis defined by an impediment to the flow of bile from the liver. Symptoms include pruritus, jaundice, dark urine, nausea, and rash[58]. Serum biomarkers alkaline phosphatase (ALP) and bilirubin are significantly increased[58,59], and medications like chlorpromazine, amoxicillin-clavulanate, cefazolin, azathioprine, ciprofloxacin, levofloxacin, cephalosporins, and terbinafine[60–65]. In many cases, if these medications are withdrawn quickly, cholestatic hepatitis usually resolves by itself[64]. Besides acute hepatitis and cholestatic hepatitis, other phenotypic representations of idiosyncratic DILI are chronic hepatitis and mixed hepatitis.

Diagnosis of DILI

Taken together, acute and chronic hepatitis, cholestatic hepatitis, and acute and chronic cholestasis are the most common phenotypic representations of DILI. However, these histological patterns are not perfectly correlated with serum biomarkers and biochemical presentations of DILI. This non-correlation, coupled with the inability to differentiate DILI from liver disease not due to medications or supplements, makes the diagnosis of DILI difficult. For instance, drug-induced acute hepatitis shares strikingly similar symptoms with acute viral hepatitis, even with an increase in ALT concentration[66]. Symptoms of DILI are similar to those found in autoimmune hepatitis fatty liver disease and hepatic necrosis[66]. Therefore, DILI is usually diagnosed based on the exclusion of seemingly related liver diseases not due to the use of medications. Currently, there are no specific biomarkers for DILI diagnosis, but measurement of these serum biomarkers ALT, AST, ALP, and bilirubin are being used as diagnostic parameters[67]. In addition to this is gamma-glutamyl transferase (GGT)[67]. These markers, used alongside Hy's law, are a tool approved for the determination of a medication's ability to cause DILI.

Hy's law was proposed by Hy Zimmerman after certain clinical observations[68]. He proposed that there is a 10% to 50% chance of mortality in patients with evidence of jaundice and hepatocellular damage[68–71]. Over time, the FDA has expanded and compiled these criteria such that a drug is determined to cause DILI if it meets the following conditions:

1. 3-fold elevation of aminotransferases (ALT or AST) above the normal upper limit
2. Alongside the previous criterion, a 2-fold increase in total bilirubin levels above the normal upper limit, barring the non-diagnosis of cholestasis
3. If there is no other diagnosis explaining the hepatocellular damage, for instance, acute viral hepatitis, congestive heart failure,

In addition to the above criteria, a separate expert panel[72] recommended the following:

1. A 5-fold elevation of aminotransferases (ALT or AST) above the normal upper limit
2. A 2-fold elevation of ALP above the normal upper limit
3. Alongside the first criterion, a 2-fold increase in total bilirubin above the normal upper limit.

Regardless of these criteria and biomarkers, determining DILI is still difficult primarily because separating drug-induced hepatocellular damage from non-drug-induced hepatocellular damage is challenging. In addition, determining causality is confounded by the use of multiple medications and insufficient information on the doses and usage of the medication(s). A pertinent challenge is these traditional biomarkers are not liver-specific, neither are they drug-specific, necessitating the need for more specific biomarkers[73]. The FDA has launched investigations into finding new and specific biomarkers including glutamate dehydrogenase (GLDH) and miRNA-122[74]. Other potential biomarkers are the histological biomarker HMGB1, macrophage colony-stimulating factor receptor, and keratin-18[75].

To address the diagnostic challenge with DILI, researchers have turned to computational approaches, in particular, machine learning[76–79]. Given the exponential growth in next-generation sequencing technologies, large biological datasets, faster computers, and more efficient computational tools, data science-driven methods to understand patterns in DILI progression are an invaluable approach to this problem.

Computational prediction of DILI has long relied on using molecular/chemical/structural information of drugs. The hypothesis is that the structure of drugs harbor information that determines how they are metabolized and that the resulting metabolites can point to DILI development. Using 3D molecular descriptors as inputs to a linear discriminant analysis and an artificial neural network algorithm, Cruz-Monteagudo and others build predictive models on 74 drugs and achieved 82% accuracy on 13 drugs used as the test data[80]. Tropsha's group developed a quantitative structure-activity relationship model using a k-nearest neighbor algorithm built on 200 molecular descriptors and tested on 37 drugs, achieving as high as 73% accuracy on the test set[28]. A challenge with these studies was that the models were never tested on large external datasets. Liu and colleagues used the chemical structure information of drugs in a CAMDA 2019 challenge to build predictive models of DILI. Using a support vector machine and a random forest algorithm, they achieved a mean balanced accuracy of 0.759 on an external test set[18]. In the same study, the authors used L1000 gene expression data for the drugs but noted that these datasets were not predictive of DILI[18].

To improve the prediction of DILI, researchers have turned to using – and integrating – diverse datasets including genetic data and toxicity information. Furlanello's group used gene expression information of two cancer cell lines treated with 276 drugs to build binary classification models. They developed a random forest model, a single-layer neural network model, and three deep learning models but obtained poor performance. Lesinski et al., in a study published in 2021 integrated gene expression data and chemical properties where their best model achieved an AUC of 0.73.

Another approach to developing classifiers for DILI is integrating available datasets, especially when individual datasets are not predictive. Piccolo's group attempted to integrate the strength of different models by aggregating many models in an ensemble approach, alongside alternative methods like class-weighting and dimensionality reduction. Regardless, their approach failed to generalize properly to the test set[16]. Voting approaches, however, have the potential to improve the prediction metrics that they are measured on[81,82].

Problem Statement

The current status of DILI prediction has room for improvement. In our approach, we had access to gene expression data across six cell lines, molecular descriptors for the drugs, toxicity information for the drugs, and patient-reported incidences of adverse drug reactions for each drug. The aim of this study was to (i) evaluate the quality of these individual datasets in predicting DILI, and (ii) to evaluate if integrating these dataset can improve DILI prediction. To this end, we developed a voting method to aggregate the strengths of models built in (i).

CHAPTER 3

METHODOLOGY

Data Processing

The overall workflow of our study is shown in Figure 1. Initially, the overlap of drugs, included in each of the gene expression cell data sets, was investigated using VennDetail [83] to create a Venn pie chart showing the various drug testing subsets across the six cell lines (Figure 2). Each of the non-gene expression datasets (FAERS, MOLD2, and TOX21) were treated as individual datasets, while the gene expression data were merged across cell lines to build classifier models. In general, we used standard preprocessing techniques, including removing zero variance features and missing values. DILI1 and DILI3 suffered from class imbalance (Table 3). For all non-gene expression data, to mitigate this issue, we attempted three oversampling techniques, including synthetic minority oversampling technique (SMOTE) [84], random oversampling examples (ROSE) [85], and a random upsampling of the minority classes. SMOTE balances data by randomly creating artificial samples between two nearest-neighbor samples, while ROSE uses a smoothed bootstrap technique to resampled the data [84,85]. For comparison, models were built using imbalanced data as well. Before training non-gene expression datasets, they were standardized to have a mean of 0 and a standard deviation of 1. Preprocessing details specific to each dataset as well as some characteristics of the data are discussed below.

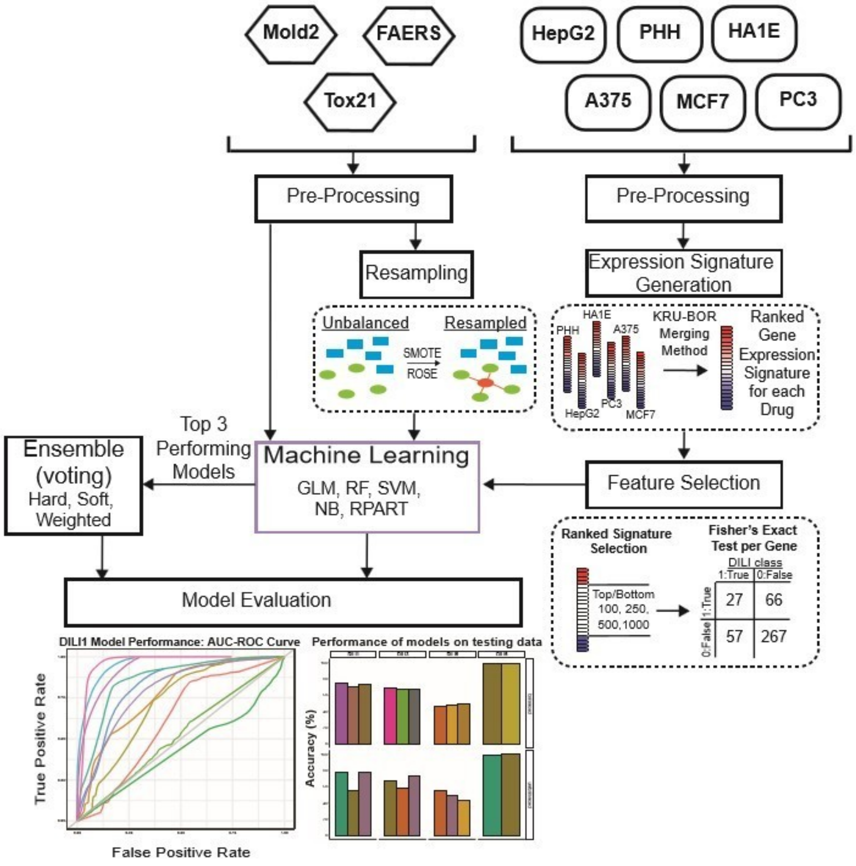

Figure 1. Study Workflow.

Data were separated into expression-based datasets and non-expression-based (FAERS, MOLD2, TOX21) for testing. Non-expression data was evaluated with resampling methods ROSE and SMOTE as well as an unbalanced dataset. Expression-based datasets were merged across cell lines into one representative expression signature per drug. These signatures were tested as the top and bottom 100, 250, 500, and 1,000 ranked genes for each drug. Following signature formation, feature selection using a fisher's exact test was used to determine significant predictors of DILI classification. Machine learning was used on predictors for both expression-based and non-expression models, which were evaluated based on training AUC curve values as well as testing performance. The top three performing models for each DILI type were utilized in ensemble voting models in an effort to incorporate both expression and non-expression datasets.

.

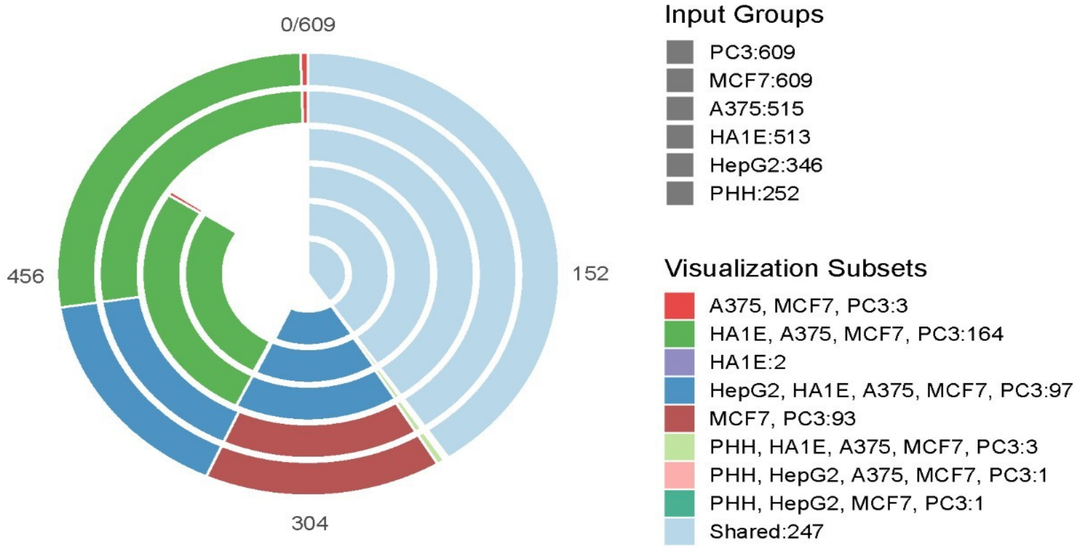

Figure 2. Drug Testing Cell Distribution.
The Venn-Pie diagram depicts the overlap of drugs tested between each of the six cell lines used in this study. Each bar within the Venn-Pie represents an individual dataset while the color of the bars indicates the overlapping group of compounds across datasets. While 247 of the 617 drugs included in the training and test data were tested in all 6 cell lines, some compounds were only tested in a singular cell line and others did not have any expression information provided.

Food and Drug Administration Adverse Event Reporting System (FAERS)

The CAMDA organizers provided us with FAERS data for all 617 drug compounds. Of these, 422 were grouped as “training data”. This dataset contains 20 features corresponding to information on the percentage of reported adverse events for each drug compound by gender and age group demographic. After removing highly correlated features, we upsampled the data to cater to the class imbalance by randomly sampling with replacement from the minority class to balance the majority class. An additional preprocessing step was to create two new variables, namely “male ratio” and “female ratio”, taking into account all reported events irrespective of the gender, all reported DILI events irrespective of gender, and the percentage of reported DILI events by gender.

Toxicology in the 21st Century (TOX21)

In addition to the FAERS dataset, we were provided with concentration-response information of 600 drugs. Of these, 412 were designated as “training data”. Thirty-two features corresponded to concentration-response curve ranks. Out of all 412 drugs for training, 57 drugs were removed for missing values. In addition, we removed highly correlated features using an arbitrary cutoff of 0.82 and catered to the class imbalance by using SMOTE.

Molecular Descriptors from 2D Structures (MOLD2)

Alongside the FAERS and TOX21 data provided, we had access to the 2D molecular descriptors or structural information of these 617 drug compounds. 422 of these drugs were designated for training. There were 777 features for each drug compound with each feature corresponding to MOLD2 descriptors. To cater to class imbalance, we upsampled minority classes, as well as ROSE, and SMOTE.

Connectivity Map L1000 Gene Expression Data

The L1000 assay data used in this study is a high-throughput gene expression assay that measures mRNA transcript abundance of 978 landmark genes based on an inference algorithm to infer the expression of 11,450 additional genes in the transcriptome [21]. Utilizing simulation, it has been observed that this reduced representation of the transcriptome can recapitulate around 80% of the relationships of measuring the entire transcriptome directly. In this study, 12,328 de-identified predictor genes were provided by the CAMDA organizers with Z scores to indicate transcript abundance. The treatment time and dosage of each drug were selected by the CAMDA committee to produce the largest available dataset for both test and training data.

Kruskal-Borda Merging

Since not all drugs were tested in each cell line data made available, we utilized the Kruskal-Borda (Kru-Bor) merging algorithm in the GeneExpressionSignature R package [86]. This approach allowed us to generate a unified drug-induced expression signature across cell types since many drugs were not tested in the PHH or HepG2 liver cell lines. The Kruskal algorithm [87] finds a minimum spanning forest of an undirected edge-weighted graph while the Borda merging method [88,89] uses ranked options in order of preference to determine the outcome. Thus each closest neighbor in rank merges one by one until a unified signature is formed. Following merging, the top and bottom 100, 250, 500, and 1,000 ranked genes were selected as drug signatures for feature selection.

Feature Selection

A method of feature selection utilized across the merged signatures produced via our Kru-Bor merging was based on a gene's significance (p-value < 0.01) in predicting the DILI score via a Fisher's exact test. If a gene is included in the top or bottom 100, 250, 500, or 1,000 ranked list, depending on the model data, for any drug it would be assigned a 1 (True), or if it fell outside of that range it would be assigned a 0 (False). The classifier for each type of DILI was also 1 (DILI positive) or 0 (DILI negative). We used these classifiers to identify if these highly influenced genes were predictive of a drug being DILI positive or DILI negative with a p-value cutoff of 0.01.

Machine Learning

The prediction of DILI was treated as a binary classification problem for each DILI type. That is, for each of DILI1, DILI3, DILI5, and DILI6, outcomes were split between 'positive' and 'negative'. We used a 5-folds cross validation repeated 100 times, and a random search strategy to search for the best parameters for each model. The data was made available such that training and test sets had been pre-identified. Importantly, we did not have access to the correct labels for the test data. Models were built using traditional machine learning algorithms within the caret [90] package in R version 4.0.0 [91].

The machine learning algorithms we used are suitable for classification tasks. They include a Logistic Regression (LR) [92], Linear Discriminant Analysis (LDA) [93], Decision Trees (DT) [94], Support Vector Machines (SVM) [95], Naïve Bayes (NB) [96], a One-layer Neural Network (Nnet), and a Random Forest (RF) algorithm. LR and LDA are generally categorized as linear classification models, with an assumption that the data follows a normal distribution. Given a set of predictors, LR aims to build a linear model of these predictors by minimizing the sum of squared residuals. LDA uses the prior probability of belonging to a class to estimate posterior probabilities by using Bayes' Theorem. DT and RF are often classified as trees and rules-based algorithms. Given a set of predictors, a decision tree works by using if-else conditions to build a definitive set of rules using splits. The challenge usually lies in determining optimal situations to apply a “then”-clause (or a split). In RF, similar conditional statements are used. However, instead of using the entire sample of data for tree-building, RF uses many independent subsamples from the training data to build small decision trees. Each small decision tree classifies an observation by voting. Neural networks and SVMs are generally grouped as non-linear algorithms. Neural networks (in our case, a multilayer perceptron i.e. a neural network with one hidden layer), are modeled after how neurons in the human brain work. The outcome or prediction is a linear combination of the hidden layer(s) transformed by a non-linear activation function. There are several activation functions used, depending on whether the problem is a regression or classification problem. In our case, we used a sigmoidal or logistic function, since we were dealing with a classification problem. SVM aims to find support vectors or data points that separate the different classes as much as possible. Intuitively, these data points are the most difficult to separate (the reasoning is that they lie very close to one another and to the hyperplane or decision boundary), and are thought of to be important in separating classes. There are different flavors of SVMs depending on the kernel used (kernels are similar to non-linear activation functions used in neural networks). In the current study, we used polynomial, linear, and a radial basis function kernels.

Model Evaluation

To evaluate the performance of our models, we focused on the area under the ROC (Receiver Operating Characteristic) curve (AUC) (Equation 6) value as well as the specificity (Equation 2), sensitivity (Equation 1), accuracy (Equation 3), and MCC (Equation 5) of the models on the test set. ROC illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. It plots two characteristics, true positive rate (TPR) against the false positive rate (FPR), at various thresholds. Therefore, the AUC value is a measurement of the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one, and it is a widely-used metric in binary classification problems. An AUC value of 1 indicates a perfect classifier, i.e. a model that is perfectly able to separate both classes, while an AUC value of 0.5 indicates a model that predicts at random. Depending on the application domain, AUC values of 0.7 and above are usually acceptable. Specificity measures the ratio of negative classes that were correctly identified by the model out of all negative classes, while sensitivity measures the ratio of positive classes that were correctly identified by the model out of all positive classes. These metrics are affected by how the target labels are structured and passed to the algorithm, and they range from 0 to 1. Additionally, we evaluated the performance of our models on the test set by calculating the balanced accuracy (Equation 4) of prediction. Balanced accuracy is the average of the sensitivity and specificity or the average of the fraction of correct labels that are predicted correctly (by the model) within each class. We used this metric because we observed that there was class imbalance within our datasets regardless of DILI type.

The MCC is particularly useful in datasets of different class distributions (or imbalanced data) because it considers all of the false and true positives and negatives. It is calculated from the confusion matrix of a model and its values range from +1 to -1, with +1 indicating a perfect classification, 0 indicating random classifications, and -1 indicating no relationships between the observed and predicted classes.

Equation 1

$$Sensitivity/TPR=\frac{TP}{TP+FN}$$

Equation 2

$$Specificity/TNR=\frac{TN}{TN+FP}$$

Equation 3

$$Accuracy\left(ACC\right)=\frac{TP+TN}{P+N}=\frac{TP+TN}{TP+FN+TN+FP}$$

Equation 4

$$Balanced\quad accuracy=\frac{TPR+TNR}{2}$$

Equation 5

$$Matthews\quad Correlation\ Coefficient\ \ \left(MCC\right)=\frac{TPxTN-FPxFN}{\sqrt{\left(TP+FP\right)\left(TP+FN\right)\left(TN+FP\right)\left(TN+FN\right)}}$$

Equation 6

$$AUC\left(Area\ under\quad the\quad curve\right)=\int_{x=0}^{1}TPR\left(FPR^{-1}\left(x\right)\right)dx\ ,$$

Where TPR and TNR are the true positive rate and true negative rate respectively, TP, FP, TN, and FN are the number of true positive, false positive, true negative, and false negative, respectively while P and N are the number of positive cases and that of negative cases in the data, respectively.

Ensemble Voting Machine Learning

In an attempt to improve the classification accuracy of our models, we used three ensemble voting approaches, namely soft voting, hard voting, and a weighted voting approach. These ensemble methods work best when there are varying algorithms of different strengths i.e. algorithms having varying underlying assumptions about the data, and when each one has reasonable predictive power [81,82]. Using the gene expression data provided by CAMDA 2018 organizers, Sumsion and colleagues [16] used hard and soft voting ensemble methods in an attempt to improve prediction accuracy on DILI risk. As an extension of their work, we hypothesized that since we have access to larger and more diverse datasets, we could capture different aspects of predicting DILI types and use these ensemble methods to improve prediction.

Hard voting, also known as majority voting, takes into account the predicted class labels of each classifier (or voter) [97]. Voting is done by counting how many class labels (for each class) were predicted among all classes. The class label with the highest count is taken to be the predicted class label for that observation. On the other hand, soft voting considers the probabilities of each class label by each classifier [98]. In other words, it considers how certain each classifier is about the class labels. For each class label, the probabilities are averaged, and the label with the highest average probability is taken as the predicted class label for the observation.

The third approach to voting involves using a weight to skew predictions towards the most certain models (Equation 7). In our approach, we used the AUC of each classifier as a weighting parameter for the output probabilities. This was done to take into account that some classifiers might have better predictive power and should be given preference in determining the outcome of the voting. To weigh each probability, we multiplied the probabilities of each predicted class by the AUC and divided this by 1 subtracted from the weight, that is, the AUC of that model. Afterward, weighted probabilities were treated just as in soft voting: by taking the average of all resulting weighted probabilities belonging to each class. The class label with the higher average was taken as the predicted class for that observation. Therefore, the predicted class, $\hat{y}$, of observation, given an output set of class membership probabilities across many models, P, is given by:

Equation 7

$$C(\hat{y}|P) = argmax_c \frac{1}{m} \sum_{i=1}^{m} \left(\frac{w_i * p_i^c}{w_i - 1} \right)$$

i.e. a class with the highest weighted average membership of the models, where m is the total number of models, i.e., $|P| = m$, $w_i$is the weighting parameter for a model i, and p_i^c is the probability of class membership of model i to a class c.

CHAPTER 4

RESULTS

FAERS Modeling

The performance of FAERS data in predicting each of the DILI types can be seen in the bar plots in Figure 3. While we built many models, we compared and picked the best three models based on the AUC values to predict DILI class on the test set. We noticed that using the raw data (without resampling), models achieved classification accuracy between 0.51 and 0.55 and MCC between 0.04 and 0.14 on the training set and did not do noticeably better on the test set (accuracy: 0.49 to 0.59, MCC: -0.03 to 0.22). On the other hand, using resampled datasets improved the accuracy of the models on the training set to a range of 0.61 to 0.94 (MCC: 0.47 to 0.89). Using these models to predict the DILI class of the test set showed a slight improvement in the accuracy (0.52 to 0.62). The MCC, however, was between 0.04 and 0.24.

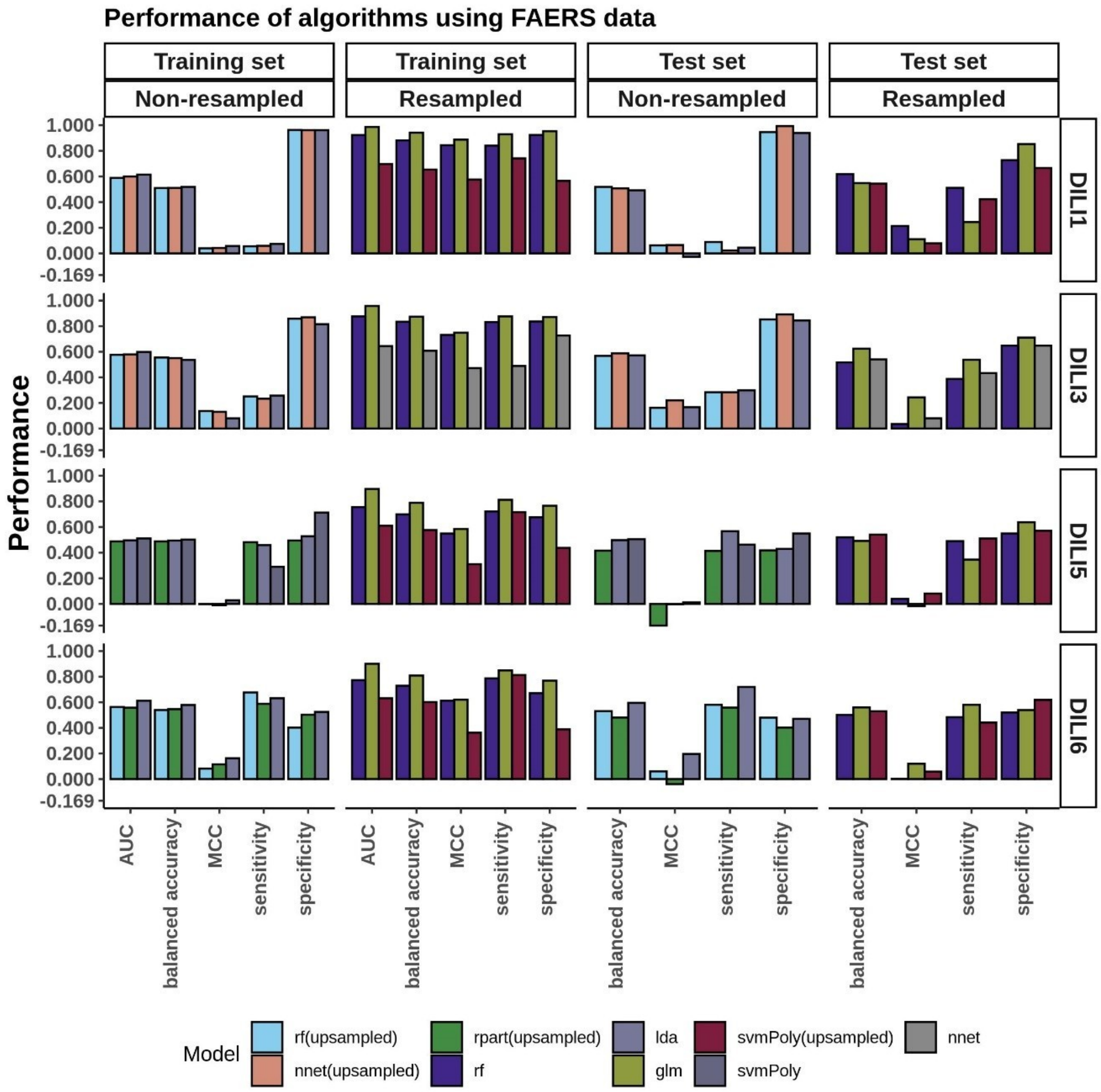

Figure 3. FAERS Model Performance.
Performance evaluation of the DILI predictive models built using the FAERS reporting data was conducted on both the original unbalanced and the resampled/balanced datasets. The best performing algorithm determined by AUC between GLM, IDA, NB, NNET, RF, RPART, and SVMPoly were selected. For DILI1 and DILI3, the highest accuracy was 0.62 with MCC values of 0.21 and 0.24, respectively.

MOLD2 Modeling

Similarly to how the FAERS data was handled, we selected the top three performing models built using MOLD2 data in each category (resampled or non-resampled) to predict the DILI class of the test data (Figure 4). Models built using the non-resampled MOLD2 dataset gave accuracies of 0.50 to 0.54, showing that the models were randomly predicting the classes (MCC: 0.00 to 0.17). This performance was similar on the test set (accuracy: 0.50 to 0.66, MCC: -0.01 to 0.36) with a slight improvement. Similarly to what we observed using FAERS data, resampling the dataset improved both the accuracy and the MCC of the training set (accuracy: 0.71 to 0.78, MCC: 0.56 to 0.76) but could not generalize better than non-resampled MOLD2 data to the test set (accuracy: 0.51 to 0.67, MCC: 0.14 to 0.36).

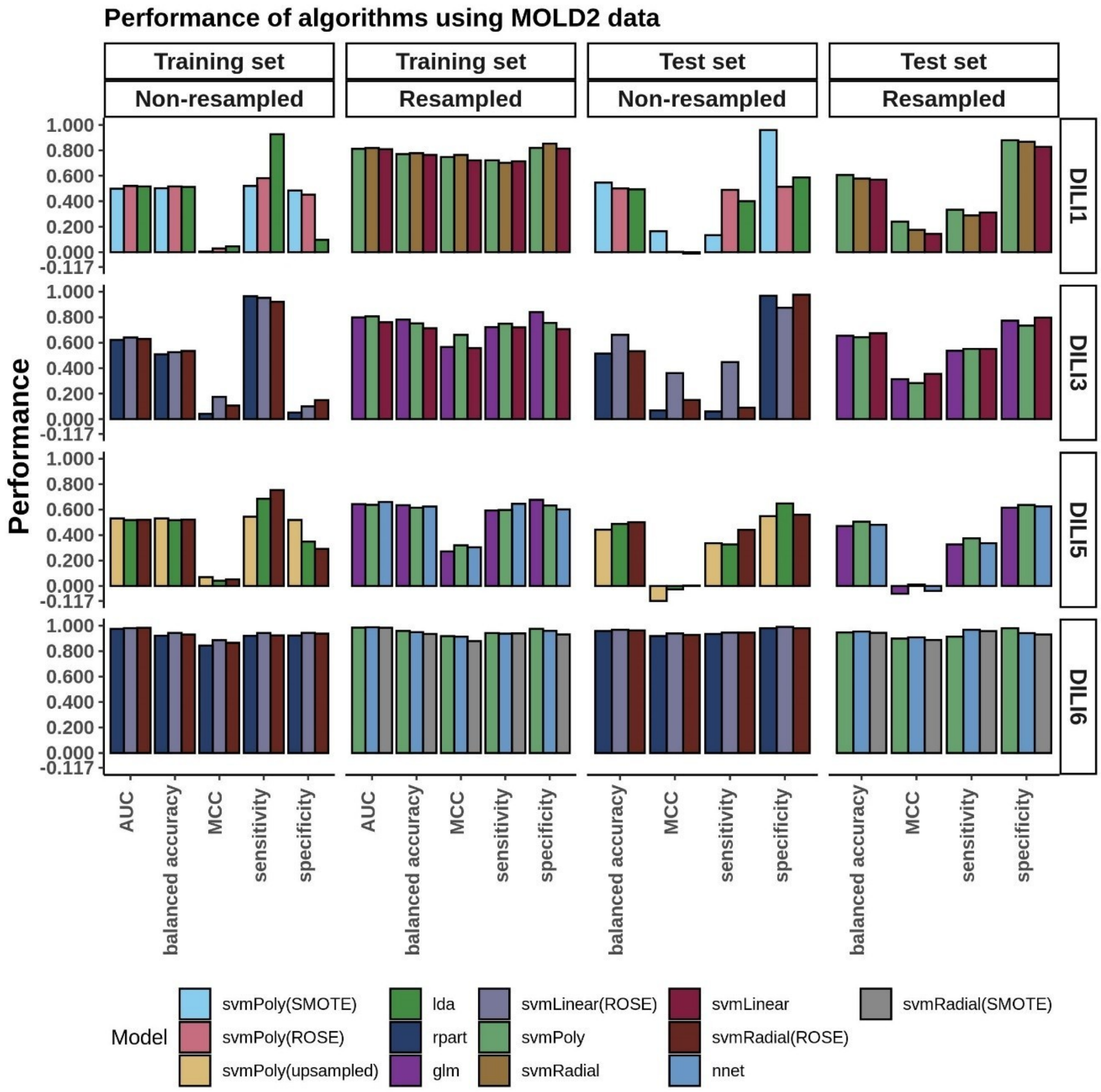

Figure 4. MOLD2 Model Performance.
The chemical structural information from MOLD2 was imbalanced between DILI positive and negative samples. Predictive models were evaluated on both the unbalanced and resampled/balanced datasets. The three best-performing models for each DILI type, based on AUC and resampling methods, are depicted in the bar graphs.

TOX21 Modeling

The top three models built using TOX21 data (using the AUC as the criterion) were evaluated on the test set (Figure 5). Using the data as is, without resampling, the accuracy of the training data was between 0.50 and 0.57 (MCC: -0.02 to 0.17). As expected, the models failed to generalize to the test set (accuracy: 0.50 to 0.59, MCC: -0.04 to 0.19). Again, we observed that resampling slightly improved the accuracies of these models on the training set (accuracy: 0.62 to 0.76, MCC: 0.25 to 0.54). Yet, there was no major improvement on the test set (accuracy: 0.50 to 0.58, MCC: -0.01 to 0.20).

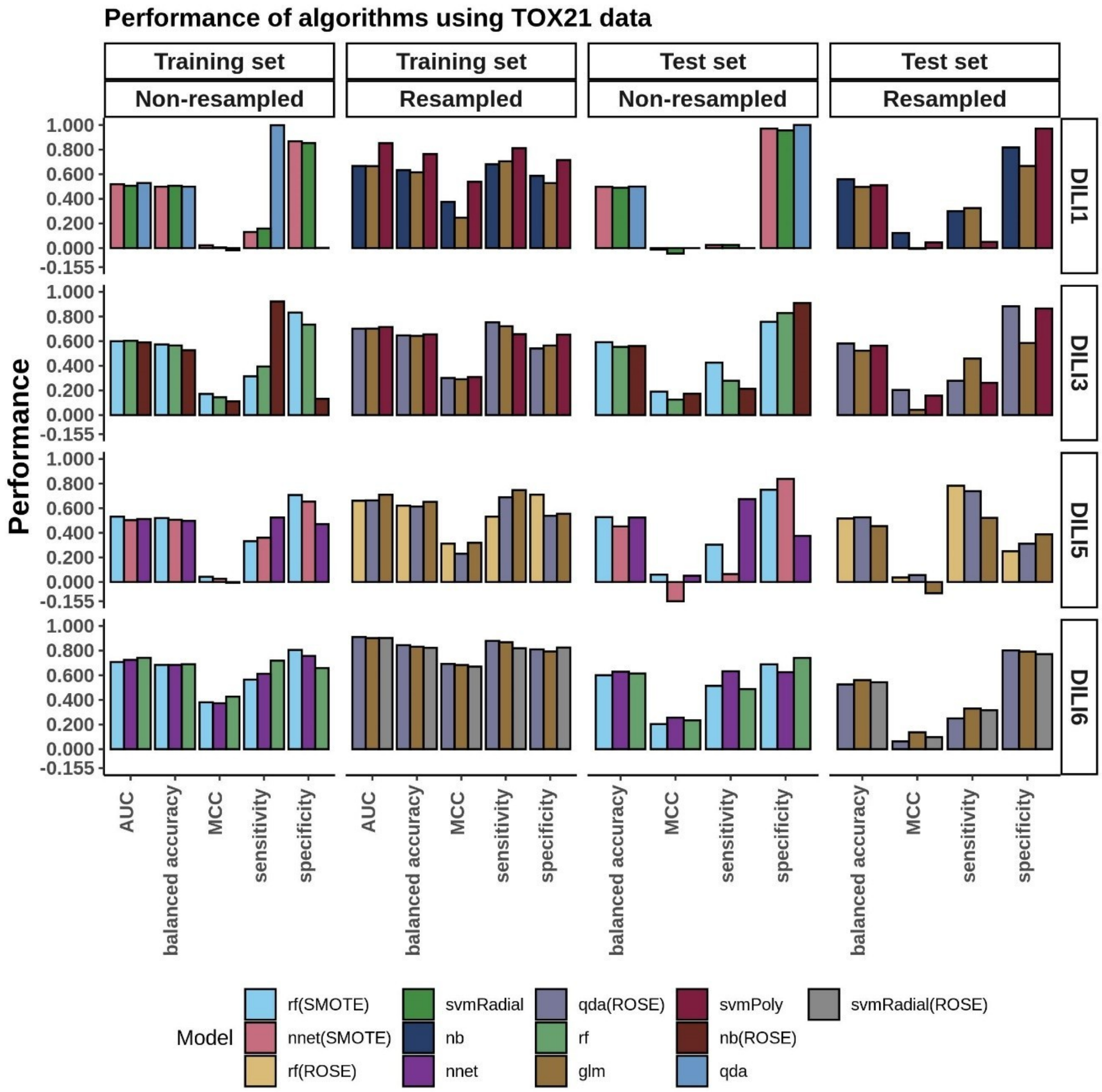

Figure 5. TOX21 Model Performance.
The performance of DILI predictive models built using the toxicology information provided from TOX21. The three best-performing algorithms, based on training AUC and based on whether resampling was used or not, are presented in the bar plots.

Connectivity Map L1000 Cell Expression Modeling

Cellular RNA expression levels in the form of microarray data have been previously investigated for their ability to predict DILI with limited predictive power [15]. In the current study, the L1000 data from the Connectivity Map was used including both the measured landmark genes as well as the inferred transcriptome. We built models using each expression data to investigate which cell lines were most successful in predicting DILI. Table 4 summarizes the model results based on our training data. However, due to the limitation of each cell only providing expression response data from a subset of drugs (Figure 2) involved in the training and test data, accuracy based on test data was not meaningful. Additional processing steps for this data involved merging across the six cell lines to generate a representative signature, testing different cutoffs for the amount of highest- and lowest-ranked genes to utilize, as well as a feature selection for determining predictor genes.

The models built using the merged expression signatures with the highest AUC from the training data were evaluated on the test set. The training and test results are summarized in the bar plots in Figure 6. None of the cell expression signatures performed well when predicting DILI3, DILI5, or DILI6 with an accuracy ranging from 0.39 to 0.64 and MCC values ranging from -0.03 to 0.1. These models did have some limited success predicting DILI1 with the merged SVM 1000 model performing the best, reaching an accuracy of 0.67 but an MCC of 0.10 (Table 5). The poor predictability of DILI3 status by these models was unexpected with the accuracy of the best model being 0.49 with 0.33

sensitivity and 0.66 specificity. The limited success in predicting DILI5 and DILI6 was expected based on the positive and negative control construction of these DILI classes, which are not reflected in the gene expression data.

Table 4. Training Performance on Independent Cell Line based Models. Each of the six cell lines with L1000 expression data were used to build predictive models of the four DILI classes. Training performance results for the best performing model for each cell type and DILI class are shown as well as the number of predictors following feature selection as described in the methods section.

DILI Class	Cell Type Tested	ML Algorithm	Predictors	AUC-ROC	Sensitivity	Specificity
DILI 1	PHH	SVM	60	0.969	0.912	0.945
	Hep G2	SVM	72	0.922	0.924	0.693
	HA1E	GLM	40	0.781	0.903	0.389
	A-375	GLM	178	0.627	0.826	0.17
	MCF7	GLM	65	0.722	0.898	0.222
	PC3	RF	315	0.589	1.000	0
DILI 3	PHH	NB	50	0.931	0.547	0.957
	Hep G2	RF	75	0.913	0.942	0.625
	HA1E	SVM	176	0.922	0.953	0.788
	A-375	SVM	3610	0.833	0.869	0.607
	MCF7	SVM	74	0.861	0.872	0.742
	PC3	SVM	345	0.844	0.863	0.606

DILI 5	PHH	GLM	8	0.723	0.484	0.761
	Hep G2	RF	17	0.719	0.984	0.229
	HA1E	GLM	20	0.711	0.693	0.513
	A-375	GLM	24	0.724	0.786	0.561
	MCF7	RF	38	0.679	0.803	0.355
	PC3	GLM	14	0.661	0.255	0.961
DILI 6	PHH	GLM	2	0.574	0.087	0.990
	Hep G2	RF	31	0.686	0.000	1.000
	HA1E	RF	27	0.688	0.247	0.949
	A-375	GLM	16	0.619	0.181	0.945
	MCF7	RF	24	0.689	0.186	0.975
	PC3	RF	53	0.724	0.159	0.986

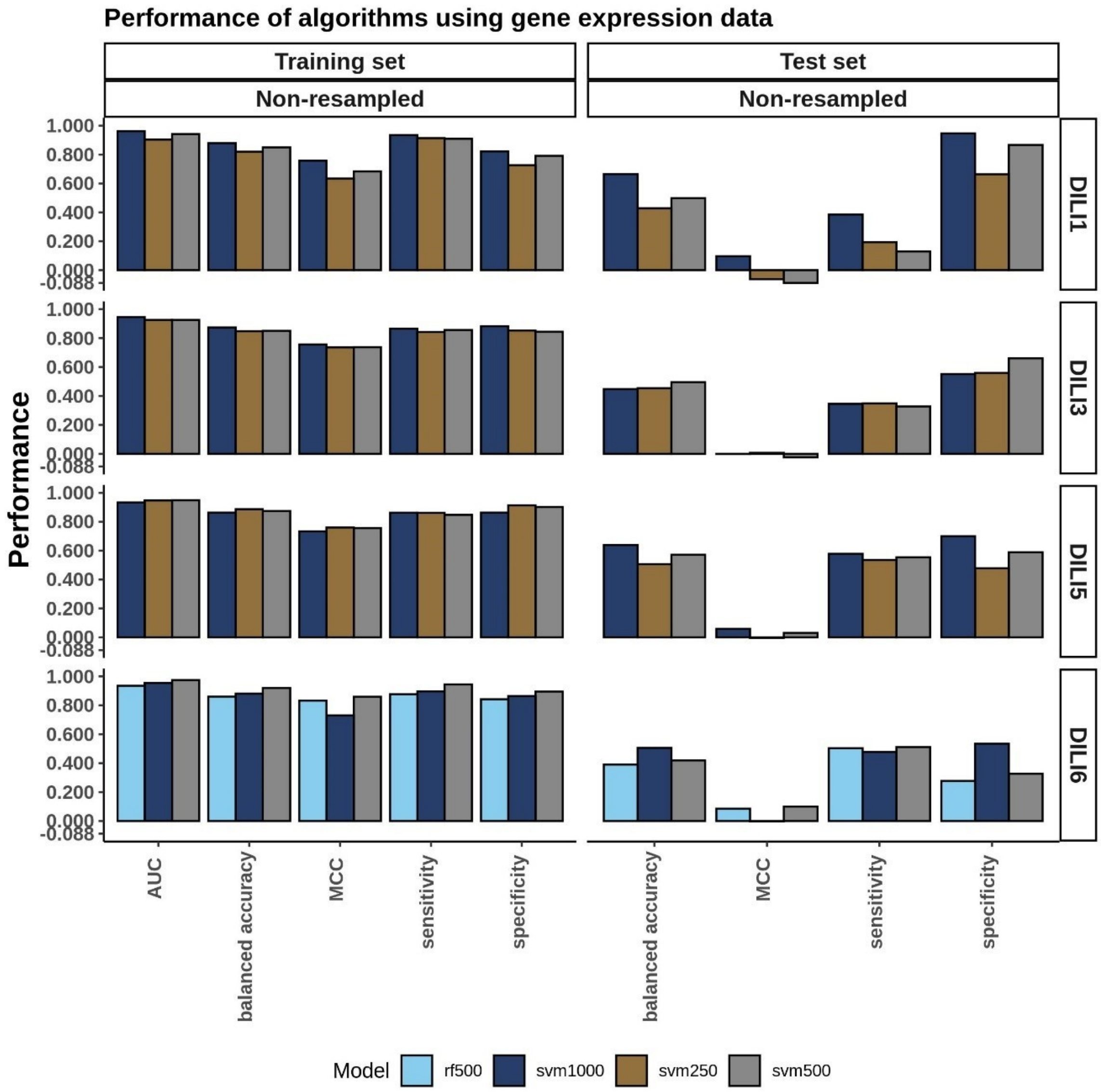

Figure 6. Cell Expression Model Performance.
A single cell expression signature for each drug was generated using Kru-Bor merging across all cell lines in which the drug was tested as described in the methods. Following merging, feature selection using a fisher's exact test was performed on expression signatures of the top and bottom 100, 250, 500, and 1,000 ranked genes. Models built on these predictors were evaluated and the top-performing ones, based on AUC, are shown in the training set bar graph.

Table 5. Testing Performance of Top Models.
The testing result metrics from the best model built using each dataset as well as the ensemble voting model.

Dataset	DILI Class	Algorithm	Test Sensitivity	Test Specificity	Test MCC	Test Balanced Accuracy
Merged Expression	DILI1	SVM	0.38	0.95	0.1	0.67
	DILI3	SVM	0.33	0.66	-0.03	0.49
	DILI5	SVM	0.58	0.7	0.06	0.64
	DILI6	SVM	0.48	0.53	0	0.51
FAERS	DILI1	NNET	0.51	0.73	0.21	0.62
	DILI3	RF	0.54	0.71	0.24	0.62
	DILI5	RPART	0.51	0.57	0.08	0.54
	DILI6	RF	0.72	0.47	0.2	0.6
MOLD2	DILI1	SVMPoly	0.33	0.88	0.24	0.61
	DILI3	SVMPoly	0.55	0.8	0.36	0.67
	DILI5	SVMPoly	0.38	0.64	0.01	0.51
	DILI6	SVMPoly	0.95	0.99	0.94	0.97

TOX21	DILI 1	NNET	0.3	0.82	0.12	0.56
	DILI 3	GLM	0.43	0.76	0.19	0.59
	DILI 5	GLM	0.3	0.75	0.06	0.53
	DILI 6	QDA	0.63	0.62	0.26	0.63
Ensemble Voting	DILI 1	Weighted voting	0.16	0.92	0.11	0.54
	DILI 3	Weighted voting	0.3	0.89	0.24	0.6
	DILI 5	Weighted voting	0.28	0.71	-0.01	0.5
	DILI 6	Weighted voting	0.96	0.97	0.93	0.96

Ensemble Voting Models Performance

Since the top three individual models did not perform well on the test set (Table 5), we asked if aggregating the top three models in an ensemble approach could improve the accuracy. To test this, we applied three ensemble voting methods namely soft voting, hard voting, and weighted voting. Hard voting gave accuracies of 0.39 and 0.37 on DILI1 and DILI3, respectively, while soft voting gave an accuracy of 0.44 and 0.40 for DILI1 andDILI3, respectively" to "Hard voting gave accuracies of 0.39 and 0.37 on DILI1 and DILI3, respectively, while soft voting gave an accuracy of 0.44 and 0.40 for DILI1 andDILI3, respectively (Figure 7). Soft voting slightly improved the accuracy of these models most likely because it considers membership probabilities rather than predicted class labels. We observed that weighted voting slightly improved the accuracy: 0.54 for DILI1 and 0.60 for DILI3. Our weighted approach considers both the probabilities and the AUC of the models and emphasizes the contribution of models with higher AUCs. Sumsion and colleagues used similar approaches (soft and hard voting) with gene expression data resulting in decreased accuracies [16]. Compared to their study, our approach improved the accuracies of the models. However, our method(s) does not report MCCs because we do not have access to the true positives, true negatives, false positives, and false negatives in the test data.

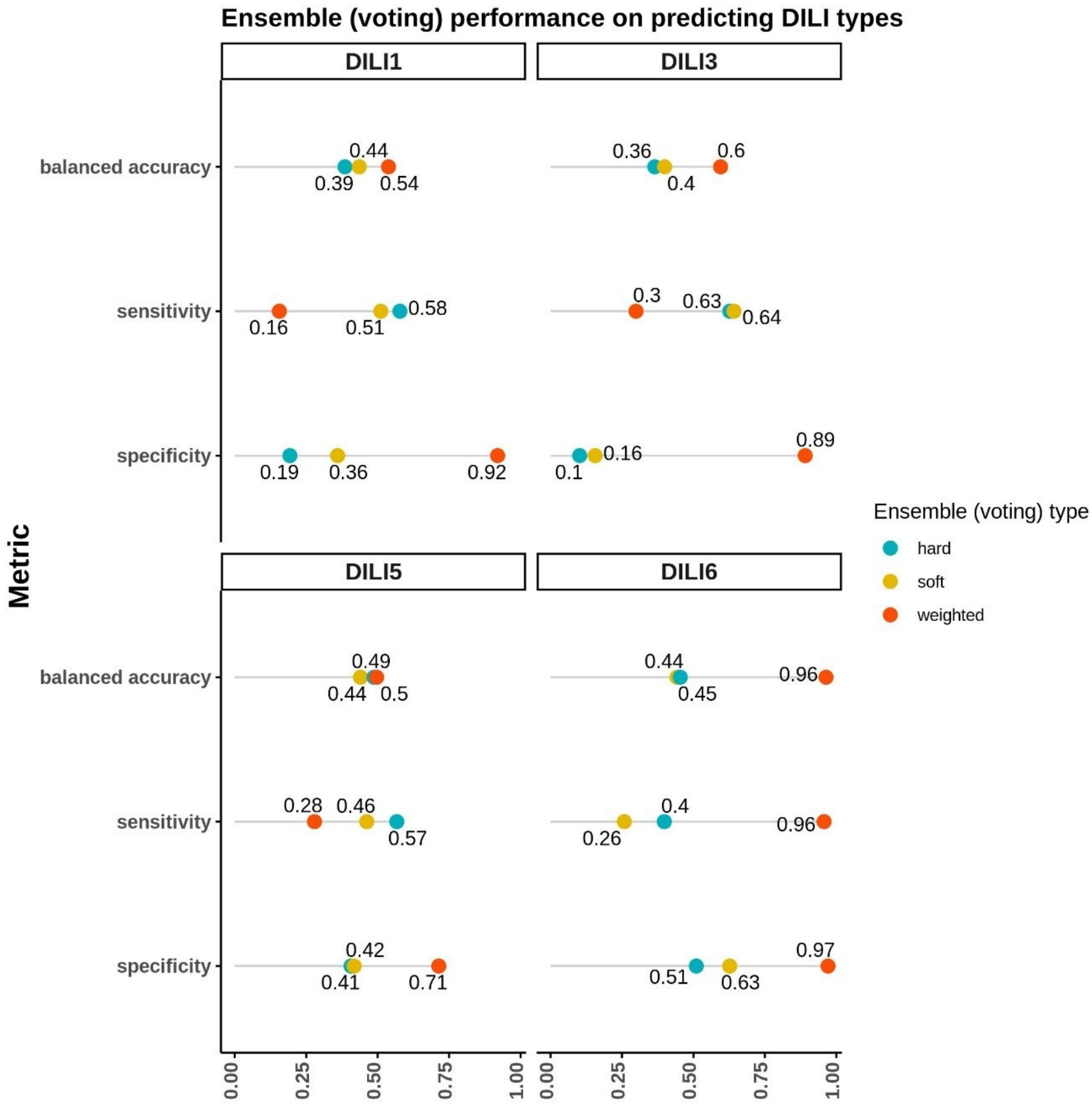

Figure 7. Ensemble Voting Method Performance.

To incorporate the various types of data provided ensemble methods including hard, soft, and weighted voting were tested using the top three performing models for each DILI type.

CHAPTER 5

DISCUSSION

CAMDA 2020 was a collaborative challenge to establish predictive models of DILI using gene expression data as well as a combination of data from clinically reported events, drug structure, and toxicology. In our study, we evaluated the predictability of these datasets on four DILI types, namely, DILI1 (severity score ≥ 6), DILI3 (withdrawn, box warning, warning, and precaution), DILI5 (negative control), and DILI6 (positive control). These datasets included gene expression/perturbation data on six cell lines (PHH, HEPG2, HA1E, A375, MCF7, and PC3), concentration-response or toxicology information, 2D molecular descriptors of the drug structure, and reported adverse events. To assess the predictive abilities of these datasets, we used various traditional machine learning algorithms. For non-gene expression datasets, we corrected the imbalance issue using well-known techniques like SMOTE, ROSE, and upsampling the minority class.

While CAMDA previously approached predicting DILI, there have been significant improvements in the data provided and scopes of the challenge each year. In 2018 the challenge data only included microarray expression data from non-liver relevant cell lines on 276 compounds with a binary DILI classification. Published results from the 2018 challenge indicate limited success from both deep learning and soft voting approaches which achieved a maximum accuracy of 0.7 and MCC values <0.2 [15,16]. When the CMap drug safety challenge was re-administered in 2019, the data expanded to L1000 transcriptomic data on 13 cell lines and allowed participants to use external data sources such as protein-protein interactions, drug-protein targets, and chemical descriptors. The DILI classifications for this challenge also changed from binary to a most, less, ambiguous, and no-DILI concern which is in line with the FDA DILIrank dataset. Predictive model rates from multiple distinct approaches to this challenge in 2019 often yielded similar accuracy results around 0.70 [17,18,99].

While it is difficult to make a direct comparison across the years of these challenges considering how the fundamental elements of predictive modeling, such as the data sources and classifications, have changed, the goal of the challenge has remained the same in modeling the risk of a drug to lead to liver injury in patients. The data structure of the challenge has also improved in each iteration attempting to expand the predictive data power as well as the data sample size to allow for more robust modeling. However, as in previous years, the highest accuracy we were able to achieve in the current study was 0.67 for DILI1 and DILI3 with the highest MCC value of 0.36. This suggests that there are still rooms to improve both in model construction as well as in developing robust predictive data, which captures the scope of DILI.

In our study, we developed models with gene expression data using individual cell lines, as well as a merging of these datasets. Each cell line dataset did not include all the drugs thereby reducing the size of the training data and making it difficult to evaluate each of them on the test set. Therefore, we merged these datasets into one expression signature across cell types. Further, we selected the 100, 250, 500, and 1,000 most upregulated and downregulated genes as an arbitrarily signature cutoff of the most perturbed genes by drug treatment. However, our approach failed to capture predictive differences between the positive and negative classes in each DILI type. Although we achieved an accuracy of 0.67 for DILI1 (on the test set), a sensitivity of 0.38 showed that our models were not learning the positive classes well enough. Usually, this problem is due to not having sufficient training examples for a particular class. In contrast, we could obtain specificity as high as 0.95, showing that the model could learn the negative classes well since there were more DILI negative drugs in the training set. Table 5 summarizes the best performances on the test set. We observed that many of these models failed to generalize to the test set i.e. they showed poor predictability on the test set (Table 5).

Since the individual models did not perform well on the test set, we attempted ensemble (voting) methods to improve prediction accuracy. We used soft voting, hard voting, and weighted voting approaches. In weighted-voting methods, there are diverse ways through which importance can be attached to each model. Weight-based ensemble methods tend to outperform single models, and even soft voting, because in addition to the posterior probabilities churned out by the models, they take into consideration some importance or weighting factor [100]. Although these methods could not improve test accuracy beyond individual models, weighted voting performed better than soft and hard voting because weighs the predicted probabilities of the test examples by the performance of each model.

One challenge we had was that the training set was perhaps too small to be further split into a training and validation set. However, machine learning algorithms benefit most from having sufficient examples. For some datasets such as the gene expression datasets, we did not have access to information on all 617 drugs, which reduced the size of the training data. Besides, the training data were largely unbalanced (Table 3). For instance, for DILI1, there were 96 positive examples and 326 negative examples. This problem resulted in many of our models having low sensitivities since the positive examples were insufficient. In an attempt to address this problem, we employed resampling techniques (SMOTE, ROSE, and upsampling minority classes) to balance the datasets. However, it was clear that models built using balanced (resampled) data were overfitting the training set. A possible reason for this was that due to our resampling approach, some training examples were also used in the validation stage during cross-validation. In addition, due to having blinded datasets, we could not explore how the features were influencing the models.

Future Work

In summary, our study suggests that currently available data, including mRNA quantification, molecular descriptors, clinically reported events, and toxicology profiles, may be inadequate to capture important information enough to separate DILI classes in real-world scenarios.

Machine learning algorithms work best when the datasets are large enough to capture all predictive spaces. The size of current DILI datasets, however, is not sufficient. Larger datasets may be needed to encourage the application of deep learning algorithms which typically do better with bigger data. Additionaly, we suspect that a limitation to DILI prediction lies in inadequate biomarker identification, and in the lack of adaption of these kind of information in predicting DILI. We hypothesize that an additional focus or challenge to predict biomarkers specific for DILI using various –omics data, for instance, single-cell data and metabolomics signatures, and incorporating these into the wider cause of DILI prediction will improve the status quo.

Another problem with DILI prediction is that it is heavily annotation-dependent. In a 2015 paper, Xu and colleagues discussed how this discrepancy affected their predictive models[13]. The inconsistency in DILI annotations have been heavily discussed in literature[101]. To mitigate this issue, the FDA began unifying annotations resulting in the DILI rank dataset[5], which was used in this study. Although today these annotations are better unified, they still present as a challenge because they are human-annotated, based on at least one reported incidence of the medication causing DILI, and may not hold information on mechanistic pathways of DILI development for each drug. To improve DILI prediction, we hypothesize that more sensitive annotations, based on some biological parameter e.g. presence or absence of some serum biomarkers, interference with an important DILI pathway, or upregulation of a set of DILI-related genes, may be needed, as opposed to annotations solely based on patients' experiences.

In future studies, rather than use de-identified datasets for prediction, we aim to use unblinded data. By knowing the identity of the features we are dealing with, we can better understand and model the predictive space. Better still, we can include more complicated, in-depth analysis of the data like network and pathway analysis. Although the current status of DILI prediction is unsatisfactory, there has been much improvement over the years, especially regarding the kinds of datasets that might be needed, alongside computational methods that can improve the DILI prediction. In the coming years, we hope that with more sensitive approaches that make use of mechanistic and molecular insights to the development of DILI, better machine learning models can be built to solve this problem.

REFERENCES

1. Daly AK. Pharmacogenomics of adverse drug reactions. *Genome Med*. Published online 2013. doi:10.1186/gm409

2. Marzano A V., Borghi A, Cugno M. Adverse drug reactions and organ damage: The skin. *Eur J Intern Med*. Published online 2016. doi:10.1016/j.ejim.2015.11.017

3. Atienzar FA, Blomme EA, Chen M, et al. Key Challenges and Opportunities Associated with the Use of in Vitro Models to Detect Human Dili: Integrated Risk Assessment and Mitigation Plans. *Biomed Res Int*. Published online 2016. doi:10.1155/2016/9737920

4. Berlin JA, Glasser SC, Ellenberg SS. Adverse event detection in drug development: Recommendations and obligations beyond phase 3. *Am J Public Health*. Published online 2008. doi:10.2105/AJPH.2007.124537

5. Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today*. Published online 2016. doi:10.1016/j.drudis.2016.02.015

6. Ozer J, Ratner M, Shaw M, Bailey W, Schomaker S. The current state of serum biomarkers of hepatotoxicity. *Toxicology*. Published online 2008. doi:10.1016/j.tox.2007.11.021

7. Andrade RJ, Chalasani N, Björnsson ES, et al. Drug-induced liver injury. *Nat Rev Dis Prim*. Published online 2019. doi:10.1038/s41572-019-0105-0

8. Messner CJ, Premand C, Gaiser C, Kluser T, Kubler E, Suter-Dick L. Exosomal microRNAs Release as a Sensitive Marker for Drug-Induced Liver Injury in Vitro. *Appl Vitr Toxicol*. 2020;6(3):77-89. doi:10.1089/aivt.2020.0008

9. García-Cortés M, Ortega-Alonso A, Lucena MI, Andrade RJ. Drug-induced liver injury: a safety review. *Expert Opin Drug Saf*. Published online 2018. doi:10.1080/14740338.2018.1505861

10. Saini N, Bakshi S, Sharma S. In-silico approach for drug induced liver injury prediction: Recent advances. *Toxicol Lett*. Published online 2018. doi:10.1016/j.toxlet.2018.06.1216

11. Shin HK, Kang MG, Park D, Park T, Yoon S. Development of Prediction Models for Drug-Induced Cholestasis, Cirrhosis, Hepatitis, and Steatosis Based on Drug and Drug Metabolite Structures. *Front Pharmacol*. 2020;11:1. doi:10.3389/fphar.2020.00067

12. Liu X, Liu X, Zheng D, et al. Machine-Learning Prediction of Oral Drug-Induced Liver Injury (DILI) via Multiple Features and Endpoints. *Biomed Res Int*. 2020;2020. doi:10.1155/2020/4795140

13. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep Learning for Drug-Induced Liver Injury. *J Chem Inf Model*. Published online 2015. doi:10.1021/acs.jcim.5b00238

14. Kohonen P, Parkkinen JA, Willighagen EL, et al. A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury. *Nat Commun*. Published online 2017. doi:10.1038/ncomms15932

15. Chierici M, Francescatto M, Bussola N, Jurman G, Furlanello C. Predictability of drug-induced liver injury by machine learning. *Biol Direct*. Published online 2020. doi:10.1186/s13062-020-0259-4

16. Sumsion GR, Bradshaw MS, Beales JT, et al. Diverse approaches to predicting drug-induced liver injury using gene-expression profiles. *Biol Direct*. 2020;15(1). doi:10.1186/s13062-019-0257-6

17. Lesiński W, Mnich K, Golińska AK, Rudnicki WR. Integration of human cell lines gene expression and chemical properties of drugs for Drug Induced Liver Injury prediction. *Biol Direct*. 2021;16(1):1-12. doi:10.1186/s13062-020-00286-z

18. Liu A, Walter M, Wright P, et al. Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure. *Biol Direct*. 2021;16(1):1-15. doi:10.1186/s13062-020-00285-0

19. Aguirre-Plans J, Piñero J, Souza T, et al. An ensemble learning approach for modeling the systems biology of drug-induced injury. *Biol Direct*. 2021;16(1):1-14. doi:10.1186/s13062-020-00288-x

20. Chen M, Borlak J, Tong W. A Model to predict severity of drug-induced liver injury in humans. *Hepatology*. 2016;64(3):931-940. doi:https://doi.org/10.1002/hep.28678

21. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*. Published online 2017. doi:10.1016/j.cell.2017.10.049

22. Hong H, Xie Q, Ge W, et al. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J Chem Inf Model*. Published online 2008. doi:10.1021/ci800038f

23. US Food and Drug Administration (FDA) Silver Spring: MD: US Department of Health and Human Services, Food and Drug Administration; 2021. FDA Adverse Event Reporting System. Available from: https://open.fda.gov/data/faers/.

24. Huang R, Xia M, Sakamuru S, et al. Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat Commun*. Published online 2016. doi:10.1038/ncomms10425

25. Lin S. Space oriented rank-based data integration. *Stat Appl Genet Mol Biol*. 2010;9(1). doi:10.2202/1544-6115.1534

26. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A*. 2010;107(33):14621-14626. doi:10.1073/pnas.1000138107

27. Almazroo OA, Miah MK, Venkataramanan R. Drug Metabolism in the Liver. *Clin Liver Dis*. 2017;21(1):1-20. doi:10.1016/j.cld.2016.08.001

28. Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. Modeling liver-related adverse effects of drugs using k nearest neighbor quantitative structure-activity relationship method. *Chem Res Toxicol*. 2010;23(4):724-732. doi:10.1021/tx900451r

29. Villanueva-Paz M, Morán L, López-Alcántara N, et al. Oxidative Stress in Drug-Induced Liver Injury (DILI): From Mechanisms to Biomarkers for Use in Clinical Practice. Published online 2021. doi:10.3390/antiox10030390

30. Hepatotoxicity D, Lee WM. Drug-Induced Hepatotoxicity. Published online 2003:474-485.

31. Roh JS, Sohn DH. Damage-associated molecular patterns in inflammatory diseases. *Immune Netw*. 2018;18(4). doi:10.4110/in.2018.18.e27

32. Pirmohamed M, Naisbitt DJ, Gordon F, Park BK. The danger hypothesis - Potential role in idiosyncratic drug reactions. *Toxicology*. 2002;181-182:55-63. doi:10.1016/S0300-483X(02)00255-X

33. Uetrecht J. Idiosyncratic drug reactions: Current understanding. *Annu Rev Pharmacol Toxicol*. 2007;47:513-539. doi:10.1146/annurev.pharmtox.47.120505.105150

34. Liu ZX, Govindarajan S, Kaplowitz N. Innate immune system plays a critical role in determining the progression and severity of acetaminophen hepatotoxicity. *Gastroenterology*. 2004;127(6):1760-1774. doi:10.1053/j.gastro.2004.08.053

35. Ishida Y, Kondo T, Ohshima T, Fujiwara H, Iwakura Y, Mukaida N. A pivotal involvement of IFN-’ in the pathogenesis of acetaminophen-induced acute liver injury. *FASEB J*. 2002;16(10):1227-1236. doi:10.1096/fj.02-0046com

36. Whitington PF, Freese DK, Alonso EM, Jane Schwarzenberg S, Sharp HL. Clinical and biochemical findings in progressive familial intrahepatic cholestasis. *J Pediatr Gastroenterol Nutr*. 1994;18(2):134-141. doi:10.1097/00005176-199402000-00003

37. Krähenbühl S, Talos C, Fischer S, Reichen J. Toxicity of bile acids on the electron transport chain of isolated rat liver mitochondria. *Hepatology*. 1994;19(2):471-479. doi:10.1002/hep.1840190228

38. Tujios S, Fontana RJ. Mechanisms of drug-induced liver injury: From bedside to bench. *Nat Rev Gastroenterol Hepatol*. 2011;8(4):202-211. doi:10.1038/nrgastro.2011.22

39. Miura Y, Imamura CK, Fukunaga K, et al. Sunitinib-induced severe toxicities in a japanese patient with the ABCG2 421 AA genotype. *BMC Cancer*. 2014;14(1):1-6. doi:10.1186/1471-2407-14-964

40. Dawson S, Stahl S, Paul N, Barber J, Kenna JG. In vitro inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. *Drug Metab Dispos*. 2012;40(1):130-138. doi:10.1124/dmd.111.040758

41. Morgan RE, Trauner M, van Staden CJ, et al. Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. *Toxicol Sci*. 2010;118(2):485-500. doi:10.1093/toxsci/kfq269

42. Funk C, Ponelle C, Scheuermann G, Pantze M. Cholestatic potential of troglitazone as a possible factor contributing to troglitazone-induced hepatotoxicity: In vivo and in vitro interaction at the canalicular bile salt export pump (Bsep) in the rat. *Mol Pharmacol*. 2001;59(3):627-635. doi:10.1124/mol.59.3.627

43. Fattinger K, Funk C, Pantze M, et al. The endothelin antagonist bosentan inhibits the canalicular bile salt export pump: A potential mechanism for hepatic adverse reactions. *Clin Pharmacol Ther*. 2001;69(4):223-231. doi:10.1067/mcp.2001.114667

44. Maddrey WC. Hepatotoxicity: The adverse effects of drugs and other chemicals on the liver. *Gastroenterology*. 2000;118(5):984-985. doi:10.1016/s0016-5085(00)70192-2

45. Donnelly MC, Davidson JS, Martin K, Baird A, Hayes PC, Simpson KJ. Acute liver failure in Scotland: changes in aetiology and outcomes over time (the Scottish Look-Back Study). *Aliment Pharmacol Ther*. 2017;45(6):833-843. doi:10.1111/apt.13943

46. Zimmerman HJ, Maddrey WC. Acetaminophen (paracetamol) hepatotoxicity with regular intake of alcohol: Analysis of instances of therapeutic misadventure. *Hepatology*. 1995;22(3):767-773. doi:10.1002/hep.1840220312

47. Reuben A, Tillman H, Fontana RJ, et al. Outcomes in adults with acute liver failure between 1998 and 2013: An observational cohort study. *Ann Intern Med*. 2016;164(11):724-732. doi:10.7326/M15-2211

48. Watkins PB, Kaplowitz N, Slattery JT, et al. Aminotransferase elevations in healthy adults receiving 4 grams of acetaminophen daily: A randomized controlled trial. *J Am Med Assoc*. 2006;296(1):87-93. doi:10.1001/jama.296.1.87

49. Chen M, Borlak J, Tong W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology*. 2013;58(1):388-396. doi:10.1002/hep.26208

50. Kaplowitz N. Idiosyncratic drug hepatotoxicity. *Nat Rev Drug Discov*. 2005;4(6):489-499. doi:10.1038/nrd1750

51. Björnsson ES, Bergmann OM, Björnsson HK, Kvaran RB, Olafsson S. Incidence, presentation, and outcomes in patients with drug-induced liver injury in the general population of iceland. *Gastroenterology*. 2013;144(7). doi:10.1053/j.gastro.2013.02.006

52. Chalasani N, Bonkovsky HL, Fontana R, et al. Features and outcomes of 899 patients with drug-induced liver injury: The DILIN prospective study. *Gastroenterology*. 2015;148(7):1340-1352.e7. doi:10.1053/j.gastro.2015.03.006

53. Andrade RJ, Lucena MI, Fernández MC, et al. Drug-induced liver injury: An analysis of 461 incidences submitted to the Spanish registry over a 10-year period. *Gastroenterology*. 2005;129(2):512-521. doi:10.1016/j.gastro.2005.05.006

54. Wei G, Bergquist A, Broomé U, et al. Acute liver failure in Sweden: Etiology and outcome. *J Intern Med*. 2007;262(3):393-401. doi:10.1111/j.1365-2796.2007.01818.x

55. Reuben A, Koch DG, Lee WM. Drug-induced acute liver failure: Results of a U.S. multicenter, prospective study. *Hepatology*. 2010;52(6):2065-2076. doi:10.1002/hep.23937

56. Schmeltzer PA, Kosinski AS, Kleiner DE, et al. Liver injury from nonsteroidal anti-inflammatory drugs in the United States. *Liver Int*. 2016;36(4):603-609. doi:10.1111/liv.13032

57. de Boer YS, Kosinski AS, Urban TJ, et al. Features of Autoimmune Hepatitis in Patients With Drug-induced Liver Injury. *Clin Gastroenterol Hepatol*. 2017;15(1):103-112.e2. doi:10.1016/j.cgh.2016.05.043

58. Gancheva D, Varbanov G. Cholestatic hepatites. *Bulg Med*. 1997;5(5-6):3-5. Accessed June 19, 2021. https://www.ncbi.nlm.nih.gov/books/NBK548914/

59. Bonkovsky HL, Kleiner DE, Gu J, et al. Clinical presentations and outcomes of bile duct loss caused by drugs and herbal and dietary supplements. *Hepatology*. 2017;65(4):1267-1277. doi:10.1002/hep.28967

60. Ahmad J, Odin JA, Hayashi PH, et al. Identification and Characterization of Fenofibrate-Induced Liver Injury. *Dig Dis Sci*. 2017;62(12):3596-3604. doi:10.1007/s10620-017-4812-7

61. Fontana RJ, Cirulli ET, Gu J, et al. The role of HLA-A*33:01 in patients with cholestatic hepatitis attributed to terbinafine. *J Hepatol*. 2018;69(6):1317-1325. doi:10.1016/j.jhep.2018.08.004

62. Björnsson ES, Gu J, Kleiner DE, Chalasani N, Hayashi PH, Hoofnagle JH. Azathioprine and 6-Mercaptopurine-induced Liver Injury. *J Clin Gastroenterol*. 2017;51(1):63-69. doi:10.1097/MCG.0000000000000568

63. Grant LM, Kleiner DE, Conjeevaram HS, Vuppalanchi R, Lee WM. Clinical and histological features of idiosyncratic acute liver injury caused by temozolomide. *Dig Dis Sci*. 2013;58(5):1415-1421. doi:10.1007/s10620-012-2493-9

64. deLemos AS, Ghabril M, Rockey DC, et al. Amoxicillin–Clavulanate-Induced Liver Injury. *Dig Dis Sci*. 2016;61(8):2406-2416. doi:10.1007/s10620-016-4121-6

65. Alqahtani SA, Kleiner DE, Ghabril M, Gu J, Hoofnagle JH, Rockey DC. Identification and characterization of cefazolin-induced liver injury. *Clin Gastroenterol Hepatol*. 2015;13(7):1328-1336.e2. doi:10.1016/j.cgh.2014.11.036

66. Fontana RJ, Seeff LB, Andrade RJ, et al. Standardization of nomenclature and causality assessment in drug-induced liver injury: Summary of a clinical research workshop. In: *Hepatology*. Vol 52. Hepatology; 2010:730-742. doi:10.1002/hep.23696

67. Meunier L, Larrey D. Drug-induced liver injury: Biomarkers, requirements, candidates, and validation. *Front Pharmacol*. 2019;10:1482. doi:10.3389/fphar.2019.01482

68. Mayoral W, Lewis JH, Zimmerman H. Drug-induced liver disease. *Curr Opin Gastroenterol*. 1999;15(3). https://journals.lww.com/co-gastroenterology/Fulltext/1999/05000/Drug_induced_liver_disease.5.aspx

69. Drug-Induced Liver Injury: Premarketing Clinical Evaluation | FDA. Accessed June 19, 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/drug-induced-liver-injury-premarketing-clinical-evaluation

70. Avigan MI, Bjornsson ES, Pasanen M, et al. Liver Safety Assessment: Required Data Elements and Best Practices for Data Collection and Standardization in Clinical Trials. *Drug Saf*. 2014;37(1):19-31. doi:10.1007/s40264-014-0183-6

71. Temple R. HY’s law: Predicting serious hepatotoxicity. *Pharmacoepidemiol Drug Saf*. 2006;15(4):241-243. doi:10.1002/pds.1211

72. Aithal GP, Watkins PB, Andrade RJ, et al. Case definition and phenotype standardization in drug-induced liver injury. *Clin Pharmacol Ther*. 2011;89(6):806-815. doi:10.1038/clpt.2011.58

73. Church RJ, Watkins PB. The transformation in biomarker detection and management of drug-induced liver injury. *Liver Int*. 2017;37(11):1582-1590. doi:10.1111/liv.13441

74. Fda, Cder. *Letter of Support for Drug-Induced Liver Injury (DILI) Biomarker(s), July 25, 2016.*; 2016. Accessed June 20, 2021. http://www.imi-safe-t.eud.

75. Church RJ, Kullak-Ublick GA, Aubrecht J, et al. Candidate biomarkers for the diagnosis and prognosis of drug-induced liver injury: An international collaborative effort. *Hepatology*. 2019;69(2):760-773. doi:10.1002/hep.29802

76. Béquignon OJM, Pawar G, van de Water B, Cronin MTD, van Westen GJP. Computational Approaches for Drug-Induced Liver Injury (DILI) Prediction: State of the Art and Challenges. In: *Systems Medicine*. Elsevier; 2021:308-329. doi:10.1016/b978-0-12-801238-3.11535-1

77. Ai H, Chen W, Zhang L, et al. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol Sci*. 2018;165(1):100-107. doi:10.1093/toxsci/kfy121

78. Fraser K, Bruckner DM, Dordick JS. Advancing Predictive Hepatotoxicity at the Intersection of Experimental, in Silico, and Artificial Intelligence Technologies. *Chem Res Toxicol*. 2018;31(6):412-430. doi:10.1021/acs.chemrestox.8b00054

79. Saini N, Bakshi S, Sharma S. In-silico approach for drug induced liver injury prediction: Recent advances. *Toxicol Lett*. 2018;295:288-295. doi:10.1016/j.toxlet.2018.06.1216

80. Cruz-Monteagudo M, Cordeiro MNDS, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity. *J Comput Chem*. 2008;29(4):533-549. doi:10.1002/jcc.20812

81. Van Erp M, Vuurpijl L, Schomaker L. An overview and comparison of voting methods for pattern recognition. In: *Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR*. ; 2002:195-200. doi:10.1109/IWFHR.2002.1030908

82. Kuncheva LI. A theoretical study on six classifier fusion strategies. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(2):281-286. doi:10.1109/34.982906

83. Guo K, McGregor B (2021). VennDetail: A package for visualization and extract details. R package version 1.8.0, https://github.com/hurlab/VennDetail.

84. Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. doi:10.1613/jair.953

85. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. *Data Min Knowl Discov*. Published online 2014. doi:10.1007/s10618-012-0295-5

86. Li F, Cao Y, Han L, et al. Geneexpressionsignature: An r package for discovering functional connections using gene expression signatures. *Omi A J Integr Biol*. Published online 2013. doi:10.1089/omi.2012.0087

87. Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Am. Math. Soc.* 7(1), 48-48. doi:10.2307/2033241

88. Saari DG. Mathematical structure of voting paradoxes: II. Positional voting. *Econ Theory*. Published online 2000. doi:10.1007/s001990050002

89. Saari DG. Mathematical structure of voting paradoxes: I. Pairwise votes. *Econ Theory*. Published online 2000. doi:10.1007/s001990050001

90. Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Softw.* 28, 1–26. doi:10.18637/jss.v028.i05.

91. R Core Team (2020). R: A language and environment for statistical computing. *R A Lang Environ Stat Comput R Found Stat Comput Vienna, Austria*. Published online 2020.

92. Cox DR. The Regression Analysis of Binary Sequences. *J R Stat Soc Ser B*. 1958;20(2):215-232. doi:10.1111/j.2517-6161.1958.tb00292.x

93. LDA (Linear Discriminant Analysis). In: *Encyclopedia of Biometrics*. Springer US; 2009:899-899. doi:10.1007/978-0-387-73003-5_349

94. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81-106. doi:10.1007/BF00116251

95. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

96. Hand DJ, Yu K. Idiot’s Bayes: Not So Stupid after All? *Int Stat Rev / Rev Int Stat*. 2001;69(3):385. doi:10.2307/1403452

97. Ruta D, Gabrys B. Classifier selection for majority voting. *Inf Fusion*. 2005;6(1):63-81. doi:10.1016/j.inffus.2004.04.008

98. Lin X, Yacoub S, Burns J, Simske S. Performance analysis of pattern classifier combination by plurality voting. *Pattern Recognit Lett*. 2003;24(12):1959-1969. doi:10.1016/S0167-8655(03)00035-7

99. Aguirre-Plans J, Piñero J, Souza T, et al. An ensemble learning approach for modeling the systems biology of drug-induced injury. *Biol Direct*. 2021;16(1):1-14. doi:10.1186/s13062-020-00288-x

100. Mu X, Lu J, Watta P, Hassoun MH. Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In: *Proceedings of the International Joint Conference on Neural Networks*. ; 2009:2168-2171. doi:10.1109/IJCNN.2009.5178708

101. Walker PA, Ryder S, Lavado A, Dilworth C, Riley RJ. The evolution of strategies to minimise the risk of human drug-induced liver injury (DILI) in drug discovery and development. *Arch Toxicol*. 2020;94(8):2559-2585. doi:10.1007/s00204-020-02763-w

APPENDIX A

Manuscript Details

The content of this thesis is from an endorsed manuscript, "Temidayo Adeluwa, Brett Anthony McGregor, Kai Guo and Junguk Hur. **Predicting Drug-Induced Liver Injury using Machine Learning on a Diverse Set of Predictors**” submitted to the Frontiers in Pharmacology.

Acknowledgments

We would like to thank the CAMDA committee for organizing this challenge and providing the opportunity to present our results at ISMB 2020.

Contribution to the Field Statement

Drug-induced liver injury, or DILI, is an umbrella term for adverse reactions that affect the liver, and which are caused by the use of medications, dietary supplements, and other xenobiotics. These reactions may be caused by exposure to toxic doses of drug compounds or may present as unpredictable and unintended consequences of drug use even within non-toxic doses. Additionally, in the process of drug development and clinical trials, it is difficult to determine if a new chemical entity can cause DILI. Currently, DILI biomarkers are unspecific for drug-related hepatocellular injuries. Therefore, there is the need to develop novel approaches to predict DILI using drug-related information. In this study, we present an evaluation of models built on a number of datasets using various traditional machine learning algorithms. These data include gene expression data, toxicology data, drug structure information, and reported cases of adverse events. Our study, consistent with other studies in this domain, showed that these data may not be sufficient to classify DILI types, and that to improve the current status of DILI unpredictability, there is a need to consider larger and more sensitive DILI-related information.

Data Availability

Data are available for download as provided by the CAMDA organizers at http://camda2020.bioinf.jku.at/doku.php/contest_dataset. The full processing code of the data for the results obtained in this manuscript can be found at https://github.com/hurlab/CAMDA-Challenge-2020-Drug-Induced-Liver-Injury.

APPENDIX B

LIST OF ABBREVIATIONS

Abbreviation	**Meaning**
DILI	Drug-Induced Liver Injury
CAMDA	Critical Assessment Of Massive Drug Analysis
MCC	Matthews Correlation Coefficient
SMOTE	Synthetic Minority Over-Sampling Technique
ROSE	Random Over-Sampling Examples
FAERS	Food And Drug Administration Adverse Event Reporting System
TOX21	Toxicology In The 21st Century
MOLD2	Molecular Descriptors From 2D Structures
AUC	Area Under The Curve
LR	Logistic Regression
RPART	Recursive Partitioning And Regression Trees
GLM	Generalized Logistic Model
RF	Random Forest
SVM	Support Vector Machine
NNET	Neural Network
ADR	Adverse Drug Reaction
FDA	Food And Drug Administration
TG-GATE	Toxicogenomics Project-Genomics Assisted Toxicity Evaluation Systems
NCI	National Cancer Institute
LTKB	Liver Toxicity Knowledge Base
ISMB	Intelligent Systems For Molecular Biology
CMap	Connectivity Map
SMILES	Simplified Molecular Input Line Entry System
SIDER	Side Effect Resource
PHH	Primary Human Hepatocytes Cell Line
HEPG2	Human Liver Cancer Cell Line
MCF7	Breast Cancer Cell Line
HA1E	Immortalized Kidney Cells
A375	Human Skin Melanoma Cell Line
PC3	Adenocarcinoma Cell Line
CYP	Cytochrome P
NAT	N-Acetyltransferase
UGT	UDP-Glucuronosyltransferase
TPMT	Thiopurine S-Methyltransferase
COMT	Catechol O-Methyltransferase
GST	Glutathione S-Transferase
SULT	Sulfotransferase
ROS	Reactive Oxygen Species
ER	Endoplasmic Reticulum

DAMP	Damage-Associated Molecular Patterns
ATP	Adenosine Triphosphate
HMGB1	High Mobility Group Box 1
DNA	Deoxyribonucleic Acid
IFN	Interferon
HLA	Human Leucocyte Antigen
BSEP	Bile Salt Export Pump
APC	Antigen Presenting Cell
PRR	Pattern Recognition Receptor
NSAID	Non-Steroidal Anti-Inflammatory Drug
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
ALP	Alkaline Phosphatase
GGT	Gamma-Glutamyl Transferase
GLDH	Glutamate Dehydrogenase
miRNA-122	Micro-Ribonucleic Acid - 122
DT	Decision Trees
NB	Naïve Bayes
ROC	Receiver Operating Characteristic
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
TPR	True Positive Rate
TNR	True Negative Rate