scientific reports

Check for updates

OPEN

Machine learning for predicting hepatitis B or C virus infection in diabetic patients

Sun–Hwa Kim[1,3], So–Hyeon Park[1,3] & Heeyoung Lee[2]✉

Highly prevalent hepatitis B and hepatitis C virus (HBV and HCV) infections have been reported among individuals with diabetes. Given the frequently asymptomatic nature of hepatitis and the challenges associated with screening in some vulnerable populations such as diabetes patients, we conducted an investigation into the performance of various machine learning models for the identification of hepatitis in diabetic patients while also evaluating the significance of features. Analyzing NHANES data from 2013 to 2018, machine learning models were evaluated; random forest (RF), support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), and least absolute shrinkage and selection operator (LASSO) along with stacked ensemble model. We performed hyperparameter tuning to improve the performance of the model, and selected important predictors using the best performance model. LASSO showed the highest predictive performance (AUC-ROC = 0.810) rather than other models. Illicit drug use, poverty, and race were highly ranked as predictive factors for developing hepatitis in diabetes patients. Our study demonstrated that a machine-learning-based model performed optimally in the detection of hepatitis among diabetes patients, achieving high performance. Furthermore, models and predictors evaluated from the current study, we expect, could be supportive information for developing screening or treatment methods for hepatitis care in diabetes patients.

Diabetes mellitus (DM) has remained one of the most problematic chronic metabolic disorders in humans over the past decades[1]. Based on the recent statistics released by the centers for disease control and prevention (CDC), DM affects approximately 34.2 million individuals in the US[2]. DM can be primarily categorized into two types such as type 1 (T1DM) and type 2 diabetes mellitus (T2DM)[3]. T1DM is an immune-mediated disease leading to absolute deficiency of endogenous insulin caused by β-cell loss in the pancreas[3]. On the other hand, T2DM is a prevalent endocrine disorder characterized by multifactorial mechanisms[3]. These mechanisms encompass insulin resistance, increased glucose production by the liver, and impaired insulin secretion[3]. In both T1DM and T2DM, a combination of genetic and environmental factors can lead to the gradual decline of β-cell mass and/or function, which is clinically manifested as hyperglycemia in T1DM and T2DM[4]. Once hyperglycemia occurs, patients with any forms of diabetes are susceptible to developing complications caused by impact on several organ systems over time[4]. Relevant to disturbed glucose homeostasis, hepatitis B virus (HBV) and hepatitis C virus (HCV) infections in populations with DM have recently been reported as an emergent comorbidity[5,6]. With a considerably high prevalence, 865 outbreaks of HBV infection were previously reported among adults who were diagnosed with diabetes[5]. Adults with DM have a 60% higher prevalence of HBV infection and are twice as likely to experience acute HBV infection compared to adults without DM[7]. According to the study by Gisi et al.[8], along with HBV, the prevalence of HCV in the diabetic group was significantly higher than that in the non-diabetic group.

While there is a high prevalence of HBV or HCV infections in diabetes, regarding asymptomatic nature in hepatitis, the challenges associated with screening in some vulnerable populations such as diabetes patients are still remained. Some patients have reported never developing histologic evidence of liver disease even after decades of infection reflecting obstacles in identifying hepatitis[9,10]. Furthermore, in those with comorbid conditions such as DM, identifying or predicting HBV or HCV accurately has proven to be also challenging, emphasizing the need for more selective screening methods[11–14]. Moreover, previous studies have yielded conflicting results regarding the important risk factors for hepatitis development in individuals with diabetes[15,16]. Therefore, further

[1]Department of Clinical Medicinal Sciences, Konyang University, Nonsan, Republic of Korea. [2]College of Pharmacy, Inje University, Gimhae, Republic of Korea. [3]These authors contributed equally: Sun–Hwa Kim and So–Hyeon Park. ✉email: phylee1@inje.ac.kr

studies conducted with improved strategies such as using machine learning models are necessary to provide essential information about predictors of hepatitis development in diabetes, aiding clinical decision-making.

In response to these unmet needs, machine learning has emerged as a promising alternative to traditional hepatitis screening strategies in recent years, making significant breakthroughs in the realm of public screening even including finance or wireless sensor network (WSN)[17–21]. In finance and WSNs that without clinical data, after tuning hyperparameters, machine learning revealed the best results among other base learning models[18,21]. Besides, given the demonstrated effectiveness of machine learning models across various fields, especially in healthcare, machine learning has emerged as a powerful approach. It enables the extraction of valuable information from imbalanced clinical datasets and facilitates decision-making through accurate predictions[17]. Doğru et al.[19] using clinical data showed the best accuracy in predicting early diabetes risk through hybrid super ensemble learning model (99.6%). Furthermore, several studies[22,23] have been conducted on predictive machine-learning models for diabetes treatment using laboratory and clinical data. Ozyilmaz et al.[24] demonstrated the possibility of accurate hepatitis identification through machine learning, and numerous previous works[25–29] have consistently used machine learning to facilitate the early prediction of a significant number of people who are at high risk for hepatitis. Yağanoğlu et al.[26] even tried to develop a model with high performance in predicting HCV with HCV dataset by comparing eight machine learning models, such as random forest (RF) and K-nearest neighbor, which yielded an overall accuracy of 96.75%. Another previous study based on clinical records of 155 hepatitis B patients predicting HBV revealed adaptive boosting as an accurate model (92%) for diagnosis after comparing other machine learning models such as eXtreme Gradient Boosting (XGBoost) and RF[29]. Given the significant previous results achieved by individual machine learning models in predicting hepatitis, to enhance predictive performance, ensemble learning has been applied for various researches as well[18–20,30,31]. By integrating various machine learning algorithms, ensemble techniques aim to make more accurate predictions compared to a single classifier[30]. When the base models are diverse and independent resulted in lower accuracy of predictions, regarding as primary goal, using ensemble models is to reduce generalization errors[18]. Despites of generally improving performance of ensemble models, several researches[25,31] in different study settings still showed lower accuracy value of ensemble learning rather than a single base model. Edeh et al.[25] still showed that an artificial intelligence-based ensemble model could predict HCV with an accuracy of approximately 94% for early diagnosis and treatment for HCV infection. Based on these discrepancies, we need to explore optimal performance metrics in single and ensemble models for predicting hepatitis development in DM patients.

To the best of our knowledge, there has not been a study evaluating machine-learning models, including ensemble models, for predicting hepatitis development exclusively among patients with diabetes. In this study, we have conducted an investigation to determine the most suitable machine-learning models and identify relevant risk factors.

Results

Characteristic analysis

We evaluated the demographics, body measurements, lipids, and questionnaire data to analyze the association between diabetes and the 12 risk factors for hepatitis. The dataset that was preprocessed from the National Health and Nutrition Examination Survey (NHANES) 2013–2018 included 29,400 participants. A total of 26,190 participants without diabetes or missing data on diabetes were excluded. Consequently, a total of 3210 diabetes identification cases remained in the complete dataset. Thus, a total of 1396 diabetic patients were included in the study (Fig. 1). The synthetic minority oversampling technique (SMOTE) balancing technique was applied to the dataset prior to establishing the model owing to the imbalanced ratio of non-hepatitis to hepatitis patients. Following data normalization, the machine-learning methods were applied to train and test the models on the training and test datasets.

A total of 1396 people with a mean age of 54.66 participated in the study, including 64 with HBV or HCV and 1332 without HBV or HCV (Table 1). The hepatitis group had a higher percentage of non-Hispanic White and Asian individuals, whereas the non-hepatitis group had a higher percentage of Mexican and other Hispanic individuals. Males comprised 70.3% of the patients in the hepatitis group. More than half of the patients in the hepatitis group administered illegal drugs.

Model performance comparison

We used a randomized search with ten iterations threefold cross-validation for each of the four models to predict hepatitis. Single machine learning models including RF, XGBoost, support vector machine (SVM), and least absolute shrinkage and selection operator (LASSO) algorithm and stacked ensemble model were created. Sixteen statistically significant clinical parameters were included within these models. The performance comparison results of the four machine-learning methods, both before and after hyperparameter tuning, are presented in Table 2. The LASSO achieved the best accuracy value (0.978) after hyperparameter tuning. The sensitivity of all four models was generally low. Specificity of LASSO was also higher than other models (0.993). Additionally, the precision and F1 score of LASSO surpassed those of the other three models. Consequently, LASSO outperformed the other models across all evaluation metrics as presented in Table 2. To improve the results, we applied the stacking ensembles algorithm to combine the predictions of individual models. In the stack-based ensemble model, the accuracy is 0.945, without showing a significant improvement compared to individual models. The specificity was 0.958, and the sensitivity was 0.500, indicating that it did not outperform the best individual model, in terms of predictive performance.

The ROC curve for the classification performance of the four machine-learning models is depicted in Fig. 2. The LASSO model performed the best among all of the classifiers, with an area under the receiver operating

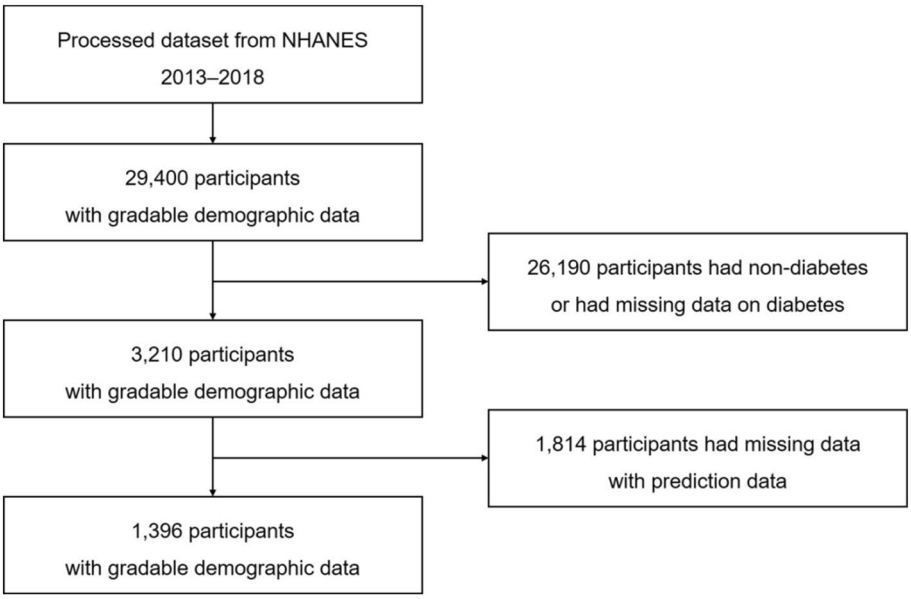

Figure 1. Flowchart of dataset creation.

Variables	Total (n = 1396)	Hepatitis B or C (n = 64)	No hepatitis B or C (n = 1332)
Demographics			
Age, mean ± SD	54.66 ± 10.85	57.69 ± 8.423	54.52 ± 10.93
Male, n (%)	717 (51.36)	45 (70.3)	672 (50.5)
Race, n (%)			
Mexican American	269 (19.27)	7 (10.94)	262 (19.67)
Other Hispanic	159 (11.39)	4 (6.25)	155 (11.64)
Non-Hispanic White	400 (28.65)	23 (35.94)	377 (28.30)
Non-Hispanic Black	375 (26.86)	17 (26.56)	358 (26.88)
Non-Hispanic Asian	135 (9.670)	10 (15.63)	125 (9.384)
Poverty, mean ± SD	2.393 ± 1.603	2.006 ± 1.651	2.411 ± 1.599
Body measure			
BMI, mean ± SD	111.2 ± 7.934	31.45 ± 7.128	33.55 ± 7.961
Waist circumference, mean ± SD	122.0 ± 17.32	107.8 ± 16.19	111.3 ± 17.36
Lipids (mg/dL)			
HDL cholesterol, mean ± SD	47.78 ± 14.93	46.27 ± 13.39	47.85 ± 15.00
Total cholesterol, mean ± SD	188.1 ± 49.21	168.4 ± 41.73	189 ± 49.36
Questionnaire data			
Receiving blood, n (%)	184 (13.18)	15 (23.44)	169 (12.69)
Hepatitis B vaccine, n (%)	430 (30.80)	14 (21.88)	416 (31.23)
General health condition, n (%)	781 (55.95)	32 (50)	749 (56.23)
Illegal drug injection, n (%)	36 (2.579)	23 (35.94)	13 (0.976)
Diagnosis			
Hepatitis B, n (%)	26 (1.862)	26 (40.63)	0
Hepatitis C, n (%)	40 (2.865)	40 (62.5)	0

Table 1. Characteristics of the study population. BMI, body measure index; HDL-cholesterol, high-density lipoprotein cholesterol.

characteristic curve (AUC-ROC) of 0.810. The AUC-ROC scores of the RF (0.794) and XGBoost (0.761) were lower than that of the LASSO.

Feature importance

The importance results for the 16 variables based on the highest-performing LASSO model are presented in Table 3 and Fig. 3. The most important predictor was illegal drug injection (71.036), which was the most reliable

Algorithm	Sensitivity	Specificity	Precision	F1 score	Accuracy
Without hyperparameter tuning					
RF	0.683	0.903	0.857	0.802	0.760
SVM	0.791	0.916	0.889	0.837	0.859
XGBoost	0.705	0.882	0.836	0.800	0.765
LASSO	0.591	0.898	0.831	0.756	0.691
With hyperparameter tuning					
RF	0.461	0.978	0.400	0.429	0.962
SVM	0.500	0.990	0.600	0.545	0.976
XGBoost	0.500	0.968	0.316	0.387	0.954
LASSO	0.500	0.993	0.667	0.571	0.978

Table 2. Performance of algorithms in diagnosis based on with and without hyperparameter tuning. RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting; LASSO, least absolute shrinkage and selection operator.

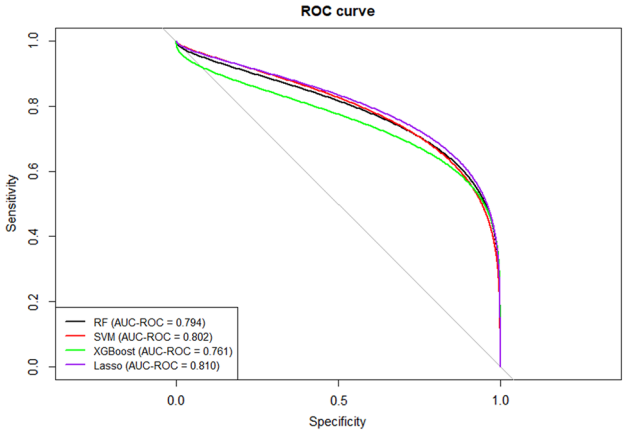

Figure 2. Performance of algorithms in diagnosis.

No	LASSO	
	Variable	Importance score
1	Illegal drug injection	71.036
2	Poverty	9.084
3	Mexican American	7.867
4	BMI	7.798
5	Age	6.930
6	Total cholesterol	6.538
7	Hepatitis B vaccine	4.565
8	General health condition	3.877
9	HDL cholesterol	1.235
10	Non-Hispanic Black	1.057
11	Other Hispanic	1.001
12	Waist circumference	0.731
13	Non-Hispanic Asian	0.681
14	Male	0.561
15	Receiving blood	0.401
16	Non-Hispanic White	0.218

Table 3. Ranked importance scores. BMI, body measure index; HDL-cholesterol, high-density lipoprotein cholesterol.

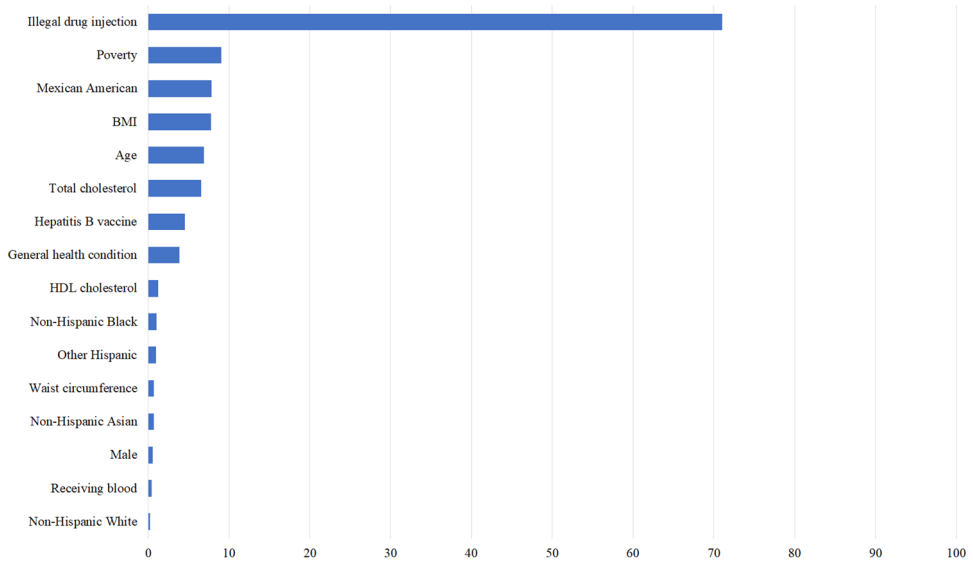

Figure 3. Ranked importance scores.

variable, and this was closely followed by the ratio of family income to poverty (9.084), Mexican Americans (7.867), the body mass index (BMI) (7.798), and age (6.930). Finally, non-Hispanic White (0.218) were placed in the lowest ranking.

Discussion

This study evaluated machine-learning models used in previous clinical studies to demonstrate the best predicting model and predictors for HBV or HCV infection in patients with DM. The results revealed that all models exhibited improved performance after the hyperparameter tuning process, with LASSO demonstrating the highest predictive ability for HBV or HCV infection development in diabetic patients. Hyperparameter optimization involves selecting the most suitable parameter values from a given parameter space, allowing for the optimization of model complexity and enhancing the performance of machine learning algorithms[32]. Hyperparameter optimization allows for the improvement of algorithm performance by identifying the correct parameter values, as they cannot be directly learned from the algorithms themselves[32]. Through hyperparameter optimization, LASSO emerged as the best-performing model in our study. Previously, LASSO has already demonstrated its clinical efficacy through several studies[21,33–36]. Ahn et al.[34] indicated that LASSO showed superior performance in predicting hepatocellular carcinoma in patients with chronic hepatitis B, particularly when utilizing twelve-month post-treatment parameters (AUC-ROC = 0.843). In another study focusing on the prediction of hepatocellular carcinoma, LASSO demonstrated excellent predictive performance when combined with first-year clinical parameters in chronic hepatitis B patients[35]. This highlights the accurate risk assessment and emphasizes the potential of LASSO for early detection in hepatocellular carcinoma prediction. Other previous studies[36,37] predicting diabetes also demonstrated that LASSO (0.84) has been shown to exhibit promising performance compared to other models such as RF, gradient boosted decision tree, deep neural network and the reference model using a logistic regression, providing evidence of its effectiveness in accurately predicting diabetes. Consequently, the current study demonstrates that the LASSO is even suitable for evaluating predictors of hepatitis in diabetic patients and highlights the importance of hyperparameter tuning in improving model accuracy. These findings provide valuable insights for developing more effective predictive models and aiding clinical decision-making.

In an attempt to demonstrate more improved performance, ensemble approach was tried in the current study, but did not achieve favorable results. One study demonstrated the efficacy of ensemble model by achieving the highest accuracy of 99.6% and superior ROC curve performances compared to other machine learning algorithms[19]. Buyrukoğlu et al.[20] showed that the individual models exhibited lower performance, with an accuracy of approximately 95% compared to the ensemble model. Particularly, among the ensemble models, the artificial neural network indicated the highest accuracy of 99.1%. While several studies[19,20] have shown better performance results using ensembles, other studies[31,32] have presented lower performance metrics than single base models, similar outcomes to our research. A prior research using similar data to our study, NHANES, has shown that there were variations of accuracy in the performing results of ensemble models depending on diseases such as CVD (0.839) and DM (0.737) and period collect data[31]. For predicting DM, the performance metrics were lower than a single model[31]. In another study[32], stacking ensembles combining models with individually excellent performance also showed lower performance results than base models such as XGboost, which is also consistent with our findings. Despites of combining high-performing models in our study, the ensemble results indicated lower performance compared to individual models. As consistently shown in previous studies, our study also demonstrated stacking ensemble always did not improve performance metrics for the predictions, suggesting the need for further future researches on ensemble performance in various settings or datasets.

The current study evaluated illegal drug injection, poverty, and race as key factors for predicting hepatitis. The results of major predictor variables related to hepatitis prevalence in diabetic patients, as observed in the study by Han et al.[38], were consistent with the findings of our study. Illegal drug use occurs globally, negatively affects the quality of life of individuals and communities, reduces productivity, and significantly increases the demands on healthcare systems[39]. Parenteral exposure with the sharing of unsterilized needles is a risk factor for the transmission of viruses in illegal drug users. Additionally, it should be noted that over 25% of individuals with diabetes administer insulin[40], which can potentially be transmitted if they share equipment, supplies, or insulin with others[41]. Furthermore, diabetes patients who are at a higher risk of experiencing elevated stress levels and immune dysfunction may also exhibit increased vulnerability to illicit drug use and transmission of HCV[38]. Moreover, multiple studies have reported a correlation between drug users and the development of HBV or HCV[2,42–47]. Eckhardt et al.[48] indicated that 343 out of 714 participants in their study on young Americans who were injected with drugs were infected with HCV. Furthermore, the hepatitis population-attributable fraction in 2013 for HBV (10% in North America and 1% in Latin America) and HCV (81% in North America and 31% in Latin America) was further increased from 1990 for HBV (6% in North America and 1% in Latin America) and HCV (60% in North America and 19% in Latin America)[47]. Although insufficient evidence is available and controversies surround the effects of diabetes on HBV or HCV infections, and further research is required, previous studies have consistently demonstrated the impact of diabetes on HBV or HCV infections[49,50]. According to Schillie et al.[49], a higher prevalence of HBV infection was observed among persons with diabetes compared to those without diagnosed diabetes (odds ratio (OR) = 1.60; 95% CI 1.30–1.90, $p < 0.05$). Furthermore, a meta-analysis study[50]. Revealed that patients with type 2 diabetes were at a higher risk of acquiring HCV infection than non-type 2 diabetic patients (OR = 3.50, 95% CI 2.54–4.82; I2 = 82.3%). Moreover, considering the problems or dysfunction of the immune system in diabetic patients, greater vulnerability to virus infection may be expected with the use of illegal drugs[51]. Therefore, this study can provide important insights in identifying a wide range of illegal drug users at high risk of HCV and HBV infection. By identifying and assessing predictive factors resulted from the current study, healthcare providers can intervene at an early stage to prevent or manage hepatitis infections in this specific population.

The next most relevant predictor in this study was poverty. People who live in poverty are generally vulnerable to infectious diseases owing to their poor living conditions and difficulties in accessing healthcare[52]. Although various factors contribute to poverty, such as age or education levels, and these are closely related to the spread of infectious diseases, poverty was only defined with the income ratio in this study[53]. In particular, based on the American Association for the Study of Liver Disease (AASLD), HBV and HCV are the leading infectious diseases that are closely related to poverty[54,55]. Greene et al.[56] analyzed surveillance data in New York City and reported that chronic hepatitis C was included in diseases that were related to severe poverty in people with a low income and hepatitis B susceptibility rates are approximately 32% among individuals with a low income in Brazil[57].

Moreover, previous studies have shown that diabetes is also associated with poverty[58]. The lack of primary healthcare owing to poverty makes it difficult for those living in poverty to access services that can reduce the incidence of hepatitis, especially in low-income countries[52]. Therefore, poverty in people with complications such as DM is a cause of increased HBV or HCV infections and is also likely to be an important factor in the incidence rate because of the cost burden[59].

Race/ethnicity has historically been an important factor in many diseases. Although the mechanism did not clearly explain the outcomes of the current study, race is an important factor in the development of hepatitis in patients with DM. We demonstrated that Mexican Americans are highly affected races. Mexican Americans were the most important variable among the races. In previous studies, the HBV infection rate was estimated to be 2.9% for Mexican Americans, which reflects almost tenfold higher prevalence rate than that in the general population[60,61]. An increased percentage of HCV infection was observed among Mexican Americans during the more recent period (2011–2016) than from 1999 to 2010 (5.6% vs. 10.6%)[62]. This analytical evidence is consistent with the finding that Mexican Americans are significantly associated with infection. Furthermore, although our study did not demonstrated as highly important factors, HBV prevalence according to the NHANES, 20.5% of Asian Americans had been infected with HBV between 2011 and 2012[61], which was also shown the association between the risk in Asian Americans and a high HBV or HCV prevalence rate in foreign-born Asian countries[63]. Regardless of the outcomes of the current study, more future studies with a larger scale data analyzing the prevalence of HCV and HBV in Asian Americans should be conducted based on this study.

Through identifying the contributing factors to disease manifestation, our findings indicate that machine learning models exhibit promising outcomes in the early detection of hepatitis among DM patients. Then, the current outcomes might contribute to model implementation in various tools such as web-based screening system, utilizing questionnaires to assess individuals' disease risk at developing hepatitis B or C. Thus, in clinical practice, with integrating survey and laboratory, models and predictors evaluated from the current study, we expect, could be supportive information for developing screening methods or treatment strategy for hepatitis care in DM patients.

Based on the encouraging findings of the current study, our research also exhibits several notable strengths. Firstly, to our knowledge, this study represents the first attempt to assess different machine learning models in predicting HBV and HCV infections in patients with DM. While a previous study examined the NHANES dataset and highlighted the susceptibility of DM patients to developing hepatitis[38], our study contributes important insights for developing screening strategies and improving treatment accessibility in identifying vulnerable DM patients at risk of various infections. Secondly, we utilized various statistical analysis methods such as SMOTE method or hyper-parameter tuning to address the imbalanced data and evaluate the optimal machine learning model. Lastly, our study presented valuable insights into important predictors using machine learning models that exhibited remarkably high accuracy. Specifically, the LASSO model demonstrated the highest accuracy level, reaching nearly 98%. Using this highly performing machine learning model, our study successfully identified

significant predictors for the development of hepatitis in patients with DM. By leveraging the superior performance of these machine learning models and the discovery of important factors, we can provide valuable information that can support the development of preventive care and policy targeting more vulnerable populations with DM.

Our study exhibits several limitations. First, the study was conducted on a population of Americans. Thus, different results may be obtained for populations from different countries or cultures. Although a machine-learning classifier was developed for use as an international instrument, it should be applied considering the culture, society, and environment of each country[64]. Therefore, future studies should include global data from other populations.

Second, as our dataset exhibited the cross-sectional nature of the NHANES dataset, it was difficult to determine the future prognosis of the patients. Although we predicted diabetes and related hepatitis at the time of investigation, longitudinal data are required to determine prognoses[65].

Third, the lack of information among the participants may have led to the exclusion of several relevant results. Owing to the limited types of risk factors that were reported in other years and the small sample size, only 1396 people from the NHANES from 2013 to 2018 were selected for this study.

Fourth, although poverty is related to scholarity, data on education levels were not included in the current study. Considering the close correlation between scholarity and the spread of infectious diseases, it is important for the education level to be evaluated. However, the definition of education levels remains controversial and may affect the results[66,67]. Thus, we anticipate further studies on the impact of education levels on HBV or HCV infections among diabetic patients in the future.

Fifth, we evaluated HBV and HCV infections without serological data owing to the inconsistency of variables among the datasets and data unavailability. The NHANES collects data of participants including serological data, which are released in two-year cycles[68]. However, discrepancies in the variables were noted among the serological data collected from 2013 to 2018 for HBV. An additional variable in the laboratory results of the hepatitis B surface antigen, namely "indeterminate", was provided for 2015 to 2018, which differed from the dataset that was collected from 2013–2014 that included only "positive" or "negative". Furthermore, limited serological data for HCV infection could be accessed owing to the data restriction with low precision[68].

Sixth, in the current study, we exclusively assessed existing machine learning models rather than undertaking the development of a new model. Through comprehensive searches of previous studies evaluating machine learning models predicting outcomes related to hepatitis, we finally selected four machine learning models shown higher performance metrics rather than other algorithms, which followed approach taken in a previous study[69]. Furthermore, it is important to note that surpassing the performance of established models falls beyond the scope of our research. In the future, we anticipate the development of novel machine learning models with the potential to enhance the accuracy of hepatitis prediction among diabetes patients. Developing new novel models improved interpretability and usability in clinical practice might contribute to the speed and accuracy of physicians' work, develop early diagnosis and treatment strategies, and improve screening protocols[70,71]. Finally, although the high prevalence of DM in patients with viral hepatitis is mediated by the development of liver cirrhosis[72,73], liver cirrhosis data were not included in the current study. Liver ultrasound data, which provide objective measures for liver cirrhosis manifestations, were not released in 2013–2016. Therefore, further studies that include liver cirrhosis data should be conducted.

Methods

Dataset

The NHANES is a program of studies that is designed to assess the health and nutritional status of adults and children in the US. The NHANES questionnaire consists of demographic, socioeconomic, dietary, and health-related questions. A nationally representative sample of approximately 5000 individuals is gathered per year using this survey through a complex and multistage sampling design and the database is up-dated every two years. Furthermore, the NHAENS gathers data on 60 years and older, African Americans, and Hispanics to produce reliable statistics by reducing bias. Common public medical databases, such as the NHANES, provide researchers with important clues to the causes of diseases based on the distribution of health problems and risk factors in the population. All data are available for download from the NHANES website www.cdc.gov/nchs/nhanes/.

In this study, the NHANES data from 2013 to 2018 were used for the model validation and prediction of HBV or HCV in diabetic patients. We extracted the demographics, body measurements, lipids, and questionnaire data to analyze the association between diabetes and the 12 risk factors for hepatitis. The demographic data included age, gender (male), race (Mexican American, other Hispanic, non-Hispanic White, non-Hispanic Black, and non-Hispanic Asian), and the ratio of family income to poverty (poverty)[74]. This ratio was calculated by dividing the family income by the poverty guidelines for the survey year[74]. Furthermore, BMI and waist circumference were measured. The lipid test indicators included high-density lipoprotein cholesterol (HDL-C) and total cholesterol. The questionnaire data included receiving blood, the hepatitis B vaccine, general health condition, and using needles to inject illegal drugs (illegal drug injection). The 12 predictors that were used in our study were selected based on HBV and HCV guidelines.

For preprocessing and normalization, we followed the following procedures as detailed. NHANES data collection employs a multistage probability sampling design, ensuring a robust and representative dataset for disease prediction and healthcare planning[31]. "Refused" and "Don't know" values were categorized as missing values to prevent potentially misleading predictions. All cases with missing data for hepatitis and variables with 50% or more missing data were excluded to prepare the data. Outliers were detected using the IQR method and then capped within the 25th to 75th percentile range[75]. We perform preprocessing by grouping variables with similar semantic meaning into the same category. For instance, preprocessing involves encoding categorical features

into a set of binary values to indicate the presence or absence of some category. For the NHANES dataset, which comprise a mixture of numerical and non-numerical features, preprocessing tasks are typically transform non-numerical variables into a format suitable for analysis. The variables in the NHANES dataset are one-hot encoded into multiple binary categorical variables where each variable (e.g., "Mexican American", "Other Hispanic") for the different race has a binary value of 0 and 1. These binary values serve to identify whether each subject falls into their respective category for each variable. One-hot encoding was performed on categorical data to implement a machine learning model based on a dataset with a total of 17 characteristics.

Evaluation of diabetes and hepatitis

Diabetes was defined as having a fasting plasma glucose (FPG) ≥ 126 mg/dL, hemoglobin A1c (HbA1c) ≥ 6.5%, or a self-reported diabetes diagnosis during the interview[3].

The participants were divided into two groups: HBV or HCV and non-HBV or non-HCV. The participants were labeled as having HBV or HCV if they answered "Yes" to hepatitis B or C infection, as represented by the question "Has a doctor or other health professional ever told you that you have hepatitis B?" or "Has a doctor or other health professional ever told you that you have hepatitis C?" If the participant answered "No" to all HBV and HCV conditions, the participant was labeled as not having HBV or HCV infection.

Statistical analysis

The continuous variables were represented as the means with standard deviations (SD, normal distribution), and the categorical variables were represented as percent-ages with frequencies. We used the SMOTE to counteract the imbalance in the number of hepatitis and non-hepatitis subjects[76].

SMOTE is an oversampling algorithm that generates additional samples based on the original dataset by setting a specific scale to balance the dataset using over-sampling methods[76]. The importance of the predictors was evaluated and plotted using an importance score to determine the best-performing model. All analyses were performed using the R statistical software 4.12 (The R Foundation for Statistical Computing, USA).

To detect outliers, the study utilized the interquartile range (IQR) method. Outliers were defined as values that exceeded the thresholds of 75th percentile + 1.5 × IQR or were lower than the 25th percentile—1.5 × IQR[75]. Once the outliers were identified, they were capped within the range of the 25th and 75th percentiles.

Machine learning model evaluation

We evaluated four machine-learning models to predict the risk of HBV or HCV infection among patients with diabetes: RF, SVM, XGBoost, and LASSO, which were shown higher performance metrics compared to other algorithms in previous studies for predicting outcomes related to hepatitis[26,29,34,69]. To achieve the best performance, we conducted hyperparameter tuning for each algorithm (Table 4). This included exploring different values for the hyperparameters of the dataset and selecting the ones that provided the most favorable outcomes. We employed a randomized search with ten iterations of three-fold cross-validation to identify the optimal hyperparameters for each of the four models. We tuned the hyperparameters of each algorithm to achieve the best performance. Cross-validation, in which 70% of the data were used to train the model and 30% were used in the prediction test, was performed to validate the prediction effects. The confusion matrix and AUC-ROC were used to select the best prediction model.

The performance metrics included the accuracy, sensitivity, specificity, precision, area under the curve (AUC), and F1 score. As the AUC tended to screen for degradation, we considered the sensitivity and specificity in conjunction with the AUC to minimize unbalanced bias[77]. The sensitivity is the ratio of true negativity that is accurately identified by the test and the specificity is the ratio of true negativity that is correctly identified by the test[78]. The F1 score is a harmonic mean of the accuracy and precision, which enabled us to compare different model performances in identifying actual disease predictions compared to false positives[79]. The importance of the predictors was evaluated and computed using the contribution importance rank of each variable to determine the best-performing model.

Algorithm	Hyperparameters	Range of value	Optimal values
RF	mtry	1 to 10	4
	ntree	100 to 1000	324
SVM	kernel	['linear', 'polynomial', 'radial', 'sigmoid']	'linear'
	cost	0.1 to 10	7.790
	degree	[3–5] except 'linear'	
XGBoost	eta	0.01 to 0.1	0.089
	gamma	0 to 1	0.129
	max_depth	1 to 10	8
	nrounds	10 to 100	40
LASSO	alpha	0 to 1	0.013
	lambda	0 to 1	0.008

Table 4. Search range for hyperparameter tuning. RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting; LASSO, least absolute shrinkage and selection operator.

Data availability

All data in the current analysis are publicly available on the NHANES website (http://www.cdc.gov/nchs/nhanes.htm).

Received: 3 April 2023; Accepted: 4 December 2023
Published online: 06 December 2023

References

1. Tanase, D. M. *et al.* Role of gut microbiota on onset and progression of microvascular complications of type 2 diabetes (T2DM). *Nutrients* **12**, 3719. https://doi.org/10.3390/nu12123719 (2020).
2. Control, C. F. D. National diabetes statistics report: Estimates of diabetes and its burden in the United States, 2014. *Atlanta, GA: US Department of Health and Human Services* (2014).
3. Classifcation and Diagnosis of Diabetes. Standards of medical care in diabetes-2020. *Diabetes Care* **43**, S14-s31 (2020).
4. Deshpande, A. D., Harris-Hayes, M. & Schootman, M. Epidemiology of diabetes and diabetes-related complications. *Phys. Ther.* **88**, 1254–1264. https://doi.org/10.2522/ptj.20080020 (2008).
5. Control, C. F. D. Use of hepatitis B vaccination for adults with diabetes mellitus: Recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Morb. Mortal. Wkly. Rep.* **60**, 1709–1711 (2011).
6. Zibbell, J. E. *et al.* Increases in hepatitis C virus infection related to injection drug use among persons aged ≤30 years - Kentucky, Tennessee, Virginia, and West Virginia, 2006–2012. *MMWR Morb. Mortal. Wkly. Rep.* **64**, 453–458 (2015).
7. Lu, P. J., Hung, M. C., Srivastav, A., Williams, W. W. & Harris, A. M. Hepatitis B vaccination among adults with diabetes mellitus, U.S., 2018. *Am. J. Prev. Med.* **61**, 652–664. https://doi.org/10.1016/j.amepre.2021.04.029 (2021).
8. Gisi, K. *et al.* Hepatitis B and C seroprevalence in patients with diabetes mellitus and its relationship with microvascular complications. *Prz. Gastroenterol.* **12**, 105–110. https://doi.org/10.5114/pg.2016.64748 (2017).
9. Puoti, C., Castellacci, R. & Montagnese, F. Hepatitis C virus carriers with persistently normal aminotransferase levels: Healthy people or true patients?. *Dig. Liver Dis.* **32**, 634–643. https://doi.org/10.1016/s1590-8658(00)80850-6 (2000).
10. Lingala, S. & Ghany, M. G. Natural history of hepatitis C. *Gastroenterol. Clin. North Am.* **44**, 717–734. https://doi.org/10.1016/j.gtc.2015.07.003 (2015).
11. Ba-Essa, E. M., Mobarak, E. I. & Al-Daghri, N. M. Hepatitis C virus infection among patients with diabetes mellitus in Dammam, Saudi Arabia. *BMC Health Serv. Res.* **16**, 313. https://doi.org/10.1186/s12913-016-1578-0 (2016).
12. Kombi, P. K. *et al.* Seroprevalence of hepatitis B and C virus infections among diabetic patients in Kisangani (North-eastern Democratic Republic of Congo). *Pan Afr. Med. J.* **31**, 160. https://doi.org/10.11604/pamj.2018.31.160.17176 (2018).
13. Greca, L. F., Pinto, L. C., Rados, D. R., Canani, L. H. & Gross, J. L. Clinical features of patients with type 2 diabetes mellitus and hepatitis C infection. *Braz. J. Med. Biol. Res.* **45**, 284–290. https://doi.org/10.1590/s0100-879x2012007500013 (2012).
14. Mekonnen, D., Gebre-Selassie, S., Fantaw, S., Hunegnaw, A. & Mihret, A. Prevalence of hepatitis B virus in patients with diabetes mellitus: a comparative cross sectional study at Woldiya General Hospital, Ethiopia. *Pan Afr. Med. J.* **17**, 40. https://doi.org/10.11604/pamj.2014.17.40.2465 (2014).
15. Merza, M. A. Seroprevalence and risk factors of hepatitis B and C viruses among diabetes mellitus patients in Duhok province, Iraqi Kurdistan. *J. Family Med. Prim. Care* **9**, 642–646. https://doi.org/10.4103/jfmpc.jfmpc_1158_19 (2020).
16. Million, Y. *et al.* Hepatitis B and hepatitis C viral infections and associated factors among patients with diabetes visiting gondar referral teaching hospital, Northwest Ethiopia: A comparative cross-sectional study. *J. Hepatocell. Carcinoma* **6**, 143–150. https://doi.org/10.2147/jhc.S222609 (2019).
17. Waljee, A. K. & Higgins, P. D. R. Machine learning in medicine: A primer for physicians. *Am. J. Gastroenterol.* **105**, 1224–1226. https://doi.org/10.1038/ajg.2010.173 (2010).
18. Akbas, A. & Buyrukoglu, S. Stacking ensemble learning-based wireless sensor network deployment parameter estimation. *Arab. J. Sci. Eng.* https://doi.org/10.1007/s13369-022-07365-5 (2022).
19. Doğru, A., Buyrukoğlu, S. & Arı, M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. *Med. Biol. Eng. Comput.* **61**, 785–797. https://doi.org/10.1007/s11517-022-02749-z (2023).
20. Buyrukoğlu, S. & Akbaş, A. Machine learning based early prediction of type 2 diabetes: A new hybrid feature selection approach using correlation matrix with heatmap and SFS. *Balkan J. Electr. Comput. Eng.* **10**, 110–117. https://doi.org/10.17694/bajece.973129 (2022).
21. Uzut, G. & Buyrukoglu, S. Hyperparameter optimization of data mining algorithms on car evaluation dataset. *Euroasia J. Math. Eng. Nat. Med. Sci.* **7**, 70–76 (2020).
22. Fregoso-Aparicio, L., Noguez, J., Montesinos, L. & García-García, J. A. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetol. Metab. Syndr.* **13**, 148. https://doi.org/10.1186/s13098-021-00767-9 (2021).
23. Zou, Q. *et al.* Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **9**, 515. https://doi.org/10.3389/fgene.2018.00515 (2018).
24. Ozyilmaz, L. & Yildirim, T. Artificial neural networks for diagnosis of hepatitis disease. *Surg. Endosc.* https://doi.org/10.1109/IJCNN.2003.1223422 (2003).
25. Edeh, M. O. *et al.* Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease. *Front. Public Health* **10**, 892371. https://doi.org/10.3389/fpubh.2022.892371 (2022).
26. Yağanoğlu, M. Hepatitis C virus data analysis and prediction using machine learning. *Data Knowl. Eng.* **142**, 102087. https://doi.org/10.1016/j.datak.2022.102087 (2022).
27. Bhingarkar, S. *A Comparison of Machine Learning Techniques for Categorization of Blood Donors Having Chronic Hepatitis C Infection* (Springer Nature Singapore, 2022). https://doi.org/10.1007/978-981-16-7610-9_54.
28. Syafa'ah, L., Zulfatman, Z., Pakaya, I. & Lestandy, M. Comparison of machine learning classification methods in hepatitis C virus. *Jurnal Online Informatika* **6**, 73–78. https://doi.org/10.15575/join.v6i1.719 (2021).
29. Obaido, G. *et al.* An interpretable machine learning approach for hepatitis B diagnosis. *Appl. Sci.* **12**, 11127. https://doi.org/10.3390/app122111127 (2022).
30. Mahajan, P., Uddin, S., Hajati, F. & Moni, M. A. Ensemble learning for disease prediction: A review. *Healthcare* **11**, 1808. https://doi.org/10.3390/healthcare11121808 (2023).
31. Dinh, A., Miertschin, S., Young, A. & Mohanty, S. D. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med. Inform. Decis. Mak.* **19**, 211. https://doi.org/10.1186/s12911-019-0918-5 (2019).
32. Li, D., Liu, Z., Armaghani, D. J., Xiao, P. & Zhou, J. Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments. *Sci. Rep.* **12**, 184. https://doi.org/10.1038/s41598-022-05594-0 (2022).
33. Raita, Y. *et al.* Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* **23**, 64. https://doi.org/10.1186/s13054-019-2351-7 (2019).
34. Ahn, S. B. *et al.* Twelve-month post-treatment parameters are superior in predicting hepatocellular carcinoma in patients with chronic hepatitis B. *Liver Int.* **41**, 1652–1661. https://doi.org/10.1111/liv.14820 (2021).

35. Choi, J., Han, S. W., Jun, D. W. & Ahn, S. B. First year clinical parameters are superior to that of the pre-treatment results for hepatocellular carcinoma prediction in patient with chronic hepatitis B. *SSRN* https://doi.org/10.2139/ssrn.3288893 (2018).
36. Wu, Y. *et al.* A prediction nomogram for the 3-year risk of incident diabetes among Chinese adults. *Sci. Rep.* **10**, 21716. https://doi.org/10.1038/s41598-020-78716-1 (2020).
37. Ou, Q. *et al.* LASSO-based machine learning algorithm to predict the incidence of diabetes in different stages. *Aging Male* **26**, 2205510. https://doi.org/10.1080/13685538.2023.2205510 (2023).
38. Han, J.-Y., Kwon, J.-H., Kim, S.-H. & Lee, H. Hepatitis risk in diabetes compared to non-diabetes and relevant factors: A cross-sectional study with national health and nutrition examination survey (NHANES). *Int. J. Environ. Res. Public Health* **20**, 4962. https://doi.org/10.3390/ijerph20064962 (2023).
39. McDonald, S. A. *et al.* Decrease in health-related quality of life associated with awareness of hepatitis C virus infection among people who inject drugs in Scotland. *J. Hepatol.* **58**, 460–466. https://doi.org/10.1016/j.jhep.2012.11.004 (2013).
40. Peyrot, M., Rubin, R. R., Kruger, D. F. & Travis, L. B. Correlates of insulin injection omission. *Diabetes Care* **33**, 240–245. https://doi.org/10.2337/dc09-1348 (2010).
41. Klonoff, D. C. & Perz, J. F. Assisted monitoring of blood glucose: special safety needs for a new paradigm in testing glucose. *J. Diabetes Sci. Technol.* **4**, 1027–1031. https://doi.org/10.1177/193229681000400501 (2010).
42. Garfein, R. S., Vlahov, D., Galai, N., Doherty, M. C. & Nelson, K. E. Viral infections in short-term injection drug users: the prevalence of the hepatitis C, hepatitis B, human immunodeficiency, and human T-lymphotropic viruses. *Am. J. Public Health* **86**, 655–661. https://doi.org/10.2105/ajph.86.5.655 (1996).
43. Lorvick, J., Kral, A. H., Seal, K., Gee, L. & Edlin, B. R. Prevalence and duration of hepatitis C among injection drug users in San Francisco, Calif. *Am. J. Public Health* **91**, 46–47. https://doi.org/10.2105/ajph.91.1.46 (2001).
44. Thomas, D. L. *et al.* Correlates of hepatitis C virus infections among injection drug users. *Medicine* **74**, 212–220. https://doi.org/10.1097/00005792-199507000-00005 (1995).
45. Tseng, F. C. *et al.* Seroprevalence of hepatitis C virus and hepatitis B virus among San Francisco injection drug users, 1998 to 2000. *Hepatology* **46**, 666–671. https://doi.org/10.1002/hep.21765 (2007).
46. Hagan, H. *et al.* Syringe exchange and risk of infection with hepatitis B and C viruses. *Am. J. Epidemiol.* **149**, 203–213. https://doi.org/10.1093/oxfordjournals.aje.a009792 (1999).
47. Degenhardt, L. *et al.* Estimating the burden of disease attributable to injecting drug use as a risk factor for HIV, hepatitis C, and hepatitis B: Findings from the Global Burden of Disease Study 2013. *Lancet Infect. Dis.* **16**, 1385–1398. https://doi.org/10.1016/s1473-3099(16)30325-5 (2016).
48. Eckhardt, B. *et al.* Risk factors for hepatitis C seropositivity among young people who inject drugs in New York City: Implications for prevention. *PLoS One* **12**, e0177341. https://doi.org/10.1371/journal.pone.0177341 (2017).
49. Schillie, S. F., Xing, J., Murphy, T. V. & Hu, D. J. Prevalence of hepatitis B virus infection among persons with diagnosed diabetes mellitus in the United States, 1999–2010. *J. Viral Hepat.* **19**, 674–676. https://doi.org/10.1111/j.1365-2893.2012.01616.x (2012).
50. Guo, X., Jin, M., Yang, M., Liu, K. & Li, J.-W. Type 2 diabetes mellitus and the risk of hepatitis C virus infection: A systematic review. *Sci. Rep.* **3**, 2981. https://doi.org/10.1038/srep02981 (2013).
51. Turk Wensveen, T., Gašparini, D., Rahelić, D. & Wensveen, F. M. Type 2 diabetes and viral infection; cause and effect of disease. *Diabetes Res. Clin. Pract.* **172**, 108637. https://doi.org/10.1016/j.diabres.2020.108637 (2021).
52. Guimarães, L. C. D. C. *et al.* Epidemiology of hepatitis B virus infection in people living in poverty in the central-west region of Brazil. *BMC Public Health* **19**, 443. https://doi.org/10.1186/s12889-019-6828-8 (2019).
53. Janjua, P. Z. & Kamal, U. A. The role of education and income in poverty alleviation: A cross-country analysis. *Lahore J. Econ.* **16**, 143–172 (2011).
54. Terrault, N. A. *et al.* Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. *Hepatology* **67**, 1560–1599. https://doi.org/10.1002/hep.29800 (2018).
55. Ghany, M. G. & Morgan, T. R. Hepatitis C guidance 2019 update: American association for the study of liver diseases-infectious diseases society of America recommendations for testing, managing, and treating hepatitis C virus infection. *Hepatology* **71**, 686–721. https://doi.org/10.1002/hep.31060 (2020).
56. Greene, S. K., Levin-Rector, A., Hadler, J. L. & Fine, A. D. Disparities in reportable communicable disease incidence by census tract-level poverty, New York city, 2006–2013. *Am. J. Public Health* **105**, e27-34. https://doi.org/10.2105/ajph.2015.302741 (2015).
57. Scarponi, C. F. O., Zolnikov, T. R. & Mol, M. P. G. Are waste pickers at risk for hepatitis B and C infections because of poverty or environmental exposures?. *Rev. Soc. Bras. Med. Trop.* **52**, e20190123. https://doi.org/10.1590/0037-8682-0123-2019 (2019).
58. Barré, T. *et al.* Cannabis use is associated with a lower risk of diabetes in chronic hepatitis C-infected patients (ANRS CO22 Hepather cohort). *J. Viral Hepat.* **27**, 1473–1483. https://doi.org/10.1111/jvh.13380 (2020).
59. Shi, L., Fonseca, V. & Childs, B. Economic burden of diabetes-related hypoglycemia on patients, payors, and employers. *J. Diabetes Complicat.* **35**, 107916. https://doi.org/10.1016/j.jdiacomp.2021.107916 (2021).
60. Wasley, A. *et al.* The prevalence of hepatitis B virus infection in the United States in the era of vaccination. *J. Infect. Dis.* **202**, 192–201. https://doi.org/10.1086/653622 (2010).
61. Roberts, H. *et al.* Prevalence of chronic hepatitis B virus (HBV) infection in U.S. households: National Health and Nutrition Examination Survey (NHANES), 1988–2012. *Hepatology* **63**, 388–397. https://doi.org/10.1002/hep.28109 (2016).
62. Zou, B. *et al.* Prevalence of viremic hepatitis C Virus infection by age, race/ethnicity, and birthplace and disease awareness among viremic persons in the United States, 1999–2016. *J. Infect. Dis.* **221**, 408–418. https://doi.org/10.1093/infdis/jiz479 (2020).
63. Chen, M. S. Jr. & Dang, J. Hepatitis B among Asian Americans: Prevalence, progress, and prospects for control. *World J. Gastroenterol.* **21**, 11924–11930. https://doi.org/10.3748/wjg.v21.i42.11924 (2015).
64. Vinuesa, R. *et al.* The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **11**, 233. https://doi.org/10.1038/s41467-019-14108-y (2020).
65. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J. Chronic Dis.* **40**, 373–383. https://doi.org/10.1016/0021-9681(87)90171-8 (1987).
66. Miech, R. A. *et al.* Trends in the association of poverty with overweight among US adolescents, 1971–2004. *Jama* **295**, 2385–2393. https://doi.org/10.1001/jama.295.20.2385 (2006).
67. Luy, M. *et al.* The impact of increasing education levels on rising life expectancy: A decomposition analysis for Italy, Denmark, and the USA. *Genus* **75**, 11. https://doi.org/10.1186/s41118-019-0055-0 (2019).
68. National Center for Health Statistics. National Health and Nutrition Examination Survey. Analytic guidelines; 2011–2014 and 2015–2016. https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx (2022)
69. Ozgur, S., Altinok, Y. A., Bozkurt, D., Saraç, Z. F. & Akçiçek, S. F. performance evaluation of machine learning algorithms for sarcopenia diagnosis in older adults. *Healthcare (Basel)* https://doi.org/10.3390/healthcare11192699 (2023).
70. Sanchez-Martinez, S. *et al.* Machine learning for clinical decision-making: Challenges and opportunities in cardiovascular imaging. *Front. Cardiovasc. Med.* **8**, 765693. https://doi.org/10.3389/fcvm.2021.765693 (2021).
71. Azmi, J. *et al.* A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data. *Med. Eng. Phys.* **105**, 103825. https://doi.org/10.1016/j.medengphy.2022.103825 (2022).
72. Mokdad, A. A. *et al.* Liver cirrhosis mortality in 187 countries between 1980 and 2010: A systematic analysis. *BMC Med.* **12**, 145. https://doi.org/10.1186/s12916-014-0145-y (2014).

73. Hsieh, P. H. *et al.* Association of type 2 diabetes with liver cirrhosis: a nationwide cohort study. *Oncotarget* **8**, 81321–81328. https://doi.org/10.18632/oncotarget.18466 (2017).
74. Moore, K. J., Gauri, A. & Koru-Sengul, T. Prevalence and sociodemographic disparities of Hepatitis C in Baby Boomers and the US adult population. *J. Infect. Public Health* **12**, 32–36. https://doi.org/10.1016/j.jiph.2018.08.003 (2019).
75. Tada, T. *et al.* Improvement of liver stiffness in patients with hepatitis C virus infection who received direct-acting antiviral therapy and achieved sustained virological response. *J. Gastroenterol. Hepatol.* **32**, 1982–1988. https://doi.org/10.1111/jgh.13788 (2017).
76. Ren, Z. *et al.* Psychological impact of COVID-19 on college students after school reopening: A cross-sectional study based on machine learning. *Front. Psychol.* **12**, 641806. https://doi.org/10.3389/fpsyg.2021.641806 (2021).
77. Symum, H. & Zayas-Castro, J. L. Prediction of chronic disease-related inpatient prolonged length of stay using machine learning algorithms. *Healthc. Inform. Res.* **26**, 20–33. https://doi.org/10.4258/hir.2020.26.1.20 (2020).
78. Sidey-Gibbons, J. A. M. & Sidey-Gibbons, C. J. Machine learning in medicine: A practical introduction. *BMC Med. Res. Methodol.* **19**, 64. https://doi.org/10.1186/s12874-019-0681-4 (2019).
79. Goutte, C. & Gaussier, E. *A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation* (Springer, 2005).

Author contributions

S.H.K., S.H.P., and H.L. conceived the study and wrote the manuscript. H.L. provided clinical expertise and supervised the study. S.H.K., S.H.P., and H.L. prepared the relevant NHANES data and carried out the analyses the data. All the authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.