IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 5, 2021, accepted February 1, 2021, date of publication February 4, 2021, date of current version February 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057196

An Ensemble Learning Approach for Enhanced Classification of Patients With Hepatitis and Cirrhosis

DAVIDE CHICCO[1] **AND GIUSEPPE JURMAN**[2]
[1]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON M5T 3M7, Canada
[2]Predictive Models for Biomedicine and Environment Unit, Fondazione Bruno Kessler, 38123 Trento, Italy

Corresponding author: Davide Chicco (davidechicco@davidechicco.it)

ABSTRACT Hepatitis C is an infectious disease that affects more than 70 million people worldwide, even killing 400 thousand of them annually. To better understand this disease and its prognosis, medical doctors can take advantage of the electronic health records (EHRs) of patients, which contain data that computer-based approaches built on statistics and computational intelligence can process to unveil new discoveries and trends otherwise unnoticeable by physicians. In this study, we analyze EHRs of 540 healthy controls and 75 patients diagnosed with hepatitis C, and use machine learning classifiers to predict their diagnosis. We employ the top classifier (Random Forests) to detect the most diagnostic variables for hepatitis C, that result being aspartate aminotransferase (AST) and alanine aminotransferase (ALT). These two enzyme levels are also employed by physicians in the AST/ALT ratio, a traditional measure commonly employed in gastroenterology and hepatology. We apply the same approach to a validation dataset of 123 patients with hepatitis C and cirrhosis, and the same two variables arose as most relevant. We therefore compared our approach with the AST/ALT ratio, and noticed that our two-features ensemble learning model outperforms the traditional AST/ALT ratio on both datasets. Our results confirm the usefulness of ensemble machine learning for hepatitis C and cirrhosis diagnosis prediction. Moreover, our discoveries can have an impact on clinical practice, helping physicians predict diagnoses of patients at risk of hepatitis C and cirrhosis more precisely.

INDEX TERMS AST/ALT ratio, cirrhosis, electronic health records, clinical feature ranking, fibrosis, hepatitis C, minimal health records, aspartate aminotransferase, alanine aminotransferase.

I. INTRODUCTION

As reported by the World Health Organization (WHO), 71 million people worldwide suffer from chronic hepatitis C, with about 400 thousand related deaths [1]. Chronic hepatitis C is a lifelong condition with no effective vaccine, often leading to the onset of severe conditions such as liver fibrosis and cirrhosis, and hepatocellular carcinoma [2]. Liver fibrosis is the results of the wound healing response to tissue damage caused by chronic hepatitis C, while cirrhosis is an advanced stage of liver fibrosis with distortion of the hepatic vasculature and architecture [3].

In standard clinical practice, an immediate, quick and non-invasive assessment of the diagnosis of potential liver damage can be indirectly derived by measuring the blood levels of a few enzymes, known as the liver tests [4], [5]. Although no test can accurately identify liver disease, nonetheless liver tests can provide essential insights towards the formulation of a diagnosis [6], usually by mean of rule-based or arithmetic-based formulæ called scores. In particular, among all the deterministic scores appeared in the literature, the DeRitis ratio still stands the test of time and remains a useful indicator of liver disease more than 60 years after its definition [7]. Defined as the ratio between the serum levels of aspartate transaminase (AST) and alanine transaminase (ALT), the DeRitis indicator owes its success to both the simple formula and the relatively high accuracy, despite its several limitations [8]. In the last few years, the ever growing public availability of electronic health records (EHRs) has paved the way for searching stochastic alternatives to the rule-based scores by machine learning approaches (see for instance [9] as a very recent example). In Sec. II we present a wider

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

overview of the landscape of the learning models appeared in the literature.

The current manuscript aims at contributing to the aforementioned research line by demonstrating the effectiveness of three classical machine learning algorithms in binary and multi-class tasks on the Lichtinghagen (discovery) cohort EHR dataset collecting 14 features for 615 individuals [10], [11] and on the Wu (validation) dataset, with 123 patients described by 13 features [12]. The contributed findings here are threefold. In detail, as a first results, Random Forests achieve good performances in both recognizing patients affected by liver diseases and, at the same time, discriminating among three conditions with increasing severity, namely hepatitis C, liver fibrosis and liver cirrhosis. Interestingly, due to the natural ranking among the three diseases, we solve the multi-class task through a regression strategy. As a second major result, two different features ranking procedures, namely univariate traditional statistical tests and Random Forests, both identify ALT and AST blood levels as the key predictive factors to discriminate between healthy controls and patients with hepatitis C in both datasets, confirming the strong link between these two enzymes and the occurrence of a liver condition. As a third and main result, we show that the two-features (AST, ALT) ensemble learning model is able to predict with high accuracy patients affected by cirrhosis and hepatitis, in both datasets, even outperforming the performance achieved by the DeRitis ratio. As a take home message, we can summarize the provided contribution stating that, in line with current practice, ALT and AST are top discriminating features in detecting liver conditions and that their predictive ability increases when used in a bivariate model rather than univariately as in the DeRitis ratio.

Finally, it is worth mentioning that to warrant reproducibility and generalization of the study we adopt a resampling strategy as indicated by the United States Food and Drug Administration (US-FDA) led initiatives MicroArray / Sequencing Quality Control (MAQC/SEQC) [13]–[15].

We organize the rest of the article as follows. After this Introduction, we briefly outline the current literature on the topic (RELATED WORKS section), then we describe the datasets we analyzed (DATASETS section) and the methods we employed (METHODS section). Finally, we report the results (RESULTS section), discuss them and outline some conclusions and future directions (DISCUSSION section).

II. RELATED WORKS

A. AST AND ALT

Aminotransferases are enzymes used by the liver to produce glycogen. Aspartate aminotransferase (AST) is present in liver, brain, pancreas, heart, kidneys, lungs, and skeletal muscles and, when damage occurs to these tissues, bloodstream AST levels increase. Alanine aminotransferase (ALT) is instead found primarily in the liver, and high ALT values are a direct indication of a liver injury. Measure of AST and ALT bloodstream values is the paramount component of the routine Liver Function Tests (LFTs), whose main is the identification of a patient's primary disorder as of hepatitic or cholestatic source [6], [16]. In clinical practice, ALT and AST values have always been the most commonly indicators in both clinical diagnosis and research involving liver damages such as viral hepatitis, fibrosis, cirrhosis, NAFLD, NASH and HCC [17], [18]. Nonetheless, their proper use and interpretation is not straightforwards [19], [20], and several studies have pointed out the limitations of serum aminotransferases for early detection of liver injury and patient outcome [21]; in particular, both have poor prognostic utility in acute liver injury and liver failure and they do not represent reliable measures of liver functions, since they may be influenced by a wide range of non-hepatic factors [22], [23]. Such considerations are leading to deeper investigations into AST and ALT dynamics reflecting into more clinically accurate insights [24], as well as to the search for alternative biomarkers [25], even coupled with additional non-invasive testing modalities, for example US/MRI/CT imaging, elastography and GWAS genomics [26]. As a final note, being not liver specific makes AST and ALT involved in very heterogeneous processes: for instance, they play a role in the detection of metabolic disturbances in the pathophysiology of Alzheimer's Disease, due to their consistent associations with cognitive performance [27].

B. LIVER SCORES AND GUIDELINES

To simplify the assessment of liver damage from LFTs and other clinical variables, a number of scoring systems have been developed with diagnostic, prognostic and therapeutic purposes: they are rule-based or defined by a simple arithmetic formula, with lab test values as their inputs [28], [29]. Among them, MELD [30] and Child-Pugh [31] and their recent improvements are the elective clinical tool in liver disease assessment. However, despite their widespread used, their accuracy is still debated [32]–[34], and also their relative effectiveness has not been sorted out yet, with the outcome of several comparative studies remaining controversial [35]. Being this a still unsettled issue is confirmed by the recent publication in the literature of proposal for further variants of the original scores [36]–[38]. In alternative to scores, a number of guidelines have been published by several institutions to drive clinicians in the dealing with patients showing abnormal LFTs results. However, such guidelines are often complex and time consuming, especially when combined with demographic and clinical data aimed at producing an electronic diagnosis and even a management plan [39]–[41].

C. THE DeRitis RATIO

As anticipated in the Introduction, the simplest score still rules against its competitors despite its age. Originally introduced in 1957 [7] and improved the following year [42], the DeRitis ratio conveys into a single figure AST/ALT the relevant association of two aminotransferases with liver damages, providing useful diagnostic and prognostic insights. An immediate and essential indicator for the clinicians,

TABLE 1. Meaning and measurement units of the clinical features of the discovery cohort Lichtinghagen dataset [10], [11]. IU: international units. L: liter. mg: milligram. g: gram. dL: deciliter. Some of these meanings were described by the authors in the original dataset article [10], and other ones were confirmed by one of the authors (G.H.) privately via email.

name	meaning	measurement unit
age	age of the patient	years
ALB	albumin level in the blood	g/L
ALP	alkaline phosphatase level in the blood	IU/L
ALT	alanine aminotransferase level in the blood	IU/L
AST	aspartate aminotransferase level in the blood	IU/L
BIL	bilirubin level in the blood	mg/dL
CHE	choline esterase level in the blood	IU/L
CHOL	cholesterol level in the blood	milligrams/dL
CREA	serum creatinine level in the blood	mg/dL
GGT	gamma-glutamyl-transferase level in the blood	IU/L
PROT	total protein level in the blood	g/L
sex	sex of the patient	binary
[target] binary condition	0: healthy control; 1: hepatitis C	boolean
[target] disease category	0: healthy control; 1: hepatitis C; 2: hepatitis C and fibrosis; 3: hepatitis C, fibrosis, and cirrhosis	category

DeRitis ratio has been widely used throughout the years also in the research setting, as confirmed by the several publications appeared and still appearing in the literature [43]–[51]. The association of AST (mainly) and ALT with biological processes involving organs other than the liver yields that the DeRitis ratio is an useful indicator for a range of different pathologies [52], [53], even including Covid19 [54], [55]. As usual, simplicity comes at a price: no generally accepted reference intervals exist for the ratio, making it hard to define an healthy interval when transaminases are abnormal [56]. Further, Reedy and colleagues [57] pointed out that the AST/ALT ratio can be useful for cirrhosis suggestion but it should not be employed for diagnosis, while in [8] limitations are found in the DeRitis ratio when predicting liver fibrosis. However, all these bounds are not undermining the fact that, to date, DeRitis ratio still represents a benchmark to compare against for any novel scoring system or predictive algorithm being proposed.

D. MACHINE LEARNING STUDIES

As for other clinical applications, the novel frontier of diagnostic and prognostic methods is represented by the predictive models provided by machine learning approaches. Such data-driven paradigm – opposite to the classical rule- or formula-based – has been boosted by the recent introduction of electronic health records (EHRs). In fact, EHRs play nowadays a major role as a data source, not only for standard shallow computational intelligence approaches but also for the growing literature trend associated with the quick development of novel deep learning algorithms [58], applied also imaging data [59], [60]. Although more classical data sources are still used (for example, insurance claims [61]), EHRs represent now a solid base for building learning models upon, both for diagnostic purposes in chronic hepatitis C [9], [62]–[64] and in derived diseases such as liver cancer [65]. Notably, in the last few years several research groups worldwide have been actively involved in assessing and predicting the progression of chronic hepatitis C into fibrosis first and then cirrhosis, proposing several methodological alternatives to carefully stage such a progressively deteriorating condition, both by classic shallow models [61], [66]–[69] as well as by more advanced deep learning approaches such as Recurrent Neural Networks [70], marking another domain where Artificial Intelligence will definitely have an impact in the near future. If better predictive performances often characterize learning based approaches, on the other hand their higher complexity (and sometimes their difficult interpretability, too) constitutes an hindrance towards their adoption in clinical practice.

The two-feature model proposed here is planned as a compromise between achieving an acceptable accuracy (for instance as compared to the DeRitis performance) but at the same time being extremely simple, usable and easily interpretable (being based on ALT and AST only) by the practitioners.

III. DATASETS

A. DISCOVERY COHORT

In this study, we analyzed a dataset of electronic health records of 615 subjects whose blood exam data were collected at Hannover Medical School (Hannover, Germany, EU) and publicly released in 2020 [10], [11]. We consider this dataset our *discovery cohort*. The clinical records of the subjects contain 14 features (Table 1), including one indicating the disease condition of the patient (disease category) and one derived factor dividing the subjects between healthy controls and subjects (binary condition). Two of the 14 clinical variables represent biological information (sex and age), while other 10 represent typical blood exam measurements, such as levels of albumin, alkaline phosphatase, bilirubin, and others (Table 1).

The dataset contains 238 men and 377 women, with an average age of 47 years (Table 2 and Table 3). We employed the binary condition feature as target factor for the binary

classification, and the disease category feature as target factor for the regression analysis.

Of the 615 subjects, 540 are healthy controls (87.8%) and 75 were diagnosed with hepatitis C (12.2%). The 75 patients with hepatitis C can be further divided this way:

- 24 have only hepatitis C (3.9%);
- 21 have hepatitis C and liver fibrosis (3.41%);
- 30 were diagnosed with hepatitis C, liver fibrosis, and cirrhosis (4.88%).

Regarding the binary condition prediction, our discovery cohort dataset results being strongly positively imbalanced, with 87.8% positive data instances and 12.2% negative data instances.

B. VALIDATION COHORT

To verify the feature ranking discoveries we made on the discovery cohort dataset, we analyzed another independent dataset, that we consider out *validation cohort*. This dataset was released by Wu *et al.* [12] in 2015, and contains electronic health records of 123 patients diagnosed with hepatitis C, collected at Kanazawa University in Japan (Table 4).

Among these patients, 83 have cirrhosis. This validation dataset has 13 clinical features, including one called ''cirrhosis'' that indicates if the patient has only hepatitis C (label: 0), or if she/he has both hepatitis C and cirrhosis (label: 1). Regarding dataset imbalance, we therefore can say that this dataset is positively imbalanced, with 67.48% positive data instances and 32.52% negative data instances (Table 10 and Table 11).

IV. METHODS

In this section, we first describe the methods used for binary classification and regression (BINARY CLASSIFICATION AND REGRESSION section), and the techniques employed for feature ranking (FEATURE SELECTION section). We represent our computational pipeline in a flowchart in Figure 1.

A. BINARY CLASSIFICATION AND REGRESSION

The analyzed discovery dataset [10], [11] contains 31 missing data related to 23 patients (section III). With 615 subjects and 14 clinical features, the amount of missing data corresponds then to 0.36%.

In the validation cohort dataset [12], instead, there are 26 missing data. Having 123 patients and 13 variables, the missing data are the 1.94% of the total.

To impute these missing data, we employed the Predictive Mean Matching (PMM) approach through the Multiple Imputation by Chained Equations (MICE) software package [71]. The PMM approach replaces each missing value with another artificial value generated through a regression prediction [72], [73], and is one of the most effective methods for this scope [74].

In the binary classification on the discovery cohort dataset, the goal of our algorithms was to distinguish between healthy controls (with target label 0) and patients with hepatitis C (with target label 1). In the regression analysis, instead, the scope of our algorithms was to correctly predict the target ordinal label (0: healthy control, 1: patient with hepatitis C; 2: patient with hepatitis C and liver fibrosis; 3: patient with hepatitis C, liver fibrosis, and cirrhosis), each corresponding to a different disease condition (DATASETS section).

TABLE 2. Quantitative characteristics of binary features of the discovery cohort Lichtinghagen dataset [10], [11]. sex: 1 means man and 0 means woman. binary condition: 1 means sick patient and 0 means healthy control. Both sex and binary condition have no missing value (NAs).

binary feature	value	#	%
sex	0	238	38.699
sex	1	377	61.301
[target] binary condition	0	540	87.805
[target] binary condition	1	75	12.195

We performed the binary classification analysis and the regression analysis by using three common machine learning techniques known for their effectiveness for clinical records [75]–[77]: Linear Regression [78], Decision Trees [79], and Random Forests [80].

Linear Regression is a statistical approach where a linear model associates some input variables to a response [81]; in our discovery dataset case, the response is 0 or 1 for the binary classification, and 0, 1, 2, or 3 for the regression analysis.

Decision Trees are tree-like models where each possible decision based on a specific dataset feature value represents a node, whose output becomes the input of the following node. The final decision node produces the algorithm response: 0 or 1 for the binary classification, and 0, 1, 2, or 3 for the regression analysis, in our case. Decision Trees are often employed by medical doctors and physicians, because they are easy to interpret and understand, and therefore can provide insightful decision processes for medical decision-making [82].

Random Forests are a popular ensemble learning technique where multiple Decision Trees are applied to subsets of features and data. Each decision tree generates an outcome, and then Random Forests return a final value which is indicated by the majority of the Decision Trees' outcomes [83].

For each algorithm execution, we split the dataset into 80% randomly selected data instances for the training set and 20% test set with the remaining data instances. We measured the binary classification with typical confusion matrix rates such as the Matthews correlation coefficient (MCC), the F_1 score, the receiver operating characteristic (ROC) area under the curve (AUC), and the regression results with traditional metrics such as the coefficient of determination R^2 and the mean squared error. We repeated our script execution 100 times, and reported the average scores and their corresponding standard deviation (TARGET PREDICTION section).

TABLE 3. Quantitative characteristics of real-valued features of the discovery cohort Lichtinghagen dataset [10], [11]. NAs: number of missing values. disease category: 0 means healthy control; 1 means patient with hepatitis C; 2 means patient with hepatitis C and fibrosis; 3 means patient with hepatitis C, fibrosis, and cirrhosis.

real-valued feature	median	mean	range	σ	NAs
age	47	47.408	[19, 77]	10.055	0
ALB	41.95	41.620	[14.9, 82.2]	5.781	1
ALP	66.2	68.284	[11.3, 416.6]	26.028	18
ALT	23	28.451	[0.9, 325.3]	25.470	1
AST	25.9	34.786	[10.6, 324]	33.091	0
BIL	7.3	11.397	[0.8, 254]	19.673	0
CHE	8.26	8.197	[1.42, 16.41]	2.206	0
CHOL	5.3	5.368	[1.43, 9.67]	1.133	10
CREA	77	81.288	[8, 1079.1]	49.756	0
GGT	23.3	39.533	[4.5, 650.9]	54.661	0
PROT	72.2	72.044	[44.8, 90]	5.403	1
[target] disease category	0	0.254	[0, 3]	0.742	0

TABLE 4. Meanings of the clinical features of the Wu validation cohort dataset [12].

feature	meaning	measurement unit
age	age of the patient	years
ALT	alanine aminotransferase level in the blood	IU/L
AST	aspartate aminotransferase level in the blood	IU/L
BMI	body-mass index	kg/m^2
cholesterol	cholesterol level in the blood	mg/dL
glucose	glucose level in the blood	mg/dL
HDL	high-density lipoprotein level in the blood	mg/dL
LDL	low-density lipoprotein in the blood	mg/dL
PathDiagNum	histopathology: 0 means needle biopsy and 1 means resection	binary
serogroup	0 means serotype 1 (genotypes I-II-V) and 1 means serotype 2 (genotypes III-IV)	binary
sex	0 means man and 1 means woman	binary
triglycerides	level of triglycerides in the blood	mg/dL
[target] cirrhosis	0 means patient without cirrhosis and 1 means with cirrhosis	binary

B. FEATURE SELECTION

1) UNIVARIATE STATISTICAL TESTS

In the following phase of our analysis, we decided to investigate which clinical features of the discovery cohort dataset were the most predictive of the healthy control or patient with hepatitis C status.

To this end, we first employed traditional univariate biostatistics tests, such as Mann-Whitney U test [84] for the real features and and the chi-square test [85] for the binary features.

We perform these tests between each feature and the target feature. For both the tests, the resulting p-value can range between 0 and 1. The closer the p-value is to the zero, the stronger is the relationship between the two processed features; the closer the p-value is to one, the less chance there is a significant relationship between them. Traditionally, a relationship between a variable and the target is considered *significant* if the p-value is lower than 0.05 (that is 5×10^{-2}), but in this study we decided to employ the stricter threshold of 0.005 (that is 5×10^{-3}), as recently suggested by Benjamin and colleagues [86]. The p-value obtained by each feature can be used to generate a ranking by significance, where the variables which obtained the lowest p-value will be at the top, and the variables which achieved the highest p-value will be at the bottom.

2) MACHINE LEARNING FEATURE SELECTION

After having explored the significance of the clinical variables of our datasets through biostatistics, we decided to employ the top performing method of the binary classification for feature selection purposes: Random Forests [80]. This ensemble learning technique, in fact, been shown to be effective for feature ranking in health informatics and bioinformatics [77], [87].

Random Forests compute the prediction accuracy of the model on the dataset, by removing a feature at each time, through a procedure known as recursive feature elimination (RFE). When computing the binary classification without a specific feature, Random Forests save the accuracy drop obtained with respect to the same classification made with all the variables instead, and associates this accuracy decrease to that feature. It repeats this procedure for each feature, and in the end it ranks all the variables based on the accuracy drop (mean decrease accuracy).

Random Forests provide the feature ranking measured with Gini purity decrease, too, but we prefer to stick with the accuracy decrease because it is a more reliable and stable measure [88].

For better understanding of this study, we depicted our computational pipeline in a flowchart in Figure 1.

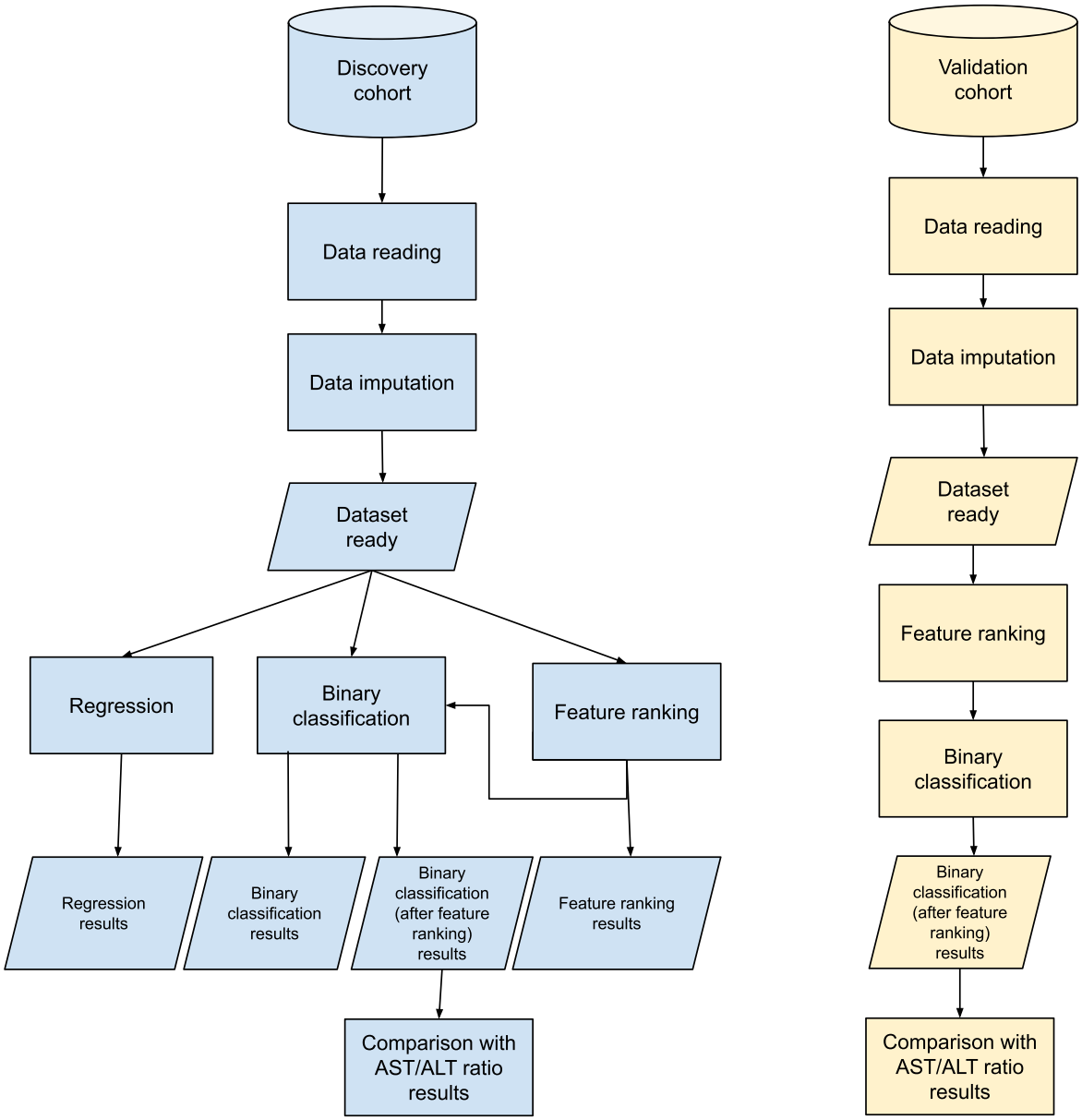


FIGURE 1. Flowchart of our computational pipeline. Cylinder shape: dataset. Rectangular shape: process. Parallelogram shape: input/output.

V. RESULTS

In this section, we first describe the results obtained for binary classification and regression (TARGET PREDICTION section), and then the results achieved for feature ranking (FEATURE RANKING section).

A. TARGET PREDICTION

We report the results achieved for the regression analysis of the patients' condition in Table 5 and for the binary classification of healthy controls and patients with hepatitis C in Table 6.

For the binary classification, we focused and ranked the results based on MCC, that is the only confusion matrix rate that generates a high score only if the classifier correctly predicts most of the data instances in proportion to the class imbalance [89], [90], rather than using other rates such as ROC AUC [91]. For the same reason, we based our regression analysis results on the coefficient of determination R^2, that takes into account the distribution of the target labels [92].

As one can notice, Random Forests were able to achieve the top results both for binary classification and regression: R^2 equal to $+0.765$ (Table 5) and MCC equal to $+0.858$ (Table 6), both scores close to $+1$, which would mean perfect prediction.

In the regression analysis, Random Forests obtained also the top scores for the other rates (MAE, MSE, and RMSE). Linear Regression was the second top performing method, with $R^2 = +0.523$ (Table 6)

TABLE 5. Results of the regression analysis made with machine learning classifiers, including standard deviation, on the discovery cohort dataset. R^2: coefficient of determination. RMSE: root mean squared error. MAE: mean absolute error. MSE: mean square error. RMSE, MAE, MSE: best value = 0.00 and worst value = $+\infty$. R^2: best value = +1.00 and worst value = $-\infty$. We highligthed in blue and with an asterisk * the top result each rate. mean: average score of 100 executions. σ: standard deviation of 100 executions.

method	value	R^2	MAE	MSE	RMSE
Random Forests	mean	***+0.765**	*0.139	*0.123	*0.346
Random Forests	σ	0.077	0.028	0.046	0.061
Linear Regression	mean	**+0.523**	0.279	0.247	0.489
Linear Regression	σ	0.189	0.032	0.102	0.089
Decision Tree	mean	**+0.467**	0.142	0.273	0.512
Decision Tree	σ	0.233	0.044	0.113	0.106

TABLE 6. Results of the binary classification made with machine learning classifiers, including standard deviation, on the discovery cohort dataset. MCC: Matthews correlation coefficient. MCC worst value = -1 and best value = $+1$. TPR: true positive rate, sensitivity, recall. TNR: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. Confusion matrix threshold τ: 0.5. F_1 score, accuracy, TP rate, TN rate, PPV, NPV, PR AUC, ROC AUC: worst value = 0 and best value = +1. We highlighted in blue and with an asterisk * the top result for each rate. mean: average score of 100 executions. σ: standard deviation of 100 executions.

method	value	MCC	F_1	accuracy	TPR	TNR	PPV	NPV	PR AUC	ROC AUC
Random Forests	mean	***+0.858**	*0.872	*0.971	*0.847	0.988	0.907	*0.979	*0.950	*0.985
Random Forests	σ	0.068	0.062	0.015	0.089	0.010	0.075	0.013	0.038	0.016
Decision Tree	mean	**+0.723**	0.748	0.943	0.742	0.970	0.777	0.965	0.637	0.856
Decision Tree	σ	0.096	0.091	0.019	0.140	0.017	0.106	0.021	0.112	0.068
Linear Regression	mean	**+0.604**	0.584	0.929	0.441	*0.996	*0.929	0.929	0.531	0.718
Linear Regression	σ	0.124	0.143	0.021	0.142	0.006	0.106	0.021	0.136	0.071

In the binary prediction, Random Forests also obtained the top F_1 score, accuracy, sensitivity, negative predictive value, PR AUC, and ROC AUC (Table 5). Linear regression was able to achieve the top specificity (TNR = 0.996) and top precision (PPV = 0.929). Here, the Decision Tree is the second top performing method, with MCC equal to +0.723.

B. FEATURE RANKING

1) DISCOVERY COHORT

After ensuring that machine learning can classify healthy controls and patients with hepatitis C in this EHR dataset, we decided to investigate which clinical features result more predictive for this task.

We first applied traditional biostatistics tests to highlight the relationship between each clinical variable and the binary condition target. Our results show that 9 features out of 12 resulted being statistically significant, by achieving a p-value lower than 0.005: ALB, BIL, CHOL, GGT, ALP, CHE, ALB, CREA, and ALT (Table 7). These results show that all these nine clinical variables are correlated to the binary condition.

After using traditional univariate statistical methods, we employed again machine learning for feature ranking. We applied Random Forests, which was the top performing method in the binary prediction (Table 6).

In the Random Forests feature ranking measured through the accuracy decrease, the level of aspartate aminotransferase (AST) and the level of alanine aminotransferase (ALT) resulted being the top most predictive variables (Figure 2a and Table 12). Random Forests also indicated sex and level of creatinine (CREA) in the blood as least predictive variable among the clinical records.

2) VALIDATION COHORT

To further investigate the results obtained by our feature ranking procedures on the discovery cohort dataset by Lichtinghagen *et al.* [10], [11], we decided to apply the same methods to an alternative dataset of patients with the same disease and with similar clinical features.

We found the public dataset of Wu and colleagues [12]. The Wu dataset contains data from 123 patients with hepatitis C (Table 4); from the original dataset, we extrapolated 12 clinical variables, where one of them indicates if the patient has hepatitis C and cirrhosis (label: 1) or she/he has hepatitis C without cirrhosis (label: 0).

We applied the univariate statistical tests to this dataset, between each variable and the cirrhosis target. Their results indicated the level of aspartate aminotransferase (AST) and the level of alanine aminotransferase (ALT) as statistically significant variables (Table 8).

These statistical results highlighted again the role of AST and ALT in these clinical records.

We therefore executed Random Forests on this validation dataset, for feature ranking purposes. Interestingly, the results indicated again AST and ALT levels as the most predictive variables, even for this validation dataset (Figure 2b and Table 13). On the other side, serogroup, age, body-mass index (BMI), and the triglyceride levels resulted being the least predictive factors, even providing a negative contribution to the overall prediction.

It is important to reaffirm that the feature ranking procedures (univariate statistical tests and Random Forests) were identical for both the datasets, but the target was different: in the Lichtinghagen discovery cohort, the goal was to discriminate patients with hepatitis C from healthy controls, while in

TABLE 7. Results of the univariate statistical tests to the Lichtinghagen discovery cohort dataset. We applied the Mann-Whitney *U* test between each real-valued feature and the binary condition (healthy control / patient), and the chi-square test between the ordinal feature sex and the same binary condition [10], [11]. We listed in blue and with and asterisk * all the features resulted being significant, that are the ones whose test result *p*-value is lower than 0.005.

rank	feature	p-value
Mann-Whitney U test		
1	*AST	0
2	*BIL	0
3	*CHOL	0
4	*GGT	0
5	*ALP	1.000×10^{-06}
6	*CHE	1.120×10^{-04}
7	*ALB	4.030×10^{-04}
8	*CREA	2.608×10^{-03}
9	*ALT	3.506×10^{-03}
10	PROT	4.071×10^{-02}
11	age	2.368×10^{-01}
chi-square test		
1	sex	7.996×10^{-02}

TABLE 8. Results of the univariate statistical tests to the Wu validation cohort dataset. We applied the Mann-Whitney *U* test between each real-valued feature and the binary condition (patient with hepatitis C and without cirrhosis / patient with hepatitis C and cirrhosis), and the chi-square test between the ordinal features and the same binary condition [12]. We listed in blue and with and asterisk * all the features resulted being significant, that are the ones whose test result *p*-value is lower than 0.005.

rank	feature	p-value
Mann-Whitney U test		
1	*AST	1.600×10^{-05}
2	*ALT	1.173×10^{-03}
3	*cholesterol	2.281×10^{-03}
4	LDL	1.491×10^{-02}
5	HDL	9.483×10^{-02}
6	triglyceride	1.369×10^{-01}
7	BMI	5.616×10^{-01}
8	age	5.927×10^{-01}
9	glucose	1
chi-square test		
1	sex	3.383×10^{-01}
2	serogroup	5.587×10^{-01}
3	PathDiagNum	7.891×10^{-01}

the Wu validation cohort the goal was to distinguish between patients with hepatitis C and cirrhosis from patients with hepatitis C and without cirrhosis.

C. TARGET PREDICTION USING ONLY AST AND ALT

To further verify the predictive power of ensemble machine learning applied to the AST and ALT clinical components of the medical records, we decided to perform a binary classification on both the datasets by employing only these two features, similar to what we did in another study related to heart failure [93]. In particular, our aim was to compare the results obtained by our ensemble learning method with the results achieved by the traditional AST/ALT ratio.

We divided each dataset into a training set of 80% randomly chosen data instances and a test set of the remaining 20% data instances. We then applied the Random Forests feature ranking method on the training set, following the previously described feature ranking procedure (subsection IV-B), repeated 100 times. Afterwards, we selected the top two final resulting features, which happened to be AST and ALT, trained a Random Forests classifier on the training set made only by 3 features: AST, ALT, and the binary target. We repeated this whole procedure (feature ranking and testing) 100 times for each test.

It is important to reaffirm that the binary target distinguish healthy controls from patients with hepatitis C in the discovery cohort Lichtinghagen dataset [10], [11], while it discriminates the patients with hepatitis C and cirrhosis from the patients with hepatitis C without cirrhosis in the validation cohort Wu dataset [12].

Once this two-variable model is trained, we applied it to the test set containing the same features, where the other factors have been discarded. We repeated the same procedure both for Lichtinghagen dataset [10], [11] and Wu dataset [12], and we reported the obtained results in Table 9.

As one can notice, the Random Forests model trained only on AST and ALT can predict hepatitis C in the discovery cohort dataset and cirrhosis in the validation cohort dataset, with a MCC of $+0.81$ in the first case and an MCC of $+0.275$ in the second case.

In the discovery cohort, the two-feature Random Forests model was able to obtain very high accuracy, specificity, and negative predictive value, and high MCC, F_1 score, and true positive rate In particular, its true negative rate and NPV resulted being close to 1, which would mean perfect prediction.

Regarding the validation cohort, the two-feature Random Forests model obtained high sensitivity (TPR $= 0.796$) and precision (PPV $= 0.724$), and a just sufficient negative predictive value (NPV $= 0.503$); The model, however, failed to correctly predict most of the true negative data instances, achieving a specificity of 0.389. This issue might be due to the imbalance of the validation cohort dataset (Table 10): with 32.62% negative data instances, the model might not be able to see enough data of patients without cirrhosis to learn well enough to perform a good prediction.

Our two-feature machine learning methods were able to outperform the AST/ALT ratio, on both the dataset (Table 9 and Figure 3). In the discovery cohort, our Random Forests obtained an MCC of $+0.781$, while AST/ALT ratio's MCC was $+0.771$, which means a 0.5% improvement in the $[-1, +1]$ interval of MCC.

In the validation cohort, our model attained MCC $= +0.202$, and the AST/ALT ratio obtained MCC $= +0.037$, which means a 8.25% increment in the $[-1, +1]$ interval of MCC. In addition to the MCC, the two-feature Random Forests outperformed the AST/ALT ratio on the F_1

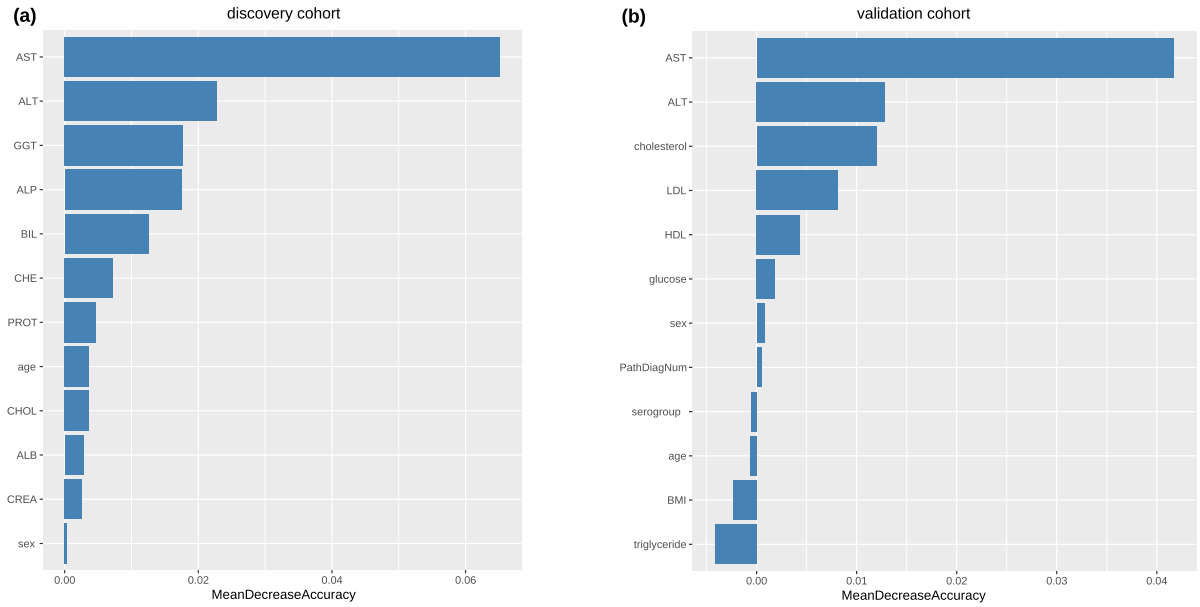


FIGURE 2. Feature ranking of the clinical variables in the discovery cohort (a) and validation cohort (b). (a): feature ranking computed through Random Forests mean decrease in accuracy, with the goal of classifying patients with hepatitis C and healthy controls, on the Lichtinghagen dataset discovery cohort [10], [11]. (b): Feature ranking computed through Random Forests mean decrease in accuracy, with the goal of classifying patients with fibrosis and patients with cirrhosis, on the Wu dataset validation cohort [12].

TABLE 9. Numerical results of the binary classification made with Random Forests after feature selection (AST and ALT) and of the classification made with the AST/ALT ratio. For the two-features Random Forests, we report the average score of 100 executions. Hepatitis C criterion: AST/ALT ratio > 3. Cirrhosis criterion among patients with hepatitis C: AST/ALT ratio > 1. RF: Random Forests. MCC: Matthews correlation coefficient. MCC worst value $= -1$ and best value $= +1$. TPR: true positive rate, sensitivity, recall. TNR: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. Confusion matrix threshold τ: 0.5. F_1 score, accuracy, TPR, TNR, PPV, NPV, PR AUC, ROC AUC: worst value $= 0$ and best value $= +1$.

method	MCC	F_1	accuracy	TPR	TNR	PPV	NPV	PR AUC	ROC AUC
discovery cohort: hepatitis C prediction									
two-features RF	**+0.781**	0.800	0.954	0.780	0.979	0.839	0.970	0.813	0.953
AST/ALT ratio	**+0.771**	0.785	0.954	0.680	0.993	0.927	0.957	0.711	0.836
validation cohort: cirrhosis prediction									
two-features RF	**+0.202**	0.752	0.658	0.796	0.389	0.724	0.503	0.775	0.638
AST/ALT ratio	**+0.037**	0.675	0.569	0.663	0.375	0.688	0.349	0.685	0.519

score, sensitivity, negative predictive value, PR AUC, and ROC AUC. It obtained approximately the same results of the AST/ALT ratio on accuracy, specificity, and was clearly outperformed by the AST/ALT ratio on precision (Table 9 and Figure 3).

In the validation cohort, our two-feature model outperformed the AST/ALT ratio in all the confusion matrix rates (Table 9 and Figure 3).

While we consider MCC as the most powerful metric to compare two classifiers, we also would like to highlight the importance of the sensitivity (true positive rate) in this scientific problem: identifying which patients have more chance to develop hepatitis C or cirrhosis, in fact, is more relevant than correctly predicting healthy controls or patients without that disease, in this setting. Patients with high likelihood of hepatitis C or cirrhosis, in fact, need more attention and have an urgent need to get under specific medical treatments. Regarding recall, our two-features Random Forests outperforms the AST/ALT ratio both in the discovery cohort and in the validation cohort, confirming its superior predictive skills.

VI. DISCUSSION

A. HEPATITIS C, FIBROSIS, AND CIRRHOSIS PREDICTION

Our results show that machine learning applied to electronic health records is capable of classifying healthy controls and patients with hepatitis C and other conditions such as fibrosis and cirrhosis, in few minutes, with low computational resources, and at a low cost.

Random Forests, in particular, resulted being effective both for the binary classification and the regression analysis. These results confirm the predictive capability of this popular ensemble learning method. Our methods can have impact on clinical activity: our techniques, in fact, can help doctors to predict if a patient will develop hepatitis C, fibrosis, or cirrhosis by analyzing just his/her medical records.

B. HEPATITIS C CLINICAL FEATURE RANKING

The two feature ranking procedures (univariate statistical tests and Random Forests) applied to the discovery cohort dataset highlighted AST, AST, GGT, ALP, and BIL as top most predictive clinical variables. Sex, instead, resulted being

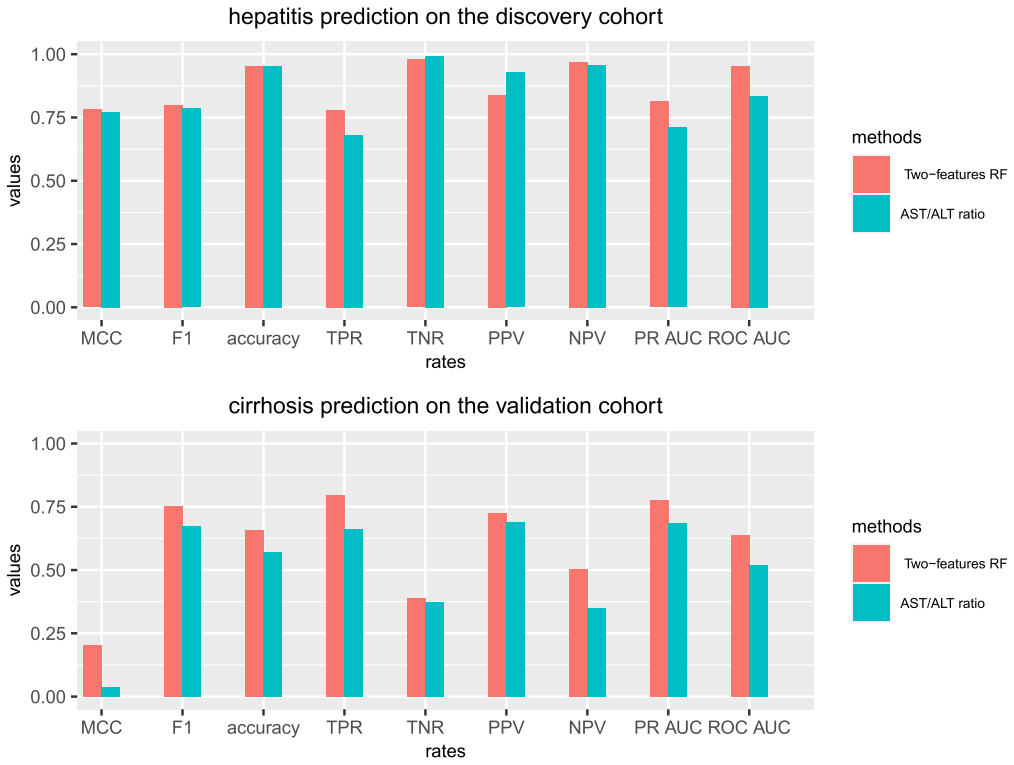


FIGURE 3. Barplots of the results of the binary classification made with Random Forests after feature selection (AST and ALT) and of the classification made with the AST/ALT ratio. Upper plot: discovery cohort [10], [11]. Lower plot: valition cohort [12]. Hepatitis C criterion: AST/ALT ratio > 3. Cirrhosis criterion among patients with hepatitis C: AST/ALT ratio > 1. RF: Random Forests. MCC: Matthews correlation coefficient. MCC worst value $= -1$ and best value $= +1$. TPR: true positive rate, sensitivity, recall. TNR: true negative rate, specificity. PR: precision-recall curve. PPV: positive predictive value, precision. NPV: negative predictive value. ROC: receiver operating characteristic curve. AUC: area under the curve. Confusion matrix threshold τ: 0.5. F_1 score, accuracy, TPR, TNR, PPV, NPV, PR AUC, ROC AUC: worst value $= 0$ and best value $= +1$.

the least important factor among the clinical records. Cholesterol was listed on top position in the statistical standing, but was ranked 9th out of 12 variables in the Random Forests ranking. These results highlighted the relevance of alanine aminotransferase (ALT) and aspartate aminotransferase (AST) to detect hepatitis C. Both these enzymes are known in the scientific literature to be associated to hepatitis C [17], [18], and employed in the AST/AST ratio for this scope [46], [48], [49].

C. CIRRHOSIS CLINICAL FEATURE RANKING

To further investigate the importance of the top features retrieved, we decided to apply the same feature ranking procedures to an external validation cohort dataset by Wu and colleagues [12], containing data of patients all having hepatitis C, where some of them have also cirrhosis. Both the univariate statistical feature ranking and the Random Forests standing listed aminotransferase (ALT) and aspartate aminotransferase (AST) as the two most predictive clinical factors in this dataset, too. These results confirmed the relevance of alanine aminotransferase (ALT) and aspartate aminotransferase (AST) to detect cirrhosis among patients with hepatitis C, and the fact that these two enzyme levels are used in the AST/ALT ratio to this end [43], [45]. In this context, it is also worth mentioning the study of

TABLE 10. Quantitative characteristics of the binary features of the Wu validation cohort dataset [12]. serogroup NAs: 25. PathDiagNum, sex, cirrhosis NAs: 0.

binary feature	value	#	%
PathDiagNum	1	104	84.553
serogroup	0	83	67.480
serogroup	1	15	12.195
sex	0	75	60.976
sex	1	48	39.024
[target] cirrhosis	0	40	32.520
[target] cirrhosis	1	83	67.480

Reedy and colleagues [57] which states that the AST/ALT ratio can be useful for cirrhosis suggestion but it should not be employed for diagnosis.

D. RANDOM FORESTS APPLIED TO AST AND ALT ALONE OUTPERFORM THE AST/ALT RATIO

To further verify the predictive power of machine learning applied AST and ALT, we performed a feature selection phase we identified the top two factors, trained a model containing only these two variables, and performed a binary classification with the trained model on a held-out subset. The results showed that Random Forests applied to AST and ALT are sufficient both to predict hepatitis C in the discovery

TABLE 11. Quantitative characteristics of the real-valued features of the Wu validation cohort dataset [12].

real-valued feature	median	mean	range	σ	NAs
age	70	69.577	[50, 86]	7.825	0
ALT	45	57.935	[9, 225]	37.275	0
AST	55	59.683	[15, 179]	30.065	0
BMI	23	23.122	[16.3, 34]	2.984	0
cholesterol	146	158.984	[86, 1151]	95.857	0
glucose	113	124.894	[72, 321]	44.256	0
HDL	43	44.358	[13, 85]	14.224	3
LDL	84.3	96.117	[14.6, 1106.6]	96.708	3
triglyceride	86	94.65	[31, 367]	44.237	0

TABLE 12. Results of the Random Forests feature ranking applied to the Lichtinghagen discovery cohort dataset [10], [11].

rank	feature	MeanDecreaseAccuracy
1	AST	0.0651332441
2	ALT	0.0228256906
3	GGT	0.0177230035
4	ALP	0.0174811164
5	BIL	0.0125196977
6	CHE	0.0072855325
7	PROT	0.0046532069
8	age	0.0036871968
9	CHOL	0.0036433966
10	ALB	0.0027952799
11	CREA	0.0025919881
12	sex	0.0003661006

cohort dataset and to predict cirrhosis in the validation cohort dataset. Additionally, we compared the results obtained this way with the results achieved by the application to the ALT/AST ratio, and we noticed that our Random Forests model outperformed this clinical criterion in both the cohorts.

This discovery reinforces the role of machine learning applied to medical data, in the direction of *minimal health records* [93], [94]. Health record data, in fact, derive from blood tests and other laboratory exams that need time and resources, which sometimes are unavailable in a hospital. The usage of a minimal clinical record, with the application of computational intelligence methods to a dataset having only 2 clinical variables, can provide results and insights with big impact, at a very low cost, in a short time.

E. LIMITATIONS AND FUTURE DEVELOPMENTS

A limitation of this study is that, unfortunately, some features of the discovery cohort dataset were absent from the validation cohort, and vice versa. If both the datasets had the same clinical variables, we would have been able to discuss additional similarities or differences among the obtained feature rankings.

In the future, we plan to employ the same approach to bioinformatics datasets of hepatitis C and cirrhosis, containing data of microarray gene expression [95]–[97]. We also plan to consider alternative data imputation methods [98] and to apply techniques to handle data imbalance [99], [100] .

SOFTWARE AND DATA AVAILABILITY

The R software code we developed and used for this study is available under the GPL-3.0 License at the following URL: https://github.com/davidechicco/hepatitis_C_virus

The discovery cohort Lichtinghagen dataset [10], [11] is publically available on the University of California Irvine Machine Learning Repository at the following URL: http://archive.ics.uci.edu/ml/datasets/HCV+data

The validation cohort Wu dataset [12] is publically available on FigShare at the following URL: https://doi.org/10.1371/journal.pone.0118297.s001

SUPPLEMENTARY INFORMATION

BINARY STATISTICAL RATES

List of statistical rates and their formulas:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \quad (1)$$

(worst value $= -1$; best value $= +1$)

$$\text{F}_1 \text{ score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (2)$$

(worst value $= 0$; best value $= 1$)

$$\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (3)$$

(worst value $= 0$; best value $= 1$)

$$\text{true positive rate = recall = sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

(worst value $= 0$; best value $= 1$)

$$\text{true negative rate = specificity} = \frac{TN}{TN + FP} \quad (5)$$

(worst value $= 0$; best value $= 1$)

$$\text{positivie predictive value = precision} = \frac{TP}{TP + FP} \quad (6)$$

(worst value $= 0$; best value $= 1$)

$$\text{negative predictive value} = \frac{TN}{TN + FN} \quad (7)$$

(worst value $= 0$; best value $= 1$)

Precision-Recall (PR) curve

$$= \begin{cases} \textit{true positive rate} & \textit{on the x axis} \\ \textit{precision} & \textit{on the y axis} \end{cases} \quad (8)$$

(worst value = 0; best value = 1)

ROC curve

$$= \begin{cases} \textit{false positive rate} & \textit{on the x axis} \\ \textit{true positive rate} & \textit{on the y axis} \end{cases} \quad (9)$$

(worst value = 0; best value = 1)

DATASETS ADDITIONAL INFORMATION

See Tables 10 and 11.

FEATURE RANKING

TABLE 13. Results of the Random Forests feature ranking applied to the Wu validation cohort dataset [12].

rank	feature	MeanDecreaseAccuracy
1	AST	0.0416608641
2	ALT	0.0128127065
3	cholesterol	0.0119807309
4	LDL	0.0081469742
5	HDL	0.0043238541
6	glucose	0.0018436839
7	sex	0.0007936335
8	PathDiagNum	0.0004960581
9	serogroup	–0.0005033249
10	age	–0.0006645135
11	BMI	–0.0023063749
12	triglyceride	–0.0041381478

See Tables 12 and 13.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and the organizers of the 1st International Workshop on Conceptual Modeling for Life Sciences (ER CMLS 2020) for their helpful feedback.

REFERENCES

[1] World Health Organization (WHO). (2020). *Hepatitis C Key Facts*. Accessed: Aug. 4, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c
[2] M. Khatun and R. B. Ray, "Mechanisms underlying hepatitis C virus-associated hepatic fibrosis," *Cells*, vol. 8, no. 10, p. 1249, Oct. 2019.
[3] K. T. Suk, "Staging of liver fibrosis or cirrhosis: The role of hepatic venous pressure gradient measurement," *World J. Hepatol.*, vol. 7, no. 3, p. 607, 2015.
[4] D. R. Dufour, J. A. Lott, F. S. Nolte, D. R. Gretch, R. S. Koff, and L. B. Seeff, "Diagnosis and monitoring of hepatic Injury. I. Performance characteristics of laboratory tests," *Clin. Chem.*, vol. 46, no. 12, pp. 2027–2049, Dec. 2000.
[5] European Association for Study of Liver, "EASL-ALEH clinical practice guidelines: Non-invasive tests for evaluation of liver disease severity and prognosis," *J. Hepatol.*, vol. 63, no. 1, pp. 237–264, 2015.
[6] P. Hall and J. Cash, "What is the real function of the liver 'function' tests?" *Ulster Med. J.*, vol. 81, no. 1, pp. 30–36, 2012.
[7] F. De Ritis, M. Coltorti, and G. Giusti, "An enzymic test for the diagnosis of viral hepatitis: The transaminase serum activities," *Clinica Chim. Acta*, vol. 2, no. 1, pp. 70–74, Feb. 1957.
[8] J. Guéchot, R. C. Boisson, J.-P. Zarski, N. Sturm, P. Calès, and E. Lasnier, "AST/ALT ratio is not an index of liver fibrosis in chronic hepatitis C when aminotransferase activities are determinate according to the international recommendations," *Clinics Res. Hepatol. Gastroenterol.*, vol. 37, no. 5, pp. 467–472, Nov. 2013.
[9] S. C. Nandipati, C. X. Ying, and K. K. Wah, "Hepatitis C virus (HCV) prediction by machine learning techniques," *Appl. Model. Simul.*, vol. 4, pp. 89–100, Mar. 2020.
[10] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—A case study," *J. Lab. Precis. Med.*, vol. 3, no. 6, p. 58, 2018.
[11] R. Lichtinghagen, D. Pietsch, H. Bantel, M. P. Manns, K. Brand, and M. J. Bahr, "The enhanced liver fibrosis (ELF) score: Normal values, influence factors and proposed cut-off values," *J. Hepatol.*, vol. 59, no. 2, pp. 236–242, Aug. 2013.
[12] Z. Wu, O. Matsui, A. Kitao, K. Kozaka, W. Koda, S. Kobayashi, Y. Ryu, T. Minami, J. Sanada, and T. Gabata, "Hepatitis c related chronic liver cirrhosis: Feasibility of texture analysis of MR images for classification of fibrosis stage and necroinflammatory activity grade," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0118297.
[13] The MAQC Consortium, "The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnol.*, vol. 28, no. 8, pp. 827–838, Aug. 2010.
[14] The SEQC/MAQC-III Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium," *Nature Biotechnol.*, vol. 32, no. 9, pp. 903–914, Sep. 2014.
[15] L. Shi, R. Kusko, R. D. Wolfinger, B. Haibe-Kains, M. Fischer, S.-A. Sansone, C. E. Mason, C. Furlanello, W. D. Jones, B. Ning, and W. Tong, "The international MAQC society launches to enhance reproducibility of high-throughput technologies," *Nature Biotechnol.*, vol. 35, no. 12, pp. 1127–1128, Dec. 2017.
[16] G. Kasarala and H. L. Tillmann, "Standard liver tests," *Clin. Liver Disease*, vol. 8, no. 1, pp. 13–18, Jul. 2016.
[17] O. Akkaya, "Clinical significance of activity of ALT enzyme in patients with hepatitis C virus," *World J. Gastroenterol.*, vol. 13, no. 41, p. 5481, 2007.
[18] P. Pradat, A. Alberti, T. Poynard, J.-I. Esteban, O. Weiland, P. Marcellin, S. Badalamenti, and C. Trépo, "Predictive value of ALT levels for histologic findings in chronic hepatitis C: A European collaborative study," *Hepatology*, vol. 36, no. 4, pp. 973–977, Oct. 2002.
[19] E. G. Giannini, "Liver enzyme alteration: A guide for clinicians," *Can. Med. Assoc. J.*, vol. 172, no. 3, pp. 367–379, Feb. 2005.
[20] F. J. C. Cuperus, J. P. H. Drenth, and E. T. Tjwa, "Mistakes in liver function test abnormalities and how to avoid them," *UEG Edu.*, vol. 17, pp. 1–5, Oct. 2017.
[21] Q. Xu, T. Higgins, and G. S. Cembrowski, "Limiting the testing of AST," *Amer. J. Clin. Pathol.*, vol. 144, no. 3, pp. 423–426, Sep. 2015.
[22] P. T. Giboney, "Mildly elevated liver transaminase levels in the asymptomatic patient," *Amer. Family Physician*, vol. 71, no. 6, pp. 1105–1110, 2005.
[23] R. C. O. Oh, T. R. Hustead, S. M. Ali, and M. W. Pantsari, "Mildly elevated liver transaminase levels: Causes and evaluation," *Amer. Family Physician*, vol. 96, no. 11, pp. 709–715, 2017.
[24] F. Anderson, "An assessment of the clinical utility of serum ALT and AST in chronic hepatitis C," *Hepatol. Res.*, vol. 18, no. 1, pp. 63–71, Jul. 2000.
[25] M. R. McGill, "The past and present of serum aminotransferases and the future of liver injury biomarkers," *EXCLI J.*, vol. 15, pp. 817–828, Dec. 2016.
[26] J. C. Connolly and J. K. Lim, "Non-invasive fibrosis assessment of patients with hepatitis C: Application of society guidelines to clinical practice," *Current Hepatol. Rep.*, vol. 18, no. 2, pp. 249–258, Jun. 2019.
[27] K. Nho, A. Kueider-Paisley, S. Ahmad, S. MahmoudianDehkordi, M. Arnold, S. L. Risacher, G. Louie, C. Blach, R. Baillie, X. Han, G. Kastenmüller, J. Q. Trojanowski, L. M. Shaw, M. W. Weiner, P. M. Doraiswamy, C. van Duijn, A. J. Saykin, and R. Kaddurah-Daouk, "Association of altered liver enzymes with alzheimer disease diagnosis, cognition, neuroimaging measures, and cerebrospinal fluid biomarkers," *JAMA Netw. Open*, vol. 2, no. 7, Jul. 2019, Art. no. e197978.
[28] S. Sareen, S. Kosey, and A. Goyal, "Various scores used for progression of liver disease," *Int. J. Pharmaceutical Sci. Res.*, vol. 8, no. 11, pp. 4852–4857, 2017.

[29] J.-S. Kang and M.-H. Lee, "Noninvasive diagnostic and prognostic assessment tools for liver fibrosis and cirrhosis in patients with chronic liver disease," in *Liver Cirrhosis-Update Current Challenges*. London, U.K.: IntechOpen, 2017.

[30] P. S. Kamath and W. R. Kim, "The model for end-stage liver disease (MELD)," *Hepatology*, vol. 45, no. 3, pp. 797–805, 2007.

[31] R. N. H. Pugh, I. M. Murray-Lyon, J. L. Dawson, M. C. Pietroni, and R. Williams, "Transection of the oesophagus for bleeding oesophageal varices," *Brit. J. Surgery*, vol. 60, no. 8, pp. 646–649, Aug. 1973.

[32] S.-Z. Wan, Y. Nie, Y. Zhang, C. Liu, and X. Zhu, "Assessing the prognostic performance of the Child-Pugh, model for end-stage liver disease, and albumin-bilirubin scores in patients with decompensated cirrhosis: A large asian cohort from gastroenterology department," *Disease Markers*, vol. 2020, pp. 1–9, Feb. 2020.

[33] G. Acharya, R. M. Kaushik, R. Gupta, and R. Kaushik, "Child-Turcotte-Pugh score, MELD score and MELD-na score as predictors of short-term mortality among patients with end-stage liver disease in northern India," *Inflammatory Intestinal Diseases*, vol. 5, no. 1, pp. 1–10, 2020.

[34] F. Durand and D. Valla, "Assessment of the prognosis of cirrhosis: Child–Pugh versus MELD," *J. Hepatol.*, vol. 42, no. 1, pp. S100–S107, Apr. 2005.

[35] Y. Peng, X. Qi, and X. Guo, "Child–Pugh versus MELD score for the assessment of prognosis in liver cirrhosis," *Medicine*, vol. 95, no. 8, p. e2877, Feb. 2016.

[36] J. Krige, R. T. Spence, E. Jonas, M. Hoogerboord, and J. Ellsmere, "A new recalibrated four-category Child–Pugh score performs better than the original Child–Pugh and MELD scores in predicting in-hospital mortality in decompensated alcoholic cirrhotic patients with acute variceal bleeding: A real-world cohort analysis," *World J. Surg.*, vol. 44, no. 1, pp. 241–246, Jan. 2020.

[37] Y. Chen, Y. Liu, W. Seto, M. Wu, Y. Yu, Y. Lam, W. Au, D. Chan, K. Sit, L. Ho, H. Tse, and K. Yiu, "Prognostic value of hepatorenal function by modified model for end-stage liver disease (MELD) score in patients undergoing tricuspid annuloplasty," *J. Amer. Heart Assoc.*, vol. 7, no. 14, Jul. 2018.

[38] U. Kartoun, K. E. Corey, T. G. Simon, H. Zheng, R. Aggarwal, K. Ng, and S. Y. Shaw, "The MELD-plus: A generalizable prediction risk score in cirrhosis," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0186301.

[39] I. Macpherson, J. H. Nobes, E. Dow, E. Furrie, M. H. Miller, E. M. Robinson, and J. F. Dillon, "Intelligent liver function testing: Working smarter to improve patient outcomes in liver disease," *J. Appl. Lab. Med.*, vol. 5, no. 5, pp. 1090–1100, Sep. 2020.

[40] P. N. Newsome, R. Cramb, S. M. Davison, J. F. Dillon, M. Foulerton, E. M. Godfrey, R. Hall, U. Harrower, M. Hudson, A. Langford, A. Mackie, R. Mitchell-Thain, K. Sennett, N. C. Sheron, J. Verne, M. Walmsley, and A. Yeoman, "Guidelines on the management of abnormal liver blood tests," *Gut*, vol. 67, no. 1, pp. 6–19, Jan. 2018.

[41] P. Y. Kwo, S. M. Cohen, and J. K. Lim, "ACG clinical guideline: Evaluation of abnormal liver chemistries," *Amer. J. Gastroenterol.*, vol. 112, no. 1, pp. 18–35, Jan. 2017.

[42] F. Wrobleski, "The clinical significance of alterations in transaminase activities of serum and other body fluids," *Adv. Clin. Chem.*, vol. 1, no. 2, pp. 313–351, 1958.

[43] S. G. Sheth, S. L. Flamm, F. D. Gordon, and S. Chopra, "AST/ALT ratio predicts cirrhosis in patients with chronic hepatitis C virus infection," *Amer. J. Gastroenterol.*, vol. 93, no. 1, pp. 44–48, Jan. 1998.

[44] T. F. Imperiale, A. T. Said, O. W. Cummings, and L. J. Born, "Need for validation of clinical decision aids: Use of the AST/ALT ratio in predicting cirrhosis in chronic hepatitis C," *Amer. J. Gastroenterol.*, vol. 95, no. 9, pp. 2328–2332, Sep. 2000.

[45] G. J. Park, B. P. Lin, M. C. Ngu, D. B. Jones, and P. H. Katelaris, "Aspartate aminotransferase : Alanine aminotransferase ratio in chronic hepatitis C infection: Is it a useful predictor of cirrhosis?" *J. Gastroenterol. Hepatol.*, vol. 15, no. 4, pp. 386–390, Apr. 2000.

[46] E. Giannini, D. Risso, and R. Testa, "Transportability and reproducibility of the AST/ALT ratio in chronic hepatitis C patients," *Amer. J. Gastroenterol.*, vol. 96, no. 3, p. 918, 2001.

[47] E. Giannini, D. Risso, F. Botta, B. Chiarbonello, A. Fasoli, F. Malfatti, P. Romagnoli, E. Testa, P. Ceppa, and R. Testa, "Validity and clinical utility of the aspartate aminotransferase–alanine aminotransferase ratio in assessing disease severity and prognosis in patients with hepatitis C virus–related chronic liver disease," *Arch. Internal Med.*, vol. 163, no. 2, p. 218, Jan. 2003.

[48] H. Nyblom, "High AST/ALT ratio may indicate advanced alcoholic liver disease rather than heavy drinking," *Alcohol Alcoholism*, vol. 39, no. 4, pp. 336–339, Jul. 2004.

[49] A. Nadeem, M. M. Hussain, and M. Aslam, "Correlation of serum alanine aminotransferase and aspartate aminotransferase levels to liver histology in chronic hepatitis C," *J. College Physicians Surgeons Pakistan*, vol. 20, no. 10, pp. 61–657, 2010.

[50] K. Parmar, G. Singh, G. Gupta, T. Pathak, and S. Nayak, "Evaluation of De Ritis ratio in liver-associated diseases," *Int. J. Med. Sci. Public Health*, vol. 5, no. 9, p. 1783, 2016.

[51] S. Rawal, A. Shahi, N. Gautam, A. Jayan, and U. Sharma, "Enzyme activity and De Ritis ratio in alcoholic and non alcoholic fatty liver patients based on ultrasonography," *J. Universal College Med. Sci.*, vol. 8, no. 1, pp. 9–13, Jul. 2020.

[52] O. Knittelfelder, D. Delago, G. Jakse, S. Reinisch, R. Partl, H. Stranzl-Lawatsch, W. Renner, and T. Langsenlehner, "The AST/ALT (De Ritis) ratio predicts survival in patients with oral and oropharyngeal cancer," *Diagnostics*, vol. 10, no. 11, p. 973, Nov. 2020.

[53] Z. Lu, G. Ma, and L. Chen, "De-ritis ratio is associated with mortality after cardiac arrest," *Disease Markers*, vol. 2020, pp. 1–13, Nov. 2020.

[54] H. Yazar, Y. Kayacan, and M. Ozdin, "De Ritis ratio and biochemical parameters in COVID-19 patients," *Arch. Physiol. Biochem.*, vol. 51, no. 1, pp. 1–5, Jul. 2020.

[55] A. Zinellu, F. Arru, A. De Vito, A. Sassu, G. Valdes, V. Scano, E. Zinellu, R. Perra, G. Madeddu, C. Carru, P. Pirina, A. A. Mangoni, S. Babudieri, and A. G. Fois, "The De Ritis ratio as prognostic biomarker of in-hospital mortality in COVID-19 patients," *Eur. J. Clin. Invest.*, vol. 51, no. 1, Jan. 2021, Art. no. e13427.

[56] M. Botros and K. A. Sikaris, "The De Ritis ratio: The test of time," *Clin. Biochemist Rev.*, vol. 34, no. 3, pp. 117–130, 2013.

[57] D. W. Reedy, A. T. Loo, and R. A. Levine, "AST/ALT ratio $\geq$ 1 is not diagnostic of cirrhosis in patients with chronic hepatitis C," *Digestive Diseases Sci.*, vol. 43, no. 9, pp. 2156–2159, 1998.

[58] I. Landi, B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, and R. Miotto, "Deep representation learning of electronic health records to unlock patient stratification at scale," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–11, Jul. 2020.

[59] Y. Yu, J. Wang, C. W. Ng, Y. Ma, S. Mo, E. L. S. Fong, J. Xing, Z. Song, Y. Xie, K. Si, A. Wee, R. E. Welsch, P. T. C. So, and H. Yu, "Deep learning enables automated scoring of liver fibrosis stages," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Oct. 2018.

[60] K. Yasaka, H. Akai, A. Kunimatsu, O. Abe, and S. Kiryu, "Deep learning for staging liver fibrosis on CT: A pilot study," *Eur. Radiol.*, vol. 28, no. 11, pp. 4578–4585, May 2018.

[61] M. A. Khan, J. E. Soh, M. Maenner, W. W. Thompson, and N. P. Nelson, "A machine-learning algorithm to identify hepatitis c in health insurance claims data," *Online J. Public Health Informat.*, vol. 11, no. 1, pp. 1–2, May 2019.

[62] G. Kurniawan and Z. Rustam, "Enhancement of hepatitis virus outcome predictions with application of K-means clustering," in *Proc. 4TH Int. Symp. CURRENT Prog. Math. Sci. (ISCPMS)*, 2019, vol. 2168, no. 1, Art. no. 020044.

[63] J. Zucker, J. G. Aaron, D. J. Feller, J. Slowikowski, H. Evans, M. L. Scherer, M. T. Yin, and P. Gordon, "Development and validation of an electronic medical record–based algorithm to identify patient milestones in the hepatitis c virus care cascade," *Open Forum Infectious Diseases*, vol. 5, no. 7, Jul. 2018, Art. no. ofy153.

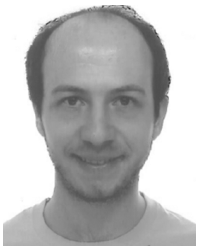
[64] A. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, and M. Bhat, "Applying machine learning in liver disease and transplantation: A comprehensive review," *Hepatology*, vol. 71, no. 3, pp. 1093–1105, Mar. 2020.

[65] S. Hashem, M. ElHefnawi, S. Habashy, M. El-Adawy, G. Esmat, W. Elakel, A. O. Abdelazziz, M. M. Nabeel, A. H. Abdelmaksoud, T. M. Elbaz, and H. I. Shousha, "Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105551.

[66] A. Arjmand, M. G. Tsipouras, A. T. Tzallas, R. Forlano, P. Manousou, and N. Giannakeas, "Quantification of liver fibrosis—A comparative study," *Appl. Sci.*, vol. 10, no. 2, p. 447, Jan. 2020.

[67] N. H. Barakat, S. H. Barakat, and N. Ahmed, "Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach," *Healthcare Informat. Res.*, vol. 25, no. 3, p. 173, 2019.

[68] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. I. Eladawy, and M. ElHefnawi, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 3, pp. 861–868, May 2018.
[69] L. Chengxi, L. Rentao, and Z. Wei, "Progress in non-invasive detection of liver fibrosis," *Cancer Biol. Med.*, vol. 15, no. 2, p. 124, 2018.
[70] G. N. Ioannou, W. Tang, L. A. Beste, M. A. Tincopa, G. L. Su, T. Van, E. B. Tapper, A. G. Singal, J. Zhu, and A. K. Waljee, "Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis," *JAMA Netw. Open*, vol. 3, no. 9, Sep. 2020, Art. no. e2015626.
[71] S. V. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations inR," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–68, 2011.
[72] L. R. Landerman, K. C. Land, and C. F. Pieper, "An empirical evaluation of the predictive mean matching method for imputing missing values," *Sociol. Methods Res.*, vol. 26, no. 1, pp. 3–33, Aug. 1997.
[73] *University of California Los Angeles (UCLA) Statistical Consulting Group. How Do I Perform Multiple Imputation Using Predictive Mean Matching in R?* Accessed: Aug. 4, 2020. [Online]. Available: https://stats.idre.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/
[74] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *Amer. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, Mar. 2014.
[75] K. Jung and N. H. Shah, "Implications of non-stationarity on predictive modeling using EHRs," *J. Biomed. Informat.*, vol. 58, pp. 168–174, Dec. 2015.
[76] B. Westra, S. Dey, G. Fang, M. Steinbach, V. Kumar, C. Oancea, K. Savik, and M. Dierich, "Interpretable predictive models for knowledge discovery from home-care electronic health records," *J. Healthcare Eng.*, vol. 2, no. 1, pp. 55–74, Mar. 2011.
[77] D. Chicco and C. Rovelli, "Computational prediction of diagnosis and feature selection on mesothelioma patient health records," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0208737.
[78] G. A. Seber and A. J. Lee, *Linear Regression Analysis*, vol. 329. Hoboken, NJ, USA: Wiley, 2012.
[79] R. Potharst and J. C. Bioch, "Decision trees for ordinal classification," *Intell. Data Anal.*, vol. 4, no. 2, pp. 97–111, Mar. 2000.
[80] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[81] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, Nov. 2003.
[82] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.
[83] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.
[84] T. W MacFarland and J. M. Yates, "Mann–whitney U test," in *Introduction to Nonparametric Statistics for the Biological Sciences using R*. Berlin, Germany: Springer, 2016, pp. 103–132.
[85] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi-Squared Testing*, vol. 280. Hoboken, NJ, USA: Wiley, 1996.
[86] D. J. Benjamin *et al.*, "Redefine statistical significance," *Nature Hum. Behav.*, vol. 2, no. 1, pp. 6–10, 2018.
[87] Y. Qi, "Random forest for bioinformatics," in *Ensemble Machine Learning*. Berlin, Germany: Springer, 2012, pp. 307–323.
[88] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.*, vol. 8, no. 1, p. 25, Dec. 2007.
[89] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2017.
[90] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020.
[91] D. Chicco and M. Masseroli, "A discrete optimization approach for SVD best truncation choice based on ROC curves," in *Proc. 13th IEEE Int. Conf. Bioinf. BioEng.*, Chania, Greece, Nov. 2013, pp. 1–4.
[92] O. Renaud and M.-P. Victoria-Feser, "A robust coefficient of determination for regression," *J. Stat. Planning Inference*, vol. 140, no. 7, pp. 1852–1862, Jul. 2010.
[93] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.
[94] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das, "Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach," *JMIR Med. Informat.*, vol. 4, no. 3, p. e28, Sep. 2016.
[95] S. Blackham, A. Baillie, F. Al-Hababi, K. Remlinger, S. You, R. Hamatake, and M. J. McGarvey, "Gene expression profiling indicates the roles of host oxidative stress, apoptosis, lipid metabolism, and intracellular transport genes in the replication of hepatitis C virus," *J. Virol.*, vol. 84, no. 10, pp. 5404–5414, May 2010.
[96] N. A. Wijetunga, M. Pascual, J. Tozour, F. Delahaye, M. Alani, M. Adeyeye, A. W. Wolkoff, A. Verma, and J. M. Greally, "A preneoplastic epigenetic field defect in HCV-infected liver at transcription factor binding sites and polycomb targets," *Oncogene*, vol. 36, no. 14, pp. 2030–2044, Apr. 2017.
[97] A. Venkatesh, X. Sun, and Y. Hoshida, "Prognostic gene signature profiles of hepatitis C-related early-stage liver cirrhosis," *Genomics Data*, vol. 2, p. 361, Dec. 2014.
[98] E. Tavazzi, S. Daberdaku, R. Vasta, A. Calvo, A. Chiò, and B. Di Camillo, "Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data with an adaptive k-nearest neighbours approach," *BMC Med. Informat. Decis. Making*, vol. 20, no. S5, p. 174, Aug. 2020.
[99] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: A package for binary imbalanced learning," *R J.*, vol. 6, no. 1, p. 79, 2014.
[100] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.



DAVIDE CHICCO received the B.Sc. and M.Sc. degrees in computer science from the Università di Genova, Genoa, Italy, in 2007 and 2010, respectively. He then started the Ph.D. program in computer engineering at the Politecnico di Milano University, Milan, Italy, where he graduated in spring 2014. He also spent a Semester as a Visiting Doctoral Scholar at the University of California Irvine, USA. From September 2014 to September 2018, he had been a Postdoctoral Researcher with the Princess Margaret Cancer Centre and a Guest with the University of Toronto. From September 2018 to January 2021, he was a Scientific Researcher first at the Peter Munk Cardiac Centre and then at the Krembil Research Institute, Toronto, ON, Canada. Since January 2021, he has been a Scientific Researcher with the University of Toronto.

GIUSEPPE JURMAN received the Ph.D. degree in algebra from the Università di Trento, Italy, in 1998. After two years as a Postdoctoral Fellow at the Australian National University (ANU) Canberra, in 2002, he moved to the Fondazione Bruno Kessler (FBK) in Trento, where he is currently a Senior Researcher of data science, working mainly on computational biology. His research interests include machine learning, mathematical modeling, and network analysis. He is also an Expert in scientific programming with R/Python and other computing languages. He teaches Data Visualization in the Master of Science course in Data Science at the Università di Trento, and since 2008, he has been the Co-Director of WebValley, the FBK summer school for dissemination of interdisciplinary research for high school students.