Received 12 April 2021; revised 22 August 2021 and 29 September 2021; accepted 30 September 2021. Date of publication 6 October 2021; date of current version 9 November 2021. The review of this article was arranged by Editor P. Bonato.

Digital Object Identifier 10.1109/OJEMB.2021.3117872

Prediction for the Risk of Multiple Chronic Conditions Among Working Population in the United States With Machine Learning Models

JINGMEI YANG [1], XINGLONG JU [2,3], FENG LIU [4], ONUR ASAN [4], TIMOTHY S. CHURCH [5], AND JEFF O. SMITH [5]

[1] Division of System Engineering, Boston University, Boston, MA 02246 USA
[2] Price College of Business, University of Oklahoma, Norman, OK 73019 USA
[3] School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853 USA
[4] School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030 USA
[5] Catapult Health Inc., Dallas, TX 75254 USA

CORRESPONDING AUTHOR: Feng Liu (e-mail: fliu22@stevens.edu)

Jingmei Yang and Xinglong Ju contributed equally to this work.

ABSTRACT *Objective:* Chronic diseases have become the most prevalent and costly health conditions in the healthcare industry, deteriorating the quality of life, adversely affecting the work productivity, and costing astounding medical resources. However, few studies have been conducted on the predictive analysis of multiple chronic conditions (MCC) based on the working population. *Results:* Seven machine learning algorithms are used to support the decision making of healthcare practitioner on the risk of MCC. The models were developed and validated using checkup data from 451,425 working population collected by the healthcare providers. Our result shows that all proposed models achieved satisfactory performance, with the AUC values ranging from 0.826 to 0.850. Among the seven predictive models, the gradient boosting tree model outperformed other models, achieving an AUC of 0.850. *Conclusions:* Our risk prediction model shows great promise in automating real-time diagnosis, supporting healthcare practitioners to target high-risk individuals efficiently, and helping healthcare practitioners tailor proactive strategies to prevent the onset or delay the progression of the chronic diseases.

INDEX TERMS Multiple chronic conditions, machine learning, predictive analysis, health informatics.

IMPACT STATEMENT A machine learning framework is implemented to automate the screening of high risk individuals for multiple chronic conditions in the working population.

I. INTRODUCTION

Chronic conditions have been the most prevalent and costly health issues in the past decades. Chronic diseases can lead to hospitalization, disability, reduced quality of life, and even mortality. According to a recent report, approximately 11.7%, 46%, 9.8%, and 36.3% of adults in the United States were inflicted by high blood pressure, high blood cholesterol, diabetes, and obesity, respectively [1], [2], [3], [4]. Besides, heart attack, heart failure, heart diseases in general, and stroke were responsible for 15%, 11%, 25%, and 5% of death in the United States, respectively [1], [5], [6], [7]. Furthermore, chronic conditions have become a key factor for high healthcare expenditures and decreased productivity for working professionals [8], [9], [10]. The prevalence of multimorbidity defined as the co-occurrence of two or more chronic diseases, has dramatically grown in the past decade [11]. Notably, in 2014, about 50% of adults in the United States lived with at least one chronic condition, and 26% of adults suffered from multiple chronic conditions (MCC) [12]. Moreover, individuals with multimorbidity are more intensive health service users. They expend significantly higher health expenditures than individuals without any chronic conditions or with a single chronic condition [13]. Therefore, proactive prevention is necessary to improve the quality of life and to reduce the cost of healthcare. In fact, the identification of high-risk individuals in the early stages is the key to provide cost-effective prevention strategies. Risk prediction for chronic diseases has been a hot topic among researchers and clinicians. With the progress in sophisticated machine learning algorithms and advanced technologies of data collection and storage, researchers have achieved promising results of predicting the risk of having a single chronic disease [14], [15]. In particular, state-of-the-art machine learning algorithms, such as deep neural network, support vector machine, logistic regression, random forest, and decision tree, have been broadly implemented and validated to predict the risk of having chronic diseases, such as hypertension [16], cardiovascular disease [17], [18], heart disease [19], and diabetes mellitus [20], [21].

Previous explorations have been focused on the prediction of a single chronic disease [22], [23], [24], [25]. For example, there are risk estimation models for hypertension patients [22], [23], [24], [25]. Also, Luo *et al.* [26], Casanova *et al.* [27], and Cahn *et al.* [28] built risk models aiming for diabetes. Seyednasrollah *et al.* and Chen *et al.* achieved a high predictive power in obesity risk prediction [29] [30]. Prediction of a single chronic disease can result in the failure of early diagnosis for other coexisting chronic diseases. Plus, building a single machine learning model for each chronic disease can be time-consuming. Currently, to the best our knowledge, there are no research done to predict the MCC. However, the characteristic of common risk factors, the tendency of co-existence for different combinations of chronic diseases, and the surging financial burden from MCC suggest that it is worthwhile to develop a risk prediction model for MCC. Part of the reason for lacking such a model can be attributed to the scarce of a dataset that covers MCC. The available datasets are usually dedicated to one particular type of disease. Fortuitously, over the past seven years, we meticulously collected a large volume of patient records in clinical practices [10], keeping track of participants' health conditions, which provides a unique opportunity for this retrospective study of early detection of MCC. The data were collected from worksite preventive checkups so that it can provide valuable information about MCC for the U.S. working population.

In this paper, we employed interpretable machine learning algorithms to address classification problem on whether or not a participant has a high risk of MCC, to identify the associated factors, and to explore the interactions among those risk factors. To the best of our knowledge, this is the first study of MCC risk prediction for a large volume of the US working population. The developed risk prediction tool for MCC was implemented in real clinical applications for yearly worksite preventive checkups. The developed tool takes a patient's biometric screening data and provides an interpretable prediction in real-time with the probability of having MCC. This model developed from a large scale dataset can effectively inform healthcare practitioners of patients' risk in order to guide prevention and intervention decisions.

II. RESULTS

A. DATA DESCRIPTION

The chronic conditions of interest are high blood pressure, high cholesterol, diabetes, and obesity. Extracted from the Catapult Health database, the dataset consists of 451,425 records, among which 137,118 records, 77,149 records, and 237,159 records were collected in 2018, in 2017, and 2012-2016.[1] All the records are collected during the worksite checkup. In particular, 301,631 worksite participants were included. Among them, 10,427 participants (3.5%) were flagged as high-risk individuals with at least one chronic condition, and 291,204 participants (96.5%) were flagged as low-risk individuals without MCC. Each participant's flag was given by a certified healthcare professional based on her/his judgment at each checkup. All records extracted from the Catapult Health database were fully de-identified, and the consent on data usage for research purposes was obtained at each checkup.

B. RISK FACTORS

We first excluded insignificant features based on t-test and chi-squared test. Then, we applied univariate selection to explore the strength of the association of each feature with the target variable individually. The features with weak associations were excluded and 20 features remained. Finally, we applied a tree-based feature selection approach to filter out less relevant predictors. Given the feature importance, top 15 important features were selected as significant predictors: diastolic blood pressure (DBP), systolic blood pressure (SBP), glucose (GLU), triglycerides (TGS), total cholesterol (TCHOL), body mass index (BMI), low density lipoprotein-cholesterol (LDL), weight, age, alanine transaminase (ALT), blood pressure assessment (BPAssessment), abdominal circumference (AbdominalCir), metabolic syndrome risk (METS_Risks), diabetes assessment (DMAssessment), and BMI assessment. The feature importance is given in Fig. 1. It is worthnoting that the selected features have overlaps with the important important features reported by other researchers. For example, Olivares *et al.* reports obesity, central obesity, higher BMI, higher waist circumference, higher triglycerides and blood glucose are contributing significantly to hypertension and diabetes [31]. In Ref [32], the authors reports sex, age, self-identified ethnicity, family history of premature cardiovascular disease, smoking status, diabetes status, systolic blood pressure, and the ratio of total cholesterol to high density lipoprotein cholesterol concentrations are contributing to cardiovascular disease. In Ref [33], the authors show the main factors of hypertension and type 2 diabetes include age, sex, smoking, exercise, family history, dietary habits, body mass index (BMI), and waist circumference.

C. MCC PREDICTION FRAMEWORK

We aim to integrate the machine learning framework into our MCC risk screening workflow. Traditionally, the workflow is

[1] Data as well as will be available upon reasonable request to the authors.

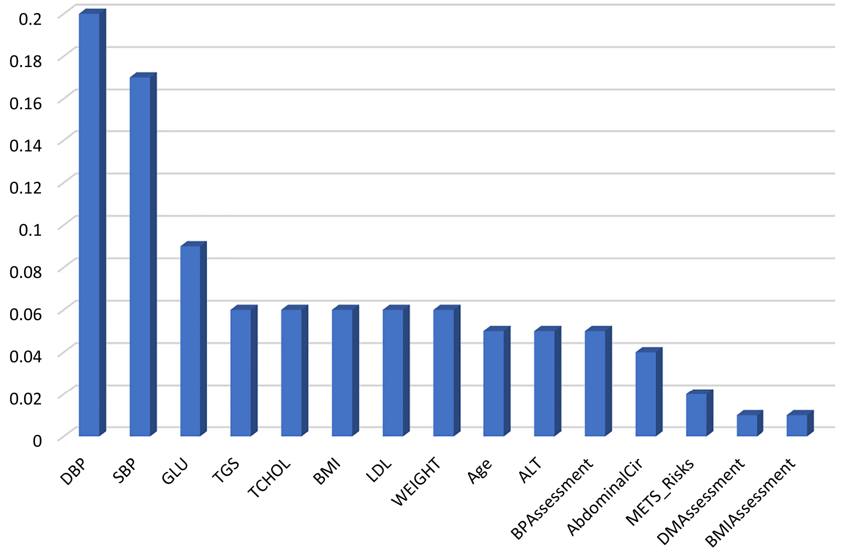

FIG. 1. **Feature importance.**

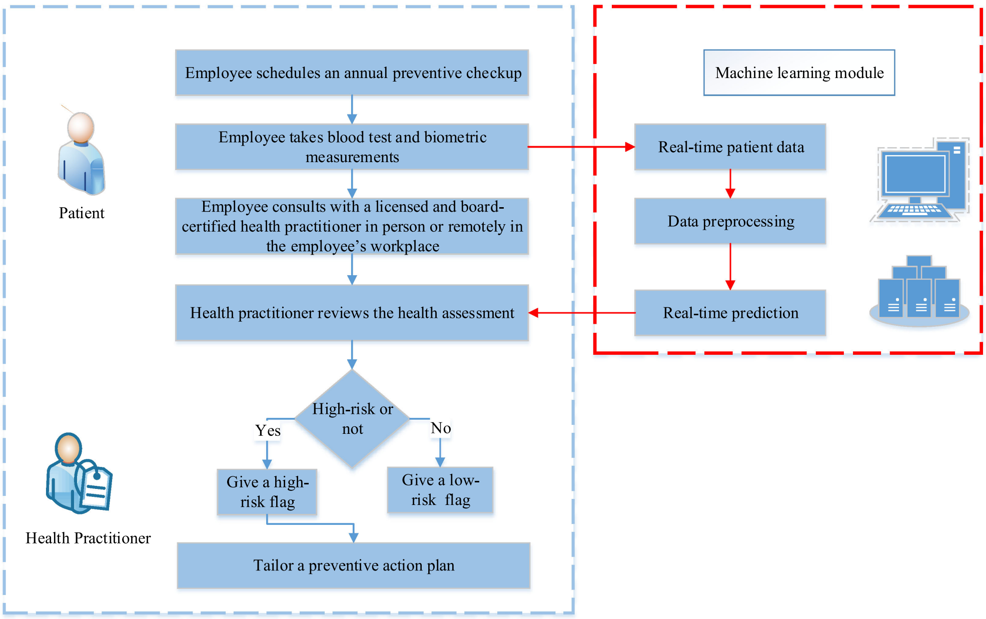
Patient
Employee schedules an annual preventive checkup
Employee takes blood test and biometric measurements
Employee consults with a licensed and board-certified health practitioner in person or remotely in the employee's workplace
Health practitioner reviews the health assessment
High-risk or not
Yes
No
Give a high-risk flag
Give a low-risk flag
Tailor a preventive action plan
Health Practitioner
Machine learning module
Real-time patient data
Data preprocessing
Real-time prediction

FIG. 2. **Proposed workflow with machine learning module for supporting decision making.**

shown in the left part of Fig. 2. With the developed machine learning module, we use predicted risk value to facilitate the healthcare practitioners' efficient screening, as illustrated in Fig. 2. The predictive model can provide real-time prediction, automating the screening procedure.

D. MACHINE LEARNING ALGORITHMS

Seven machine learning algorithms were used and evaluated, including k-nearest neighbors (kNN), decision tree (DT), random forest (RF), gradient boosting tree (GBT), logistic regression (LR), support vector machine (SVM), and Naive Bayes (NB). It is also worth noting that deep learning models can make predictions with high accuracy in numerous scenarios [34]. However, in our study, predictive accuracy was not the only factor we took into consideration. We prioritized model interpretability, computational efficiency, and applicability to our existing platform. As a result, deep learning was not considered in this research.

E. HYPERPARAMETER TUNING

We applied the grid search approach to tune the hyperparameters of each algorithm, and we found that each of the fine-tuned models achieved better predictive performance than default parameter settings.

TABLE I Hyperparameter Tuning

Algorithm	Hyperparameter	Optimal value	Range
kNN	Neighbors	19	[1, 2, 3, 4, 5, 6, 7, 8, 9, $\cdots$, 18, 19, 20]
	P	1	[1, 2, 3, 4]
DT	Criterion	entropy	[entropy, gini]
	Max depth	6	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
RF	Boostrap	true	[true, false]
	Max depth	9	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	Max features	sqrt	[auto, sqrt]
	N estimators	81	[1, 3, 5, 7, 9, 11, 13, $\cdots$, 91, 93, 95, 97, 99]
GB	Learning rate	0.1	[0.001,0.01,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1,10,100]
	Max depth	4	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	N estimatiors	39	[1, 3, 5, 7, 9, 11, 13, $\cdots$, 91, 93, 95, 97, 99]
SVM (rbf)	C	10	[0.0001, 0.001, 0.01, 0.05, 0.1, 1, 10, 100]
	Gamma	0.0001	[0.0001, 0.001, 0.01, 0.05, 0.1, 1, 10, 100]
SVM (linear)	C	0.1	[0.0001, 0.001, 0.01, 0.05, 0.1, 1, 10, 100]
LR	C	100	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	Max iteration	1000	[100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
	Penalty	ℓ_2	[ℓ_1, ℓ_2]

As can be seen from Table I, kNN achieved the best performance when the number of neighbors was set to 19, and the Manhattan distance was used to measure the similarity. The DT tends to overfit as DT is constructed greedily. Pruning strategies, such as setting the maximum depth of a tree, are necessary to alleviate this problem. Built on the entropy criterion, the DT model with a maximum level of 6 achieved the best performance. The RF classifier achieved the best performance when 81 parallel tree learners were combined. Within each tree, the maximum depth was set to 9. The value of bootstrap was set to be true. In contrast to the RF classifier, the GBT classifier was sequentially built with 39 weak decision tree learners. The maximum depth of each tree was set to 4, and a learning rate of 0.1 achieved the best performance. In an effort to build a parsimonious LR model, a ℓ_2 penalty term was added to the objective function, and the hyperparameter C that controls the regularization power, was set to 100. Similarly, in the SVM classifier, C is responsible for smoothing decision boundary. The fine-tuned SVM model was the model built with a radial basis function (RBF) kernel, a gamma value of 0.0001, and a C value of 10. Regarding the NB algorithm, there were no hyperparameters needed to be tuned.

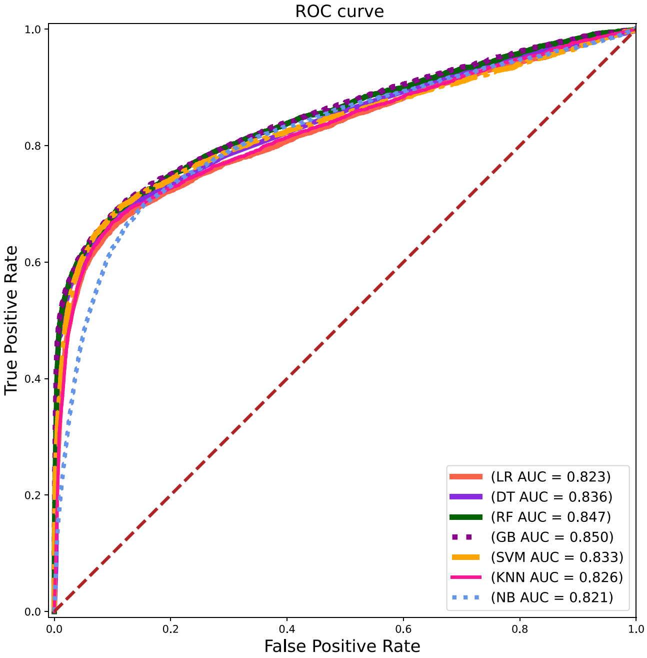

FIG. 3. ROC curves.

TABLE II Performance Metrics for Optimized Models

Algorithm	AUC	Sensitivity	Accuracy
GB	0.850	0.692	0.882
RF	0.846	0.683	0.888
SVM (RBF)	0.833	0.686	0.885
DT	0.837	0.678	0.890
NB	0.821	0.664	0.867
LR	0.823	0.717	0.809
kNN	0.826	0.699	0.851

F. MODEL EVALUATION

Fig. 3 shows that the AUC values achieved by the GBT classifier, RF classifier, SVM classifier, DT classifier, NB classifier, LR classifier, and kNN classifier were 0.850, 0.846, 0.833, 0.837, 0.821, 0.823, 0.826 respectively.

As can be seen from Table II, the kNN model yielded the least accurate result compared to other models. We reasoned that its predictive performance might be hampered by the presence of noisy data, as the kNN algorithm was sensitive to the local structure of neighboring data points. Additionally, its computation time was significantly high because it required the calculation of pair-wise similarity (defined by distance metrics) between a new instance and all training instances. When it comes to the performance of the DT classifier, although it was easy to understand and decipher by visualizing the tree structure, it insufficiently outlined the relationship between the predictors and the risk of having MCC. Similar to the DT classifier, the LR classifier achieved a poor predictive performance. Consistent with other research, the ensemble models showed better performance compared to other single learners. Both the RF classifier and the GB classifier achieved high predictive accuracy. Typically, the GB classifier outperformed the RF classifier in terms of AUC and sensitivity, partly because of its distinguishing characteristic, that new tree learners were added to correct the mistakes made by previous learners.

III. DISCUSSION

It is well known that chronic conditions have become the top public health concerns. Chronic conditions, such as high blood pressure, high cholesterol, diabetes, and obesity, are often preventable and manageable [35]. Numerous studies have stated the importance of identifying associated risk factors and the benefits of detection, prevention, and intervention of chronic conditions in the early stages [36], [37], [38]. A risk assessment tool plays a key role in cost-effective preventive programs and personalized interventions. Driven by this motivation, we applied seven popular machine learning algorithms to identify risk factors, to explore the effect of these factors, and then to develop a risk prediction model based on the recognized factors. This predictive tool can provide a more comprehensive picture of an individual's risk of having MCC, empowering healthcare practitioners to make a tailored recommendation for each patient.

To our best knowledge, this is one of the very few studies applying machine learning to electronic health record data from such a large working-age population, demonstrating the predictive powers of a wide range of algorithms in risk assessment of MCC. Besides, a broad spectrum of risk factors was taken into consideration, including demographic characteristics, family medical history, and modifiable lifestyle factors, as well as the laboratory test results. This study also demonstrated the effectiveness of rebalancing techniques to address the imbalanced class distribution problem. Concerning the sample size, we included a total of 306,631 participants. It is noted that a large dataset paves the way for a more robust model, which is less susceptible to underfitting [39]. Apart from the relatively large sample size, it is worthwhile to mention that the data were collected from multiple cities across the nation, which lays a solid ground for model generalization. It is also well recognized that as the number of risk factors increases, the complexity of models increases. A complex model is prone to overfitting [40]. By tuning model hyperparameters, we mitigated the overfitting problem.

Moreover, some studies only focused on a specific chronic condition. Compared with disease-specific models to estimate the risk of developing specific chronic outcomes, our study presents a multifaceted risk tool that can predict MCC concurrently. If implemented in clinical practices, this risk prediction model can be used as a supportive tool to quantify individuals' risk and to aid healthcare practitioners' decision making.

As the medical data is dramatically growing, the demand of healthcare practitioners with domain knowledge to decipher the clinical data surges. Additionally, as the population is aging, the healthcare practitioner shortage becomes pronounced. A large number of population has limited access to healthcare services in developing countries due to the dearth of medical education, the shortage of medical professionals, and the insufficiency of resources [41], [42], [43]. Encouragingly, machine learning algorithms offer possibilities to fill this gap. This model is capable of automatically identifying patients at high risk of having MCC. Medical practitioners could be trained to use this system to alleviate the shortage of healthcare practitioners and extend access to healthcare services.

The study has several limitations. We evaluated our models based on the performance comparison of different models and validated the optimal model performance on the test dataset. External validation in clinical practices would yield more accurate and reliable predictions [44]. Moreover, during the data preprocessing step, the removal of incomplete records inevitably led to information loss. More sophisticated imputation strategies, such as the k-nearest neighbors based imputation strategy, would mitigate the undesirable effect. Also, the high risk labels are given by the certified nurses and doctors, the labels may subject to subjective judgement of different individuals. For future research, it is definitely valuable to use the time series and other state-of-the-art models such as recurrent neural networks, Transformers to train the longitudinal time series data and make predictions. This research focus on predicting the risk of having MCC, and the specific chronic disease can be further identified by other downstream machine learning models or preferably by doctors.

IV. CONCLUSION

In this study, we aimed to develop a prognostic model to identify high-risk individuals with MCC. We built the predictive models for MCC using the undersampling approach for the training dataset, and we evaluated with the imbalanced testing dataset using accuracy, AUC, and sensitivity. Overall, all proposed predictive models for MCC achieved satisfactory AUC values, ranging from 0.809 to 0.8307. Among the seven models, the gradient boosting model yielded the best performance with the highest AUC and accuracy. Our real-time risk prediction model has been deployed in the staging environment and is improving the screening efficiency in clinical practices. The incorporation of our model into clinical practices offers great potential in supporting the healthcare practitioners to target high-risk patients and tailor personalized recommendations to high-risk individuals, thus improving decision making, enhancing the effectiveness of prevention programs, hopefully promoting individual's quality of life and productivity, and eventually reducing healthcare expenditure.

V. MATERIAL AND METHODS

A. DATA PREPROCESSING

The features include self-reported information (demographic characteristics, modifiable lifestyle risk factors, and family medical history) and clinical measurements. We observed that for 15 features, there are more than 50% of the total records have missing values. Those 15 features were dropped, with a

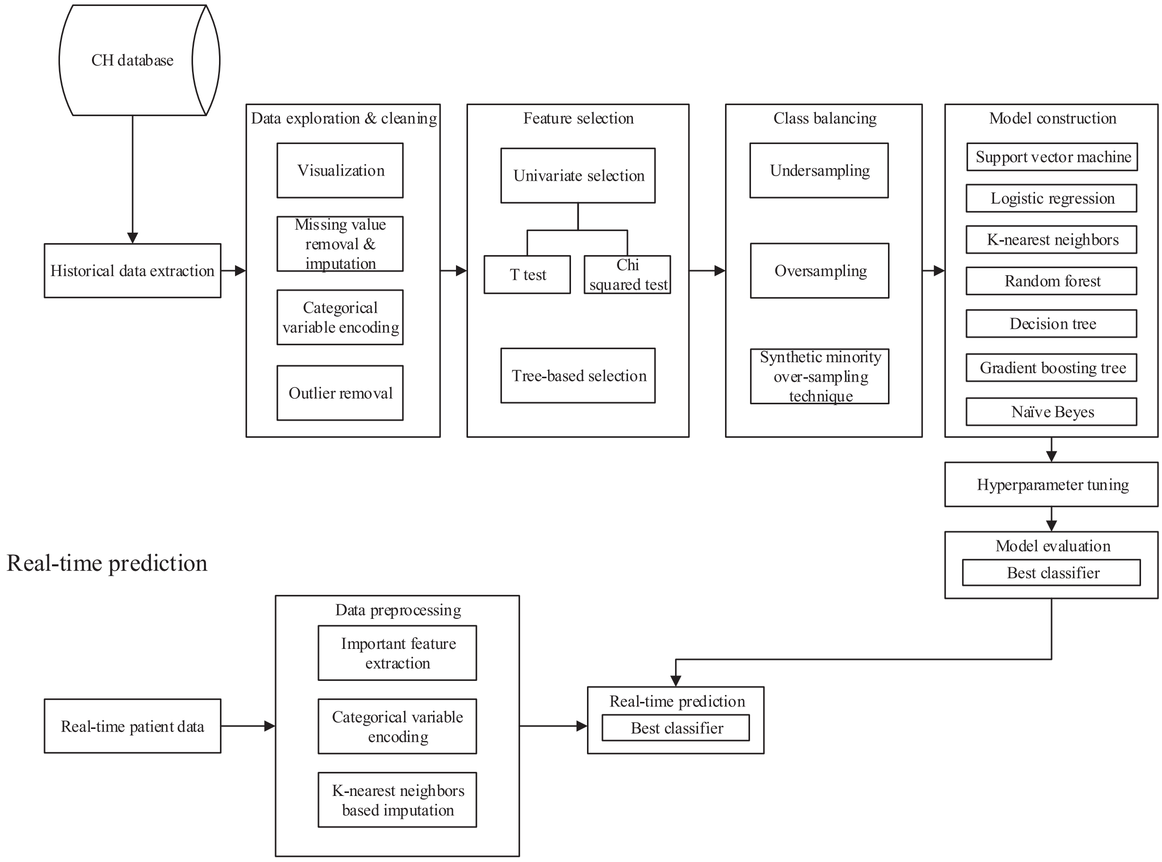

FIG. 4. **Machine learning framework.**

total of 39 features remaining, listed in Table 1 of the supplemental material.

For this new dataset, we first performed a descriptive analysis. Contingency tables were used to summarize the statistics of categorical features. Histograms and box plots were used to identify data anomalies in features with continuous values. Additionally, χ^2 tests and t-test were performed to examine statistical significance. In the second step, categorical data were encoded, and missing values are imputed, or the whole records were dropped during the data cleaning process.

Furthermore, the dataset contains heterogeneous features with different magnitudes and ranges. Hence, the min-max scaling method was applied to improve computational efficiency. After the rescaling step, samples were randomly split into a training set (70%) and a testing set (30%). Besides, 10-fold cross-validation was used to tune hyperparameters. The detailed illustrations for the machine learning pipeline with the data preprocessing procedure and the data splitting step are shown in Fig. 4 and Fig. 5. We use the historical records to predict the MCC risk at the studied checkup. The average time difference between two checkups is around one year.

B. IMBALANCED CLASSES

One difficulty in our prediction is the highly skewed class distribution, a typical phenomenon in medical analytics, with the number of low-risk participants without MCC far exceeding the number of high-risk participants. Poor performance in identifying the minority yet significant class can be attributed to the imbalanced class distribution, given that most machine learning algorithms expect an equal number of observations in each class, and they were inherently designed to maximize the overall accuracy.

In an attempt to mitigate the effect of imbalanced class distribution, we use three techniques, including randomly oversampling the minority class technique, undersampling the majority class technique, and synthetic minority oversampling technique (SMOTE) [45], [46]. The oversampling method randomly replicates the observations in the minority class [47]. Conversely, the undersampling technique randomly samples a subset of observations from the majority class. SMOTE is an advanced oversampling method that searches k nearest neighbors of each observation in the minority class and then generates synthetic samples. We run a set of experiments, and the results show that the undersampling method outperformed other approaches in terms of model generalization and the computational cost. Specifically, the oversampling method led to model overfitting, and the SMOTE method had a higher computation cost, as our dataset is considerably large. Table III summarize the experiment results of the undersampling method. As can be seen from the results, all classifiers achieve decent performance when the ratio of

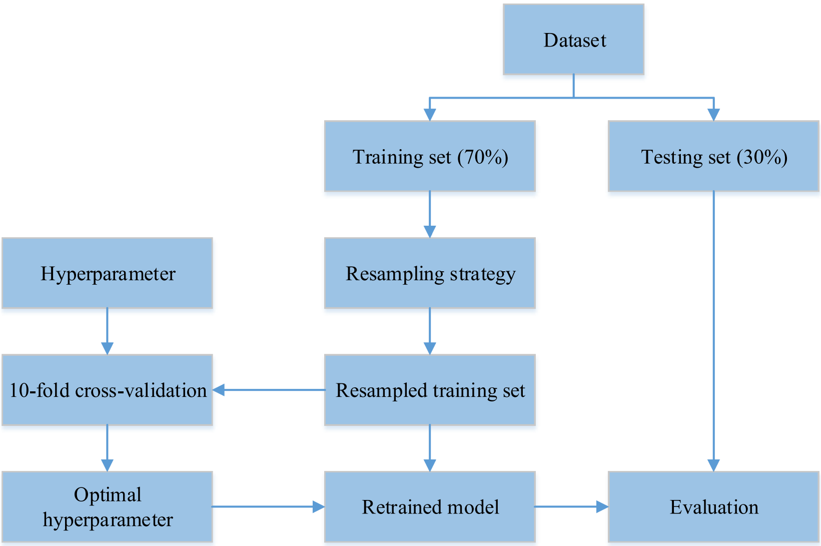

FIG. 5. Data splitting procedure.

TABLE III Performance Metrics for Seven Algorithms Using Different Ratios

Algorithm	Ratio	AUC	Sensitivity	Accuracy
DT	0.2	0.830	0.520	0.968
	0.4	0.833	0.561	0.955
	0.6	0.834	0.596	0.945
	0.8	0.837	0.686	0.884
	1	0.837	0.678	0.890
RF	0.2	0.848	0.515	0.970
	0.4	0.848	0.576	0.955
	0.6	0.848	0.626	0.933
	0.8	0.847	0.665	0.902
	1	0.846	0.683	0.888
GB	0.2	0.848	0.525	0.968
	0.4	0.849	0.583	0.954
	0.6	0.849	0.629	0.931
	0.8	0.849	0.672	0.902
	1	0.850	0.692	0.882
kNN	0.2	0.817	0.506	0.959
	0.4	0.822	0.606	0.932
	0.6	0.824	0.648	0.905
	0.8	0.826	0.678	0.879
	1	0.826	0.699	0.851
LR	0.2	0.818	0.467	0.961
	0.4	0.821	0.586	0.937
	0.6	0.822	0.640	0.903
	0.8	0.822	0.683	0.860
	1	0.823	0.717	0.809
NB	0.2	0.823	0.665	0.869
	0.4	0.824	0.672	0.865
	0.6	0.823	0.686	0.857
	0.8	0.823	0.697	0.843
	1	0.821	0.664	0.867
SVM	0.2	0.792	0.529	0.964
	0.4	0.818	0.597	0.945
	0.6	0.828	0.673	0.894
	0.8	0.832	0.685	0.886
	1	0.833	0.686	0.885

the number of samples in the minority class over the number of samples in the majority class reaches to 1. Therefore, the ratio was set to 1.

C. PERFORMANCE METRICS

Accuracy is defined as the proportion of correct predictions in a population [48]. Along with accuracy, the area under the curve (AUC) measures the capability that a model differentiates high-risk participants with MCC from low-risk participants [49]. Sensitivity measures the proportion of participants actually with MCC who are correctly identified by the models [50], [51].

REFERENCES

[1] E. J. Benjamin *et al.*, "Heart disease and stroke statistics—2017 update: A report from the American Heart Association," *Circulation*, vol. 135, no. 10, pp. e146–e603, 2017.

[2] E. J. Benjamin *et al.*, "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e66, 2019.

[3] C. D. Fryar, T.-C. Chen, and X. Li, *Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, 1999–2010*. US Department of Health and Human Services, Centers Disease Control Prevention, Washington DC, USA: 2012. [Online]. Available: https://www.cdc.gov/nchs/data/databriefs/db103.pdf

[4] R. Merai, "CDC grand rounds: A public health approach to detect and control hypertension," *Morbidity Mortality Weekly Rep.*, vol. 65, no. 45, pp. 1261–1264, 2016.

[5] D. Mozaffarian *et al.*, "Heart disease and stroke statistics–2016 update a report from the American Heart Association," *Circulation*, vol. 133, no. 4, pp. e38–e48, 2016.

[6] CDC. Multiple Cause of Death 1999-2018 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, 2020. Accessed: Oct 28, 2021. [Online]. Available: https://wonder.cdc.gov/mcd-icd10.html

[7] Q. Yang *et al.*, "Vital signs: Recent trends in stroke death rates—United States, 2000–2015," *Morbidity Mortality Weekly Rep.*, vol. 66, no. 35, p. 933, 2017.

[8] T. D. Vuong, F. Wei, and C. J. Beverly, "Absenteeism due to functional limitations caused by seven common chronic diseases in us workers," *J. Occup. Environ. Med./Amer. College Occup. Environ. Med.*, vol. 57, no. 7, pp. 779–784, 2015.

[9] G. R. B. Asay, K. Roy, J. E. Lang, R. L. Payne, and D. H. Howard, "Peer reviewed: Absenteeism and employer costs associated with chronic diseases and health risk factors in the us workforce," *Preventing Chronic Dis.*, vol. 13, no. E-141, 2016, Art. no. 150503.

[10] J. Yang *et al.*, "Blood pressure states transition inference based on multistate markov model," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 1, pp. 237–246, Jan. 2021.

[11] A. J. K. Pefoyo *et al.*, "The increasing burden and complexity of multimorbidity," *BMC Public Health*, vol. 15, no. 1, pp. 1–11, 2015, Art. no. 415.

[12] B. W. Ward, "State and regional prevalence of diagnosed multiple chronic conditions among adults aged $\geq$ 18 years—United States, 2014," *Morbidity Mortality Weekly Rep.*, vol. 65, no. 29, pp. 735–738, Jul. 2016.

[13] K. Thavorn *et al.*, "Effect of socio-demographic factors on the association between multimorbidity and healthcare costs: A population-based, retrospective cohort study," *BMJ open*, vol. 7, no. 10, 2017, Art. no. e017264..

[14] T. B. Murdoch and A. S. Detsky, "The inevitable application of Big Data to healthcare," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[15] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017.

[16] A. Wang, N. An, Y. Xia, L. Li, and G. Chen, "A logistic regression and artificial neural network-based approach for chronic disease prediction: A case study of hypertension," in *Proc. IEEE Int. Conf. Internet Things, IEEE Green Comput. Commun. IEEE Cyber, Phys. Social Comput.*, 2014, pp. 45–52.

[17] E. Y. F. Wan *et al.*, "Classification rule for 5-year cardiovascular diseases risk using decision tree in primary care chinese patients with type 2 diabetes mellitus," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 15238.

[18] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, 2017, Art. no. e0174944.

[19] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," in *Proc. IEEE Conf. Inf. Commun. Technol.*, 2013, pp. 1227–1231.

[20] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," *BMC Med. Informat. Decis. Mak.*, vol. 10, no. 1, pp. 1–7, 2010.

[21] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, 2018, Art. no. 515.

[22] N. I. Parikh *et al.*, "A risk score for predicting near-term incidence of hypertension: The framingham heart study," *Ann. Intern. Med.*, vol. 148, no. 2, pp. 102–110, 2008.

[23] N. P. Paynter, N. R. Cook, B. M. Everett, H. D. Sesso, J. E. Buring, and P. M. Ridker, "Prediction of incident hypertension risk in women with currently normal blood pressure," *Amer. J. Med.*, vol. 122, no. 5, pp. 464–471, 2009.

[24] C. Ye *et al.*, "Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning," *J. Med. Internet Res.*, vol. 20, no. 1, 2018, Art. no. e22.

[25] F. López-Martínez, A. Schwarcz, E. R. Núñez-Valdez, and V. García-Díaz, "Machine learning classification analysis for a hypertensive population as a function of several risk factors," *Expert Syst. Appl.*, vol. 110, pp. 206–215, 2018.

[26] G. Luo, "Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction," *Health Inf. Sci. Syst.*, vol. 4, no. 1, pp. 1–9, 2016.

[27] R. Casanova *et al.*, "Prediction of incident diabetes in the jackson heart study using high-dimensional machine learning," *PLoS One*, vol. 11, no. 10, 2016, Art. no. e0163942.

[28] A. Cahn, A. Shoshan, T. Sagiv, R. Yesharim, I. Raz, and R. Goshen, "Use of a machine learning algorithm improves prediction of progression to diabetes," Diabetes vol. 67, (Supplement 1), pp. 1286–P, 2018.

[29] F. Seyednasrollah *et al.*, "Prediction of adulthood obesity using genetic and childhood clinical risk factors in the cardiovascular risk in young finns study," *Circulation: Cardiovasc. Genet.*, vol. 10, no. 3, 2017, Art. no. e001554.

[30] H. Chen *et al.*, "Using blood indexes to predict overweight statuses: An extreme learning machine-based approach," *PLoS One*, vol. 10, no. 11, 2015, Art. no. e0143003.

[31] D. E. Olivares, F. R. Chambi, E. M. Chañi, W. J. Craig, S. O. Pacheco, and F. J. Pacheco, "Risk factors for chronic diseases and multimorbidity in a primary care context of central Argentina: A web-based interactive and cross-sectional study," *Int. J. Environ. Res. Public Health*, vol. 14, no. 3, 2017, Art. no. 251.

[32] R. Pylypchuk *et al.*, "Cardiovascular disease risk prediction equations in 400000 primary care patients in New Zealand: A derivation and validation study," *Lancet*, vol. 391, no. 10133, pp. 1897–1907, 2018.

[33] V. Martín *et al.*, "Comparison of body mass index (bmi) with the cunbae body adiposity estimator in the prediction of hypertension and type 2 diabetes," *BMC Public Health*, vol. 16, no. 1, pp. 1–8, 2015.

[34] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Informat.*, vol. 69, pp. 218–229, 2017.

[35] S.-Y. Wu and A. Green, "The growing crisis of chronic disease in the united states," RAND Corporation, 2000.

[36] A. K. Parekh, R. A. Goodman, C. Gordon, H. K. Koh, and H. I. W. on Multiple Chronic Conditions, "Managing multiple chronic conditions: A strategic framework for improving health outcomes and quality of life," *Public Health Rep.*, vol. 126, no. 4, pp. 460–471, 2011.

[37] T. M. Dall, M. V. Storm, A. P. Semilla, N. Wintfeld, M. O'Grady, and K. V. Narayan, "Value of lifestyle intervention to prevent diabetes and sequelae," *Amer. J. Prev. Med.*, vol. 48, no. 3, pp. 271–280, 2015.

[38] A. L. Albright and E. W. Gregg, "Preventing type 2 diabetes in communities across the us: The national diabetes prevention program," *Amer. J. Prev. Med.*, vol. 44, no. 4, pp. e346–S351, 2013.

[39] P.-H. C. Chen, Y. Liu, and L. Peng, "How to develop machine learning models for healthcare," *Nat. Mater.*, vol. 18, no. 5, pp. 410–414, 2019.

[40] E. W. Steyerberg, "Overfitting and optimism in prediction models," in *Clinical Prediction Models*. Springer, 2019, pp. 95–112.

[41] J. L. Clarke, S. Bourn, A. Skoufalos, E. H. Beck, and D. J. Castillo, "An innovative approach to health care delivery for patients with chronic conditions," *Popul. Health Manage.*, vol. 20, no. 1, pp. 23–30, 2017.

[42] A. Darzi and T. Evans, "The global shortage of health workers-an opportunity to transform care," *Lancet*, vol. 388, no. 10060, pp. 2576–2577, 2016.

[43] S. Naicker, J. Plange-Rhule, R. C. Tutt, and J. B. Eastwood, "Shortage of healthcare workers in developing countries—Africa," *Ethnicity Dis.*, vol. 19, no. 1, 2009, Art. no. 60.

[44] K.-H. Yu and I. S. Kohane, "Framing the challenges of artificial intelligence in medicine," *BMJ Qual Saf*, vol. 28, no. 3, pp. 238–241, 2019.

[45] N. Rosso and P. Giabbanelli, "Accurately inferring compliance to five major food guidelines through simplified surveys: Applying data mining to the UK national diet and nutrition survey," *JMIR Public Health Surveill.*, vol. 4, no. 2, 2018, Art. no. e9536.

[46] G. E. Batista, R. C. Prati, and M. C. Monard, "Balancing strategies and class overlapping," in *Proc. Int. Symp. Intell. Data Anal.*, Springer, 2005, pp. 24–35.

[47] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.

[48] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, 1997.

[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[50] D. M. Powers, "*Evaluation: From precision, recall and f-measure to Roc, informedness, markedness and correlation*," 2011, *arXiv:2010.16061*.

[51] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2011.