Journal of Biosocial Science (2024), 56: 3, 426–444
doi:10.1017/S0021932024000063

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Machine learning models for prediction of double and triple burdens of non-communicable diseases in Bangladesh

Md. Akib Al-Zubayer[1], Khorshed Alam[2,3], Hasibul Hasan Shanto[1], Md. Maniruzzaman[1], Uttam Kumar Majumder[1] and Benojir Ahammed[1]

[1]Statistics Discipline, Khulna University, Khulna, Bangladesh, [2]School of Business, University of Southern Queensland, Toowoomba, QLD, Australia and [3]Centre for Health Research, University of Southern Queensland, Toowoomba, QLD, Australia

Corresponding author: Benojir Ahammed; Emails: benojir@stat.ku.ac.bd; benojirstat@gmail.com

(Received 30 March 2023; revised 24 October 2023; accepted 22 January 2024; first published online 20 March 2024)

Abstract

Increasing prevalence of non-communicable diseases (NCDs) has become the leading cause of death and disability in Bangladesh. Therefore, this study aimed to measure the prevalence of and risk factors for double and triple burden of NCDs (DBNCDs and TBNCDs), considering diabetes, hypertension, and overweight and obesity as well as establish a machine learning approach for predicting DBNCDs and TBNCDs. A total of 12,151 respondents from the 2017 to 2018 Bangladesh Demographic and Health Survey were included in this analysis, where 10%, 27.4%, and 24.3% of respondents had diabetes, hypertension, and overweight and obesity, respectively. Chi-square test and multilevel logistic regression (LR) analysis were applied to select factors associated with DBNCDs and TBNCDs. Furthermore, six classifiers including decision tree (DT), LR, naïve Bayes (NB), k-nearest neighbour (KNN), random forest (RF), and extreme gradient boosting (XGBoost) with three cross-validation protocols (K2, K5, and K10) were adopted to predict the status of DBNCDs and TBNCDs. The classification accuracy (ACC) and area under the curve (AUC) were computed for each protocol and repeated 10 times to make them more robust, and then the average ACC and AUC were computed. The prevalence of DBNCDs and TBNCDs was 14.3% and 2.3%, respectively. The findings of this study revealed that DBNCDs and TBNCDs were significantly influenced by age, sex, marital status, wealth index, education and geographic region. Compared to other classifiers, the RF-based classifier provides the highest ACC and AUC for both DBNCDs (ACC = 81.06% and AUC = 0.93) and TBNCDs (ACC = 88.61% and AUC = 0.97) for the K10 protocol. A combination of considered two-step factor selections and RF-based classifier can better predict the burden of NCDs. The findings of this study suggested that decision-makers might adopt suitable decisions to control and prevent the burden of NCDs using RF classifiers.

Keywords: classification; machine learning; non-communicable diseases

Introduction

The world is inundated with non-communicable diseases (NCDs) which have emerged as one of the most serious public health concerns (Biswas *et al.*, 2019; Bista *et al.*, 2020). NCDs are any illness that lasts a long time or has lengthy consequences and is caused by a non-infectious and non-transmissible aetiology (WHO, 2020). NCDs are the prime and most important causes of mortality and infirmity worldwide (Vos *et al.*, 2020). Therefore, reducing the incidence of NCDs is one of the most significant priorities of the UN Sustainable Development Goals (SDGs). In addition, NCDs are responsible for approximately 41 million deaths worldwide annually, with

approximately 77% of those occurring in low- and middle-income countries (WHO, 2022). More than one-third of these deaths (15 million people out of a total of 41 million) occurred in people aged 30–69 years (WHO, 2020). Moreover, it is anticipated that the number of victims will increase from 38 to 52 million between 2012 and 2030 (Bigna and Noubiap, 2019). Based on the projections by the Harvard Public Health School, the accumulated production loss caused by NCDs will exceed approximately US$47 trillion by 2030 (Bloom *et al.*, 2011). The increased life expectancy and declining fertility in South Asia caused by demographic transition and economic growth elevated the incidence of NCDs (Islam *et al.*, 2014). The crucial risk factors for the rapid growth of NCDs comprise high cholesterol, elevated blood pressure, unsatisfactory intake of fruit and vegetables, expanded fat in the blood, unhealthy diet, overweight or obesity, physical inactivity, smoking, and overconsumption of alcohol (Saeed, 2013). In a progression of steps, these risk factors constitute severe NCDs, such as respiratory diseases, diabetes, cardiovascular diseases, and cancers, which are the 'group of four' and are liable for 80% of all fatalities brought on by NCDs (Bista *et al.*, 2020). On the contrary, most of these risk factors can be prevented or controlled (Zaman *et al.*, 2015). The worldwide strategic programme of WHO for the control and prevention of NCDs concentrated on nine elective global objectives with a target of achieving a 25% considerable decline in early death from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases concerning premature mortality from NCDs by 2025 (WHO, 2013).

As one of the most densely populated countries with a population of over 169 million, Bangladesh recently graduated to a lower-middle-income country (BBS, 2022). NCDs are the leading cause of mortality and morbidity in Bangladesh, accounting for 67% of total deaths (WHO, 2016). In the past few decades, the incidence of NCDs has grown at an alarming rate in Bangladesh (Islam *et al.*, 2021). In addition, Bangladesh is currently in a more aggressive manifestation of epidemiological transition, and the number of deaths caused by NCDs is anticipated to rise (Islam *et al.*, 2021). Recent evidence from a community-based survey revealed an expansion in the occurrence of NCDs as compared with preceding years (Khalequzzaman *et al.*, 2017). Notably, 26%, 21%, and 5% of the population are overweight and have hypertension and diabetes, respectively (Zaman *et al.*, 2015). The Bangladesh Demographic and Health Survey (BDHS) 2011 revealed that the incidence rate of hypertension, diabetes as well as overweight or obesity was 48.0%, 11.0%, and 25.3%, respectively (Al Kibria *et al.*, 2021). Catastrophic health expenditure for screening, diagnosis, and treatment of NCDs deter various individuals in developing countries like Bangladesh from seeking the care they need (Fottrell *et al.*, 2018).

Similar to other developing countries, NCDs are a widespread threat to public health in Bangladesh. Thus, it is now one of the government's top priorities and various international organisations working in Bangladesh to reduce the morbidity and mortality caused by NCDs (Riaz *et al.*, 2020). Although Bangladesh has made progress in various aspects towards achieving SDGs, NCDs continue to pose a significant challenge to its general public health (Khalequzzaman *et al.*, 2017). A number of studies have investigated the prevalence and risk factors of NCDs in Bangladesh using traditional techniques such as univariate, bivariate, and, in some instances, regression analysis to understand the situation (Saeed, 2013; Khalequzzaman *et al.*, 2017; Russell *et al.*, 2019; Bista *et al.*, 2020; Yosef, 2020). Most of the work focused on a specific burden. To date, however, no studies have yet investigated the coexistence of multiple NCDs and made predictions using the machine learning (ML)-based algorithms. While the traditional methods are widely used, the insight from these types of analysis is limited to exploratory and inferential analyses. To control the health risk, however, it is important to predict the prevalence of these diseases using ML-based methods in addition to conventional techniques. Recently, disease prediction has increased substantial devotion from the information learning research community. Recently, the presence of multiple diseases is common in the human body and this is significantly increasing worldwide. The double and the triple burden of NCDs (DBNCDs and TBNCDs) as a form of the presence of multiple diseases in a body are relatively recent phenomena, and only a limited number of studies have investigated the risk factors of DBNCDs and TBNCDs (Al-Zubayer *et al.*, 2021). Although the use of ML has significantly

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

improved the diagnosis, treatment, and prognosis of several diseases, the application of ML-based approaches for predicting the DBNCDs and TBNCDs has not yet been properly investigated (Maniruzzaman *et al.*, 2019; Maniruzzaman *et al.*, 2020; Islam *et al.*, 2022). Thus, this research was inspired to simultaneously study on both DBNCDs and TBNCDs to overcome the challenges of NCDs in Bangladesh. Therefore, this study aimed to: (i) explore the prevalence of DBNCDs and TBNCDs, (ii) identify the most significant determinants of DBNCDs and TBNCDs, and (iii) suggest an ML-based classifier to predict the DBNCDs and TBNCDs.

This study has made significant technical contributions to the field of NCDs. A novel methodology has been implemented to assess the risk factors associated with DBNCDs and TBNCDs in Bangladesh, thereby addressing a critical research gap. It has utilised a range of ML techniques to predict DBNCDs and TBNCDs and has pioneered a balanced dataset generation method for enhancing model accuracy. Additionally, a robust 10-fold cross-validation (CV) process has been established to rigorously evaluate the model performance. A comprehensive set of performance metrics has been applied, ensuring a thorough assessment of the classifier efficacy. These contributions have significantly advanced NCDs research in Bangladesh and in similar jurisdictions, providing a valuable framework for further research in public health domains.

Materials and methods

Study design

This study used secondary data extracted from the BDHS 2017–2018, which was a nationwide representative cross-sectional household survey. The survey was administered by the National Institute of Population Research and Training between October 2017 and March 2018. It collected individual participant data using a two-stage stratified cluster sampling process. First, a total of 675 enumeration areas (EAs) (250 in urban regions and 425 in rural areas) were selected based on a probability that was proportionate to the size of each EA. Second, 30 households were selected from each EA. This was done to obtain a statistically reliable assessment of important demographic and health characteristics for the whole country. The procedure is described in depth in the BDHS 2017–2018 report (NIPORT and ICF, 2020). The survey selected a total of 20,250 households and information from 89,819 individuals within those households was gathered. All the adult males and females who were at least 18 years old had their blood glucose and blood pressure levels checked. Finally, the participants in the study were limited to 12,151 adults (5,238 and 6,913 men and women, respectively) aged 18 years and above. Finally, BDHS-2011 dataset was also used to check the efficiency of our proposed system.

Outcome variable

The study primarily used two outcome variables, namely DBNCDs and TBNCDs, which were calculated from three NCDs including diabetes, hypertension as well as overweight and obesity. If any two of the diseases existed in a person's body, he/she was classified to have DBNCDs. On the contrary, if a person suffered from the considered three diseases of diabetes, hypertension, and being overweight or obese, then the person was classified as having TBNCDs.

Diabetes

The respondents' fasting plasma glucose levels and whether they took any diabetic medication were taken into account to determine whether or not the person had diabetes. If an individual had a fasting plasma glucose reading of more than 7.0 mmol/L and/or she/he was taking any medication for diabetes, then the subject was classified to have diabetes disease; otherwise, the subject was classified as normal (WHO, 2016).

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Hypertension

The value of blood pressure was utilised to assess hypertension. The interviewers took the respondents' blood pressure thrice over the course of each interview: at the very start, at the exact centre and at the final closing of the session. The average value of these measurements was included in the BDHS, 2017–2018 dataset, which was used to measure hypertension. A respondent was classified to have hypertension if she/he had an average systolic blood pressure of $\geq$ 140mmHg and/or average diastolic blood pressure of $\geq$ 90 mmHg and/or was taking any medicine or drug to lower blood pressure (Ahammed *et al.*, 2021).

Overweight and obesity

Body mass index (BMI) was estimated by taking participants' weight in kilograms and dividing it by the square of their height in metres. An adult BMI of 25.0 to <30 kg/m^2 is considered overweight, and 30.0 or higher is obese. In this study, if a respondent's BMI is $\geq$25 kg/m^2, then the respondent is classified as overweight or obese otherwise stated not (Bista *et al.*, 2020; NIPORT and ICF, 2020; Ahammed *et al.*, 2022).

Explanatory variables

Several explanatory variables were included in this study to find the associated risk factors for DBCNDs and TBNCDs in Bangladesh (Saeed, 2013; Khalequzzaman *et al.*, 2017; Russell *et al.*, 2019; Bista *et al.*, 2020; Yosef, 2020). The variables which were considered for the analysis were respondent's age (<35, 35–44, 45–54, 55–64, and $\geq$65 years), sex (male and female), marital status (never married, currently married, and formerly/ever married), education level (no education/ preschool, primary, secondary, and higher), employment status (working and not working), family size ($\leq$4 and >4), wealth index (poorest, poorer, middle, richer, and richest), height (short, medium, and tall), caffeinated drink (no and yes), smoking status (no and yes), place of residence (urban and rural), regionality (Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur, and Sylhet), community poverty (low and high), and community literacy (low and high). The community poverty and community literacy were generated by aggregating the wealth index and education level, respectively, and then categorised as high or low depending on the distribution of the ratio values that were evaluated for each cluster. Moreover, the ratio value was examined using a histogram, and if the data were normally distributed, then the mean value was used as the cut-off point for the category; otherwise, the median value was utilised (Al-Zubayer *et al.*, 2021).

Risk factors selection techniques

The selection of risk factors is either through variable selection, feature selection, or a subset of features in the fields of statistics and ML. Several risk factor selection techniques are used to choose the variables that provide the most valuable information to enhance the performance of ML-based algorithms. Thus, it is important to select the most important and significant factors for easy operation of an ML-based system, and these include clear interpretation of the findings, minimise the amount of expense and time spent on computations, removing the dimensionality issue, optimise the accuracy of the classification, and minimise the problem of over-fitting (James *et al.*, 2013; Liu and Motoda, 2012; Maniruzzaman *et al.*, 2022). Chi-square analysis was used in this study to measure the association of explanatory variables with DBNCDs and TBNCDs. A multilevel logistic regression (LR) model was used to select the important risk factors for DBNCDs and TBNCDs. All the risk factors were selected using a p-value of <0.05.

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Imbalanced maintenance procedure and formation of balanced datasets

A dataset is imbalanced if one class label exceeds the other class label in size. The imbalanced outcome variable in data poses practical challenges for the community of ML-based research (Libbrecht and Noble, 2015). An ML-based system favours the majority class when classifying imbalanced data. Therefore, this study adopted a combination of oversampling and under-sampling techniques to address this issue. Oversampling is an approach in which samples from the minority class are randomly chosen with replacements and added to the training dataset. Consequently, ML-based classifier performance is enhanced (Matsuoka, 2021; Maniruzzaman *et al.,* 2022). Under-sampling is another strategy in which samples from the majority class are randomly chosen without replacement until the label's balance is attained (Bunkhumpornpat *et al.,* 2011).

The study used a mixture of over- and under-sampling strategies to balance the outcome variables category label (No vs. Yes) for both DBNCDs and TBNCDs. In the case of DBNCDs, the database that was used for the investigation included a total of 1,735 (14.3%) and 10,416 (85.7%) individuals who had and who did not have DBNCDs, respectively. Herein, the ratio between yes and no was 1:6. The study took 3.501 times the positive class (Yes) $(3.501 \times 1735) = 6{,}075$ respondents having DBNCDs using oversampling and took 6,076 respondents who did not have DBNCDs from 10,416 using under-sampling to minimise the disparity between the numbers of samples found in each category. In terms of TBNCDs, the database that was used for the investigation consisted of a total of 278 (2.3%) and 11,873 (97.7%) respondents who had and did not have TBNCDs, respectively. Herein, the ratio between Yes and No was too imbalanced. Thus, the study took 21.85 times the number of individuals in the positive class (Yes) $(21.85 \times 278) = 6{,}075$ respondents having TBNCDs using oversampling and also took 6,076 respondents who did not have TBNCDs from 11,873 using under-sampling to lessen the difference between the numbers of samples found in each category.

Data partitioning

The process of data partitioning is the CV protocol. It was used to create two distinct subsets from the original dataset, which are the training and the validation/test sets. Several CV techniques are available that may be used to partition the dataset into smaller segments to minimise variability (Maniruzzaman *et al.*, 2020; Islam *et al.*, 2022). The 10-fold CV procedure was frequently used to partition the data (Maniruzzaman *et al.*, 2020; Islam *et al.*, 2022). This protocol involves dividing the dataset into 10 equivalent portions, 9 of which were utilised as a training set, while the remaining portion served as a validation/test set. Then, ML-based methods were trained on the training set and predicted the class label on the test set, and then the classification accuracy of each protocol was computed. This procedure was repeated 10 times to minimise the variability and then computed the average classification accuracy of ML-based methods. This procedure is the K10 CV protocol, wherein 10 represents the total number of partitions that occur throughout the ML-based process. Similarly, K2 and K5 data partition protocols were the most popular data splitting procedures, which were based on the training set's accessible percentage of 50% as well as 80%, respectively, whereas the remaining portions were the validation or test set. Three partition protocols were employed in the present study, sequentially labelled as K2, K5, and K10.

ML approach

The application of an ML-based technique is to make predictions about DBNCDs and TBNCDs. Several ML-based techniques can be potentially employed for classification and regression. Among them, the study implemented the following six classifiers: decision tree (DT) (Quinlan, 1986), logistics regression (LR) (Maniruzzaman *et al.*, 2018), k-nearest neighbours (KNN) (Hastie *et al.*, 2009), naïve Bayes (NB) (Hossain and Chetty, 2011), random forest (RF) (Breiman, 2001),

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

and extreme gradient boosting (XGBoost) (Bentéjac *et al.*, 2021). The considered ML-based techniques were the most popular and essential classification method for biomedicine investigations (Liao *et al.*, 2016; Shah *et al.*, 2019) and also specify dummy indicators and may be extended for the classification of different NCDs such as diabetes, hypertension, and overweight or obesity (Maniruzzaman *et al.*, 2018). Finally, the seven performance parameters were applied to measure the performance of the classifiers, which are accuracy (ACC), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-measure (FM), and area under the curve (AUC).

Performance evaluations

Several statistical parameters can be used to evaluate the performance of different ML-based classifiers. This study primarily used accuracy (ACC) and AUC. In addition, SE, SP, PPV, NPV, and FM were used to measure the performance of ML-based classifiers. However, all the parameters were calculated using true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). The statistical parameters of the classifiers were defined as follows:

$$\text{ACC (\%)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \quad (1)$$

$$\text{SE (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (2)$$

$$\text{SP (\%)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (3)$$

$$\text{PPV (\%)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (4)$$

$$\text{NPV (\%)} = \frac{\text{TN}}{\text{TN} + \text{FN}} \times 100 \quad (5)$$

$$\text{F} - \text{measure (\%)} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \times 100 \quad (6)$$

$$\text{AUC} = \int_0^1 \text{ROC(t)dt} \quad (7)$$

where ROC is the receiver operating character, t = (1 – specificity), and ROC (t) is sensitivity. The value of AUC ranged from 0 to 1.

Statistical analysis

A summary of the entire analysis system is exhibited in Fig. 1, while the overall flowchart of the proposed ML-based approach is shown in Fig. 2. For all the explanatory variables, the basic characteristics of the current study are shown as frequency and percentage. Then, the dataset was properly verified and weighed for further analysis. The weighted prevalence of DBNCDs and TBNCDs was presented in the bivariate analysis. Chi-square tests showed the initial relationship between DBNCDs and TBNCDs with explanatory variables. Furthermore, a multilevel LR analysis was conducted after adjusting the covariates to examine the explanatory variables' associations with DBNCDs and TBNCDs. Multilevel analysis is beneficial when samples are produced from a complicated survey design that includes multistage sampling such as the BDHS data because it exposes more accurate findings and lessens the effects of dependence across sampling clusters (Merlo *et al.*, 2005; Rabe-Hesketh and Skrondal, 2006; Ma *et al.*, 2017). The multilevel LR results

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

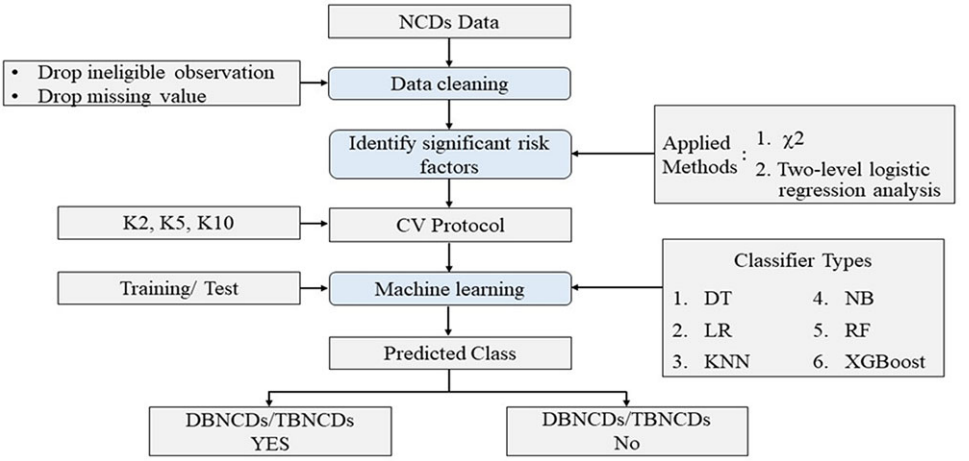

Figure 1. Overview of the entire analysis system.

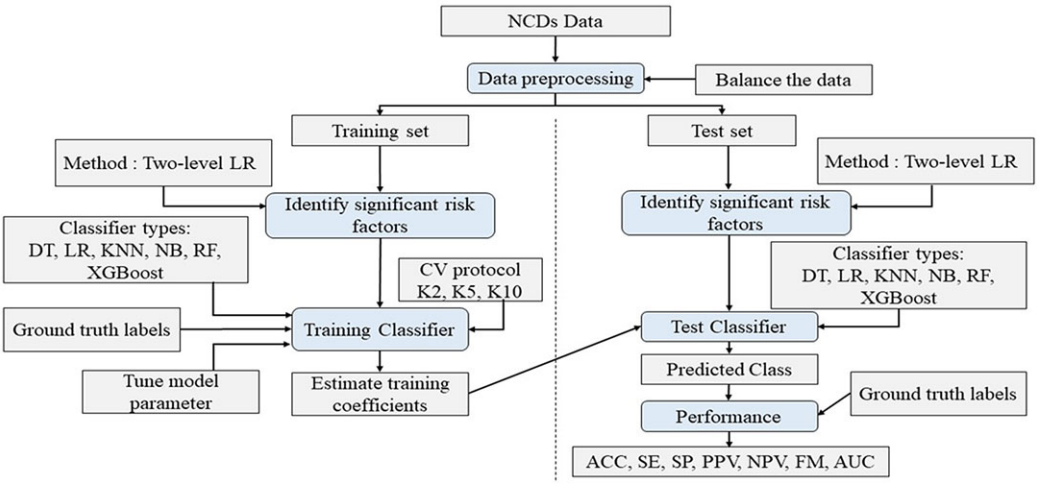

Figure 2. The training/test set paradigm of the ML-based system.

were presented using a p-value (<0.05). Finally, the significant variables obtained from multilevel LR analysis were further incorporated into ML-based algorithms to predict the performance of classifiers for both DBNCDs and TBNCDs. All the statistical analyses of this study were performed using Stata 16 and R version 4.2.2.

Results

Basic characteristics of DBNCDs and TBCNDs

The basic characteristics of the respondents are presented in Table 1. Approximately, 45% of individuals were below 35 years old. More than half of the adults in the survey were females (56.9%). Most adults (80%) were currently married and working in some capacity (61%). Primary education was completed by 30.3% of respondents, while about one-fifth of respondents belonged to the richest quintile (22.2%). Approximately half of the respondents had more than four family members (57.6%) and had medium height (48.6%). Most adults neither consume caffeine (92.8%)

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 1. Basic characteristics and prevalence of DBNCDs and TBNCDs in Bangladesh

	Distribution	DBNCDs	Prevalence of DBNCDs		TBNCDs	Prevalence of TBNCDs	
Variables	*Total, n* (%)	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value
Total	12151(100)		14.3(13.5–15.2)			2.3(2.0–2.6)	
Initial-level factors							
Age group (in years)				**<0.001**			**<0.001**
<35	5451(44.9)	402(7.4)	7.3(6.5–8.1)		36(0.7)	0.6(0.4–0.9)	
35–44	2455(20.2)	458(18.7)	18.7(17.0–20.6)		61(2.5)	2.6(1.9–3.4)	
45–54	1710(14.1)	368(21.5)	21.9(19.5–24.4)		75(4.4)	4.4(3.4–5.7)	
55–64	1376(11.3)	297(21.6)	21.2(18.8–23.8)		61(4.4)	4.2(3.2–5.6)	
≥65	1159(9.5)	210(18.1)	16.8(14.4–19.4)		45(3.9)	3.1(2.0–4.6)	
Sex				**<0.001**			**0.017**
Male	5238(43.1)	610(11.6)	11.6(10.6–12.8)		100(1.9)	1.9(1.6–2.4)	
Female	6913(56.9)	1125(16.3)	16.0(14.9–17.0)		178(2.6)	2.4(2.0–2.8)	
Marital status				**<0.001**			**<0.001**
Never married	1263(10.4)	56(4.4)	4.2(3.1–5.7)		4(0.3)	0.3(0.1–1.0)	
Currently married	9717(80.0)	1468(15.1)	14.9(14.0–15.9)		234(2.4)	2.3(2.0–2.7)	
Formerly/ever married	1171(9.6)	211(18.0)	17.4(15.1–19.8)		40(3.4)	2.8(2.0–4.1)	
Education level				**<0.001**			**0.022**
No education, preschool	2962(24.4)	365(12.3)	12.3(11.0–13.7)		51(1.7)	1.6(1.1–2.1)	
Primary	3677(30.3)	486(13.2)	13.4(12.2–14.8)		77(2.1)	2.2(1.7–2.8)	
Secondary	3519(29.0)	544(15.5)	15.2(13.8–16.7)		98(2.8)	2.7(2.1–3.4)	
Higher	1993(16.4)	340(17.1)	16.5(14.6–18.5)		52(2.6)	2.3(1.7–3.2)	
Employment status				**<0.001**			**<0.001**
Not working	4733(39.0)	814(17.2)	16.8(15.6–18.2)		139(2.9)	2.7(2.2–3.4)	
Working	7418(61.0)	921(12.4)	12.4(11.5–13.4)		139(1.9)	1.9(1.5–2.2)	
Family size				**0.027**			0.49
≤4	5146(42.4)	761(14.8)	14.6(13.4–15.9)		112(2.2)	2.1(1.7–2.7)	
>4	7005(57.6)	974(13.9)	13.7(12.7–14.8)		166(2.4)	2.2(1.9–2.7)	
Wealth index				**<0.001**			**<0.001**
Poorest	2364(19.5)	143(6.0)	6.2(5.1–7.5)		14(0.6)	0.6(0.4–1.1)	
Poorer	2301(18.9)	192(8.3)	8.6(7.4–9.9)		19(0.8)	0.9(0.6–1.5)	
Middle	2408(19.8)	281(11.7)	11.6(10.2–13.1)		25(1.0)	1.1(0.7–1.7)	
Richer	2384(19.6)	373(15.6)	15.9(14.2–17.8)		54(2.3)	2.2(1.6–3.0)	
Richest	2694(22.2)	746(27.7)	27.7(25.7–29.9)		166(6.2)	6.1(5.1–7.2)	
Height				**0.002**			0.56

(*Continued*)

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 1. (*Continued*)

	Distribution	DBNCDs	Prevalence of DBNCDs		TBNCDs	Prevalence of TBNCDs	
Variables	*Total, n* (%)	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value
Short	4008(33.0)	635(15.8)	15.8(14.5–17.2)		91(2.3)	2.2(1.7–2.7)	
Medium	5902(48.6)	795(13.5)	13.1(12.1–14.2)		142(2.4)	2.2(1.9–2.7)	
Tall	2241(18.4)	305(13.6)	13.6(12.0–15.5)		45(2.0)	2.1(1.5–2.9)	
Caffeinate drinks				**<0.001**			**<0.001**
No	11277(92.8)	1547(13.7)	13.6(12.8–14.5)		238(2.1)	2.0(1.7–2.4)	
Yes	874(7.2)	188(21.5)	21.9(18.7–25.4)		40(4.6)	4.7(3.3–6.7)	
Smoking status				0.472			0.239
No	10301(84.8)	1461(14.2)	14.0(13.1–14.9)		243(2.4)	2.2(1.9–2.6)	
Yes	1850(15.2)	274(14.8)	14.7(12.9–16.7)		35(1.9)	1.9(1.3–2.8)	
Cluster-level factors							
Place of residence				**<0.001**			**<0.001**
Urban	4350(35.8)	797(18.3)	18.9(17.3–20.5)		150(3.4)	3.4(2.8–4.2)	
Rural	7801(64.2)	938(12.0)	12.4(11.4–13.4)		128(1.6)	1.7(1.4–2.1)	
Division				**<0.001**			**0.003**
Barisal	1265(10.4)	191(15.1)	14.1(11.8–16.6)		27(2.1)	1.9(1.3–2.7)	
Chittagong	1647(13.6)	307(18.6)	18.5(16.3–20.9)		51(3.1)	3.1(2.3–4.1)	
Dhaka	1592(13.8)	265(16.6)	16.0(14.0–18.3)		51(3.2)	3.3(2.2–4.2)	
Khulna	1678(13.1)	278(16.6)	15.5(13.6–17.6)		45(2.7)	2.2(1.6–3.2)	
Mymensingh	1376(11.3)	138(10.0)	9.5(7.8–11.5)		18(1.3)	1.3(0.8–2.0)	
Rajshahi	1591(13.1)	199(12.5)	11.7(9.6–14.1)		30(1.9)	1.4(0.9–2.1)	
Rangpur	1569(12.9)	187(11.9)	10.6(8.7–12.8)		28(1.8)	1.3(0.8–2)	
Sylhet	1433(11.8)	170(11.9)	11.2(9.2–13.6)		28(2)	1.6(1–2.4)	
Community poverty				**<0.001**			**<0.001**
Low	6132(50.6)	1136(18.5)	18.3(17.0–19.6)		208(3.4)	3.3(2.8–3.9)	
High	6019(49.5)	599(9.9)	10.1(9.2–11.1)		70(1.2)	1.2(0.9–1.5)	
Community literacy				**<0.001**			**<0.001**
Low	6351(53.3)	655(10.3)	10.4(9.6–11.3)		81(1.27)	1.2(1.0–1.5)	
High	5800(47.7)	1080(18.6)	18.4(17.0–19.8)		197(3.4)	3.3(2.8–3.9)	
Diabetes							
No	10936(90.0)						
Yes	1215(10.0)						
Hypertension							
No	8821(72.6)						
Yes	3330(27.4)						

(*Continued*)

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 1. (*Continued*)

	Distribution	DBNCDs	Prevalence of DBNCDs		TBNCDs	Prevalence of TBNCDs	
Variables	*Total, n* (%)	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value	Yes, *n* (%)	Yes, % (95% CI)	*p*-Value
Overweight and obesity							
No	9194(75.7)						
Yes	2957(24.3)						

Bold refers to significant results.

nor were involved in any source of smoking (84.8%). Moreover, nearly two-thirds of respondents resided in rural areas (64.2%), while most adults (13.8%) came from the Dhaka division. Furthermore, the study found that 10%, 27.4%, and 24.3% of respondents had diabetes, hypertension, and overweight or obesity, respectively.

Herein, Table 1 revealed that the prevalence of DBNCDs and TBNCDs was 14.3% (95% CI: 13.5%–15.2%) and 2.3% (95% CI: 2.0%–2.6%), respectively. The prevalence of DBNCDs and TBNCDs were higher among respondents aged 45–54 years (21.9% and 4.4%), female (16.0% and 2.4%), formerly/ever married (17.4% and 2.8%), not working (16.8% and 2.7%), richest families (27.7% and 6.1%), short-heighted (15.8% and 2.2%), and those who drank caffeine (21.9% and 4.7%). Higher educated respondents (16.5%) and family size of less than four (14.6%) had the maximum prevalence of DBNCDs, whereas the TBNCDs were superior among the secondary-educated respondents (2.7%) and those with a family size of greater than four (2.2%). Consequently, DBNCDs and TBNCDs were both greatly prevalent among urban respondents (18.9% and 3.4%), adults with low community poverty (18.3% and 3.3%) and high community literacy (18.4% and 3.3%). DBNCDs were higher among the respondents in the Chittagong division (18.5%), whereas TBNCDs were the most prevalent among the respondents in the Dhaka division (3.3%).

Associated risk factors of DBNCDs and TBNCDs

This study focuses on the two-step analysis to find out the associated factors of DBNCDs and TBNCDs. First, the findings from the Chi-square analysis found that all factors were significantly associated with either DBNCDs or TBNCDs, except smoking status (Table 1). Then, a multilevel LR model was used to select the potential risk factors for both DBNCDs and TBNCDs. Results of multilevel LR analysis for both DBNCDs and TBCNDs were presented in Table 2. Age, sex, marital status, wealth index, education and major administrative region significantly predicted DBNCDs and TBNCDs ($p < 0.05$). Moreover, family size and community literacy were associated with only DBNCDs, whereas caffeinated drinks and community poverty were linked only with TBNCDs ($p < 0.05$). These significant factors were entered into ML-based algorithms for predicting both DBNCDs and TBNCDs.

Performance evaluation of six ML-based techniques

Figure 3 demonstrates the comparison of the accuracy of six different classifiers for three protocols (K2, K5, and K10). The findings were presented with the value of average classification accuracy. The results showed that with an increase in the number of protocols from K2 to K5 to K10, the classification accuracy of most of the classifiers also increased. In addition, the study found that the RF-based classifier performed better comparing the others considered classifier for the three protocols to predict both DBNCDs and TBNCDs. The RF-based classifier provided the highest

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 2. Two-level logistic regression analysis of study factors associated with DBNCDs and TBNCDs

	DBNCDs (Model 4)		TBNCDs (Model 4)	
	Two-level logistic regression analysis		Two-level logistic regression analysis	
Variables	*p*-Value	Decision	*p*-Value	Decision
Initial-level factors				
Age group	**<0.001**	Approved	**<0.001**	Approved
Sex	**<0.001**	Approved	**0.008**	Approved
Marital status	**<0.001**	Approved	**0.006**	Approved
Family size	**0.009**	Approved	0.826	Rejected
Wealth index	**<0.001**	Approved	**<0.001**	Approved
Education level	**<0.001**	Approved	**0.004**	Approved
Employment status	0.273	Rejected	0.893	Rejected
Height	0.216	Rejected	0.802	Rejected
Caffeinate drinks	0.701	Rejected	**0.011**	Approved
Cluster-level factors				
Place of residence	0.695	Rejected	0.926	Rejected
Division	**0.005**	Approved	**0.011**	Approved
Community poverty	0.349	Rejected	**0.002**	Approved
Community literacy	**0.025**	Approved	0.065	Rejected

Bold refers to significant results.

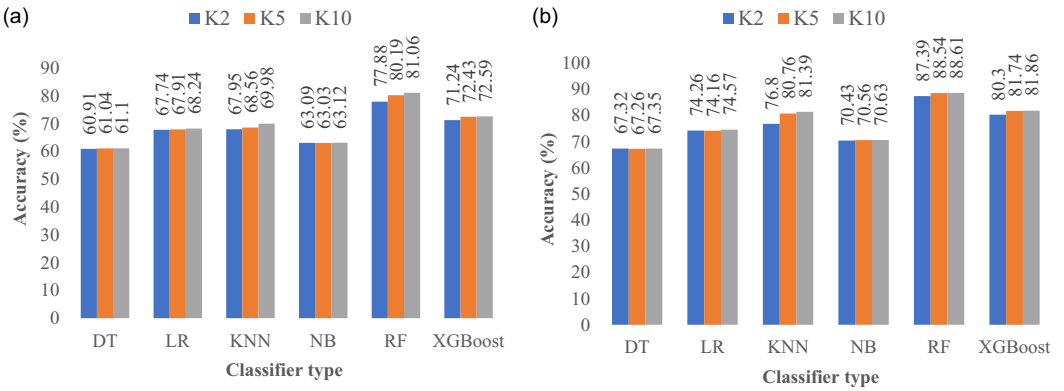

Figure 3. Comparison of accuracy of six classifiers for DBNCDs and TBNCDs over K2, K5, and K10 protocols.

classification accuracy as compared to other classification models for both DBNCDs and TBNCDs. Moreover, the RF-based classifier archived a classification accuracy of 77.88%, 80.19% and 81.06% for K2, K5 and K10 for DBNCDs, whereas the classification accuracy of 87.39%, 88.54% and 88.61% for K2, K5 and K10 was obtained by the RF-based classifier for TBNCDs. The correspondence results of both DBNCDs and TBNCDs and their violin plots of accuracy were also presented in Table 3 and Fig. 4, respectively. As shown in Fig. 4, the RF-based classifier delivered the highest accuracy, followed by XGBoost, whereas DT offered the lowest accuracy for both DBNCDs and TBNCDs.

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 3. Comparison of ACC (in %) and AUC of six classifiers for DBNCDs and TBNCDs over K2, K5, and K10 protocols

	DBNCDs						TBNCDs					
	K2		K5		K10		K2		K5		K10	
Classifier type	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
DT	60.91	0.69	61.04	0.64	61.10	0.64	67.32	0.76	67.26	0.77	67.35	0.76
LR	67.74	0.74	67.91	0.73	68.24	0.74	74.26	0.81	74.16	0.82	74.57	0.82
KNN	67.95	0.78	68.56	0.80	69.98	0.81	76.80	0.94	80.76	0.95	81.39	0.95
NB	63.09	0.67	63.03	0.67	63.12	0.69	70.43	0.76	70.56	0.77	70.63	0.76
RF	**77.88**	**0.91**	**80.19**	**0.91**	**81.06**	**0.93**	**87.39**	**0.96**	**88.54**	**0.96**	**88.61**	**0.97**
XGBoost	71.24	0.84	72.43	0.85	72.59	0.86	80.30	0.94	81.74	**0.96**	81.86	0.95

Bold refers to significant results.

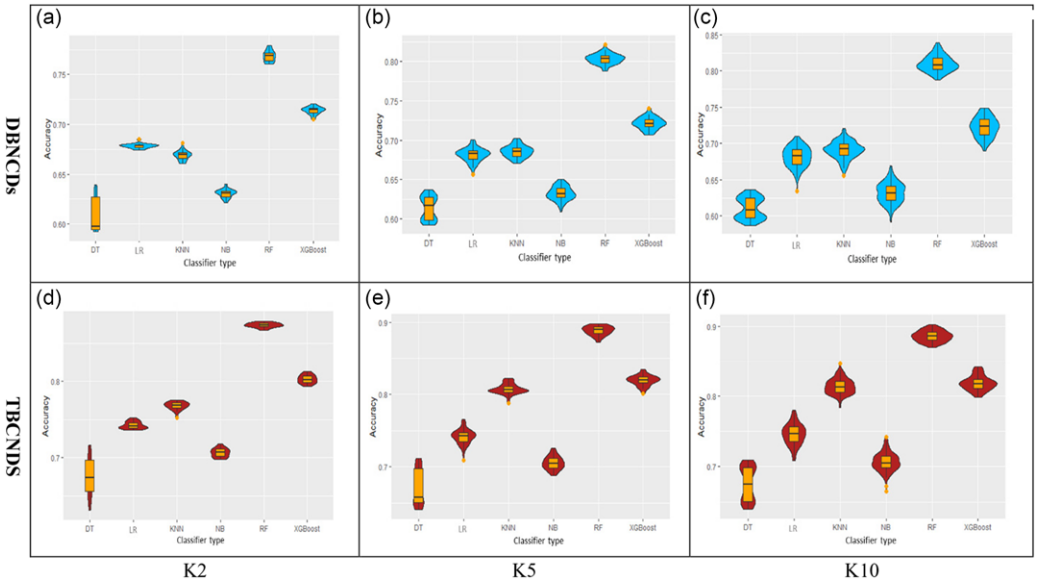

Figure 4. Violin plot of the three partitions (K2, K5, and K10) CV for DBNCDs and TBNCDs.

Table 4 represents the other five performance assessment parameters, denoted as SE, SP, PPV, NPV, and FM, for six different classifiers of three different partition systems for DBNCDs and TBNCDs. In the case of the K10 partition protocol, the RF-based classifier gives the highest SE (82.38% and 92.12%), SP (79.76% and 85.16%), PPV (80.00% and 85.92%), NPV (82.16% and 91.67%), and FM (81.18% and 88.92%). On the contrary, DT had the lowest SE (43.02% and 49.98%), NPV (58.47% and 63.20%), and FM (52.30% & 60.28%) for both DBNCDs and TBNCDs. Furthermore, K2 and K5 partition protocols also provided almost similar results for both DBNCDs and TBNCDs.

Receiver operation characteristics (ROC) evaluation

Figure 5 represents the ROC curves of six classifiers of three different protocols (K2, K5, and K10) for both DBNCDs and TBNCDs. The RF-based classifier performed a better AUC of each

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 4. Five other performance evaluation parameters (in %) for six classifiers of DBNCDs and TBNCDs over K2, K5, and K10 protocols

		Performance Evaluation parameters									
		DBNCDs					TBNCDs				
Protocol Type	Classifier type	SE	SP	PPV	NPV	FM	SE	SP	PPV	NPV	FM
K2	DT	45.36	76.19	65.19	58.65	53.50	51.65	83.14	75.07	63.63	61.20
	LR	66.57	68.88	67.77	67.70	67.17	72.49	76.00	74.81	73.75	73.63
	KNN	67.80	66.11	66.29	67.62	67.04	88.57	65.23	71.46	85.31	79.10
	NB	51.77	74.61	66.72	61.15	58.30	69.40	71.87	70.80	70.49	70.09
	RF	**77.44**	**76.33**	**76.28**	**77.48**	**76.85**	**91.62**	**83.23**	**84.30**	**90.99**	**87.81**
	XGBoost	69.13	73.32	71.81	70.73	70.44	80.44	80.16	79.94	80.65	80.19
K5	DT	45.58	76.24	65.35	58.77	53.70	49.19	85.02	76.35	63.00	59.84
	LR	66.51	69.28	68.04	67.79	67.26	72.46	75.84	74.67	73.69	73.55
	KNN	71.35	65.82	67.24	70.04	69.23	91.51	66.26	73.56	91.76	83.11
	NB	50.83	75.03	66.68	60.82	57.69	68.94	72.14	70.87	70.26	69.89
	RF	**81.75**	**78.65**	**79.01**	**81.43**	**80.36**	**92.50**	**85.25**	**86.04**	**92.04**	**89.15**
	XGBoost	70.30	74.54	73.07	71.85	71.66	82.74	81.14	81.18	82.71	81.95
K10	DT	43.02	78.88	66.69	58.47	52.30	49.98	84.42	75.93	63.20	60.28
	LR	67.05	69.41	68.30	68.18	67.67	72.30	76.81	75.40	73.83	73.82
	KNN	73.17	64.87	67.18	71.09	70.05	91.18	66.84	74.04	91.68	83.67
	NB	51.85	74.20	66.40	61.06	58.23	69.66	71.59	70.68	70.59	70.16
	RF	**82.38**	**79.76**	**80.00**	**82.16**	**81.18**	**92.12**	**85.16**	**85.92**	**91.67**	**88.92**
	XGBoost	70.46	74.29	72.93	71.90	71.67	82.70	81.04	81.09	82.65	81.88

Bold refers to significant results.

protocol for both DBNCDs and TBNCDs as compared to other classification models and their correspondence AUC values were illustrated in Table 3. As shown in Table 3, for K10 partition protocol, the RF-based classifier gave the highest AUC of 0.93, followed by XGBoost (0.86), KNN (0.81), LR (0.74), NB (0.69), and DT (0.64) for DBNCDs. In addition, the RF-based classifier gave the highest AUC of 0.97 for TBNCDS. However, results of AUC for K2 and K5 partition protocols, and RF classifier provide the highest AUC.

Validation of methods

This study also utilised the BDHS 2011 dataset to validate the suggested procedure. The dataset consisted of a total of 5,223 participants, where the prevalence of DBNCDs and TBNCDs was 10.0% and 1.7%, respectively. Table 5 depicts the validation accuracy of the suggested method for DBCNDs and TBNCDs for all three protocols (K2, K5, and K10). In the case of K10, the RF-based classifier contributed to a better accuracy of 73.36% for DBNCDs and 83.80% for TBNCDs compared to DT, LR, KNN, NB, and XGBoost. Thus, the study may claim that the suggested procedure is a better classifier for DBNCDs and TBNCDs. It implies that the RF classifier is good and reliable for predicting both DBNCDs and TBNCDs.

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Table 5. Validation of our proposed method using BDHS-2011 data over K2, K5, and K10 protocols

Classifier type	AUC of DBNCDs			AUC of TBNCDs		
	K2	K5	K10	K2	K5	K10
DT	68.29	68.31	68.34	77.24	77.18	77.24
LR	55.58	55.58	55.63	58.76	58.80	58.80
KNN	52.46	52.38	52.54	70.09	70.13	70.15
NB	51.14	51.50	51.50	73.08	73.12	73.12
RF	**73.09**	**73.31**	**73.36**	**83.65**	**83.78**	**83.80**
XGBoost	65.06	64.96	65.06	75.99	76.74	76.34

Bold refers to significant results.

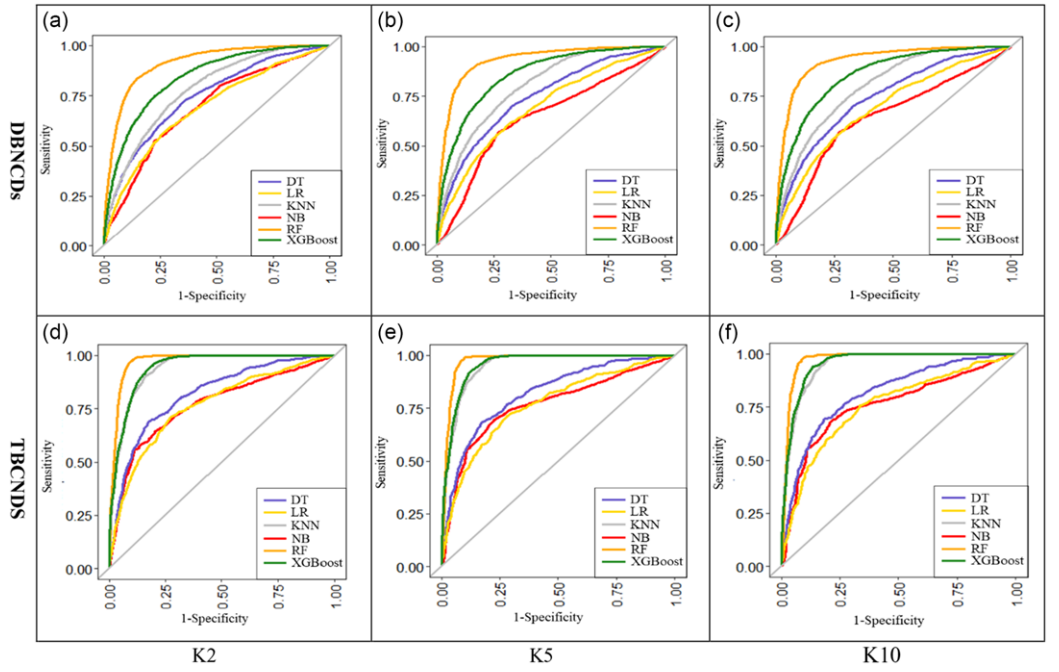

Figure 5. ROC curves of six classifiers for DBNCDs and TBNCDs over K2, K5, and K10 protocols.

Discussion

The increasing prevalence of NCDs and the associated burden have become a major public health concern for society and national governments across all countries including Bangladesh. Extant literature focused on the identification of risk factors and prediction of individual NCDs such as diabetes, hypertension, as well as overweight and obesity. Thus, this study primarily aimed to find out the most significant factors of DBNCDs and TBNCDs and their prediction using an ML-based technique. The two most popular methods, Chi-square test and two-level LR model, were used to find out the most significant factors of DBNCDs and TBNCDs. Furthermore, the study included six ML-based classifiers for prediction.

The prevalence of diabetics, hypertension as well as overweight and obesity was 10.0%, 27.4% and 24.3%, respectively. Moreover, DBNCDs and TBNCDs were prevalent among 14.3% and 2.3%

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

of adults, which was greater than a preceding study conducted in Bangladesh, where except for diabetes (12%), the prevalence of hypertension and overweight and obesity were 27% and 22%, and the outcome of related types such as DBNCDs and TBNCDs accounted for 14% and 1.3%, respectively (Biswas *et al.*, 2019).

Six ML-based classifiers delivered an accuracy ranging from 61.10% to 81.06% for DBNCDs and an AUC that ranged from 0.64 to 0.93, whereas for TBNCDs, the classification accuracy ranged from 67.26% to 88.61% and AUC ranged from 0.76 to 0.97 in the case of K10 protocol. RF-based classifiers achieved a remarkably higher accuracy of 81.06% and 88.61% and an AUC of 0.93 and 0.97 for DBNCDs and TBNCDs, respectively.

Limited research has been conducted to predict individual NCDs such as diabetes, hypertension, and overweight in Bangladesh (Guo *et al.*, 2002; Maniruzzaman *et al.*, 2019; Islam *et al.*, 2022). However, the prediction of DBNCDs and TBNCDs using 2017–2018 BDHS data has not yet been employed. For instance, a study was conducted in Bangladesh to detect and classify diabetes using the BDHS 2011, and among six ML-based classifiers, bagged classification and regression tree (Bagged CART) classifiers provide the highest ACC and AUC of 94.3% and 0.60, respectively (Islam *et al.*, 2020). In another study conducted in the Kurmitola General Hospital, Bangladesh, DT, KNN, RF, and NB classifiers were used to categorise diabetes (Pranto *et al.*, 2020). In that study, RF and NB classifiers performed well on diabetes datasets. A study adopted classifiers such as NB, DT, Adaboost, and RF classifiers to predict diabetic patients using the 2009–2012 National Health and Nutrition Examination Survey dataset in the USA. The combination of LR-based feature selection and RF-based classifier gives the highest accuracy of 94.25% and AUC of 0.95 for the K10 protocol (Maniruzzaman *et al.*, 2019). Moreover, the RF-based classifier can better predict diabetes with an AUC value of 0.999 (Cheng *et al.*, 2020).

Previously, some studies were conducted in Bangladesh as well as elsewhere to predict hypertension using ML techniques. BDHS 2017–2018 data were considered to predict hypertension using artificial neural network (ANN), DT, RF, and gradient boosting (GB) techniques. The performance of the GB technique gives the maximum accuracy of 66.98% and AUC of 0.669 as compared to others (Islam *et al.*, 2021). Another study conducted in a private university of Vitoria da Conquista, Bahia, Brazil, predicted that increased blood pressure was related to hypertension and found that classification tree analysis performed best (Golino *et al.*, 2014). Another study also found that neural network models perform better in predicting hypertension (AUC = 0.766) in several rural villages of Xinxiang County, Henan province in Central China (Zhang *et al.*, 2020). In Hyderabad and India, researchers sought to develop an ML-based algorithm for the risk stratification of NCDs diseases like diabetes and hypertension. The study considered five ML-based models, namely DT, KNN, Adaboost, RF, and LR, as well as the results indicated that the highest performance scores were outperformed for both diseases by the RF-based model (Boutilier *et al.*, 2021). In addition, another study was carried out to predict hypertension in three South Asian countries of Bangladesh, Nepal, and India using GB, RF, DT, and ANN ML-based techniques. The GB provided the highest accuracy (66.98%), FM (78.99%), and AUC (0.669) as compared to other methods (Islam *et al.*, 2021).

In modern times, being overweight and obese has become a significant threat worldwide; hence, their early prediction is very important. A study in Bangladesh used KNN, RF, LR, multilayer perceptron, support vector machine (SVM), NB, adaptive boosting, DT, and GB classifier to predict obesity and found that the LR algorithm achieves the highest accuracy of 97.09% as compared to the other classifiers (Ferdowsy *et al.*, 2021). Meanwhile, another study on obesity was conducted in the United Kingdom using its millennium cohort data and found that the multilayer perceptron algorithm resulted in a minority class accuracy of 54% for the imbalanced dataset but jumped over 90% in the case of balanced data (Singh and Tawfik, 2020). Furthermore, another study predicts obesity using publicly available genetic profiles. Some most popular ML techniques, including GB, generalised linear model, CARTs, KNN, SVMs, RF, and multilayer perceptron neural network, are used to predict obesity and found that SVM generated

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

the highest AUC value of 90.5% (Montañez *et al.*, 2017). A study on overweight or obesity conducted in China found that GB machine (ACC = 0.9454) performed best when compared among the considered ML-based techniques including LR, DT, SVM, RF, KNN, gradient boosting machine (GBM), XGBoost, light gradient boosting machine (LGBM), and NB (Wang *et al.*, 2022).

Another study was also conducted in Bangladesh based on NCDs and found that the gradient boosting decision tree (GBDT)-based model yielded the greatest AUC of 0.91 with an accuracy of 67.5% (Hu *et al.*, 2018). Meanwhile, another study focused on smoking-induced NCDs prediction and used the National Health and Nutrition Examination Survey datasets from South Korea (KNHANES) and the United States (NHANES). The study included the following three feature selection techniques and six classifiers: LR, RF, KNN, MLP, NN, and XGBoost. Under hybrid feature selection, XGBoost provided the highest accuracy of 88.12% with an AUC value of 0.84 (Davagdorj *et al.*, 2020). Further study concentrated on predicting and diagnosing NCDs by adopting six classifiers (ANN, SVN (RBF), DT, LSTM, NB, and RF). The study used a total of 26 attributes and found that DT provided the highest accuracy of 99% (Fatou *et al.*, 2020).

Strengths and limitations of this study

In addition to validated indicators as well as biomarker analyses of the wealth index, the principal strengths of the present investigation were the utilisation of a demographically representative survey that gathered information on blood glucose, blood pressure, body height, and weight measurements by qualified professionals based on established standards. Second, from the latest available information, this study utilises a two-level LR approach on DBNCDs and TBNCDs. Finally, the use of the three-partition CV protocol with selected classifiers provided an accurate performance measurement of NCDs. However, despite the various positive aspects of the study, there are also some drawbacks. First, the causal route of this study could not be constructed because it was a cross-sectional study; thus, it simply provides the association between explanatory and outcome variables. Second, the information on fruit and vegetable consumption was not accessible considering that is one of the nine voluntary agreed-upon global objectives that the WHO has announced. Third, BMI was the only procedure that was used to ascertain the dietary status of individuals following the WHO standards. However, this procedure is not as accurate as some of the others that are available, such as DEXA methods, waist–hip ratio as well as bioelectrical impedance, which are used to measure the status of being overweight and obese.

Conclusion

This study offered the latest and most detailed knowledge related to NCDs in Bangladesh and concludes that age group, sex, marital status, wealth index, education level, and division were significantly associated with both DBNCDs and TBNCDs. Moreover, family size and community literacy were substantially related to DBNCDs, whereas a notable connection was observed between TBNCDs with caffeinate intake and community poverty. Furthermore, the use of an RF-based classifier on all three CV protocols (K2, K5, and K10) provided the best performance for both DBNCDs and TBNCDs by considering the selected risk factors. Thus, a population-based approach utilises the healthcare sector and draws attention to the trend of these illnesses within demographics to detect and treat diseases at an early stage, as well as lower the possibility of getting DBNCDs or TBNCDs. In addition, non-health strategies such as multisectoral partnership, information and knowledge management, as well as innovations need to be set as priorities to address determinants of DBNCDs and TBNCDs.

Acknowledgements. The authors of this study would like to thank the DHS programme, ICF International for providing us with the dataset for analysis. The authors also would like to acknowledge the contribution of Statistics Discipline, Science, Engineering and Technology School, Khulna University, Khulna-9208, Bangladesh.

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Funding statement. This research received no specific grant from any funding agency, commercial entity, or not-for-profit organisation.

Competing interests. The authors have no conflicts of interest to declare.

Ethical standard. The BDHS data collected from secondary sources and obtained the required ethical approvals from National Ethics Committee of the Bangladesh Medical Research Council. We registered and requested access to data from the DHS website. DHS programmes collect data following written informed consent from each individual. The survey was conducted in accordance with relevant guidelines and regulations.

References

Ahammed B, Maniruzzaman M, Talukder A and Ferdausi F (2021). Prevalence and risk factors of hypertension among young adults in Albania. *High Blood Pressure & Cardiovascular Prevention* **28**, 35–48.

Ahammed B, Sarder MA, Kundu S, Keramat SA and Alam K (2022). Multilevel exploration of individual-and community-level factors contributing to overweight and obesity among reproductive-aged women: a pooled analysis of Bangladesh Demographic and Health Survey, 2004–2018. *Public Health Nutrition* **25**(8), 2074–2083.

Al Kibria GM, Hashan MR, Hossain MM, Zaman SB and Stennett CA (2021). Clustering of hypertension, diabetes and overweight/obesity according to socioeconomic status among Bangladeshi adults. *Journal of Biosocial Science* **53**(2), 157–166.

Al-Zubayer MA, Ahammed B, Sarder MA, Kundu S, Majumder UK and Islam SM (2021). Double and triple burden of non-communicable diseases and its determinants among adults in Bangladesh: evidence from a recent demographic and health survey. *International Journal of Clinical Practice* **75**(10), e14613.

Bangladesh Bureau of Statistics. Population & Housing Census (2022). URL: http://www.bbs.gov.bd/site/page/47856ad0-7e1c-4aab-bd78-892733bc06eb/Population-&-Housing.2022.

Bentéjac C, Csörgő A and Martínez-Muñoz G (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* **54**, 1937–1967.

Bigna JJ and Noubiap JJ (2019). The rising burden of non-communicable diseases in sub-Saharan Africa. *The Lancet Global Health* **7**(10), e1295–e1296.

Bista B, Dhungana RR, Chalise B and Pandey AR (2020). Prevalence and determinants of non-communicable diseases risk factors among reproductive aged women of Nepal: results from Nepal Demographic Health Survey 2016. *PloS One* **15**(3), e0218840.

Biswas T, Townsend N, Islam MS, Islam MR, Gupta RD, Das SK and Al Mamun A (2019). Association between socioeconomic status and prevalence of non-communicable diseases risk factors and comorbidities in Bangladesh: findings from a nationwide cross-sectional survey. *BMJ Open* **9**(3), e025538.

Bloom DE, Cafiero ET, Jané-Llopis E, Abrahams-Gessel S, Bloom LR, Fathima S, Feigl AB, Gaziano T, Mowafi M, Pandya A, Prettner K, Rosenberg L, Seligman B, Stein AZ and Weinstein C (2011). *The Global Economic Burden of Noncommunicable Diseases*. Geneva: World Economic Forum.

Boutilier JJ, Chan TC, Ranjan M and Deo S (2021). Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis. *Journal of Medical Internet Research* **23**(1), e20123.

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

Bunkhumpornpat C, Sinapiromsaran K and Lursinsap C (2011). MUTE: Majority under-sampling technique. In *2011 8th International Conference on Information, Communications & Signal Processing*, IEEE, Singapore, pp. 1–4.

Cheng D, Ting C, Ho C and Ho C (2020). Performance evaluation of explainable machine learning on non-communicable diseases. *Solid State Technology* **63**, 2780–2793.

Davagdorj K, Pham VH, Theera-Umpon N and Ryu KH (2020). XGBoost-based framework for smoking-induced noncommunicable disease prediction. *International Journal of Environmental Research and Public Health* **17**(18), 6513.

Fatou NG, Ibrahima FA, Camara MS and Alassane BA (2020). A study on predicting and diagnosing non-communicable diseases: case of cardiovascular diseases. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, IEEE, Morocco, pp. 1–8.

Ferdowsy F, Rahi KS, Jabiullah MI and Habib MT (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences* **2**, 100053.

Fottrell E, Ahmed N, Shaha SK, Jennings H, Kuddus A, Morrison J, Akter K, Nahar B, Nahar T, Haghparast-Bidgoli H and Khan AA (2018). Distribution of diabetes, hypertension and non-communicable disease risk factors among adults in rural Bangladesh: a cross-sectional survey. *BMJ Global Health* **3**(6), e000787.

Golino HF, Amaral LS, Duarte SF, Gomes CM, Soares TD, Reis LA and Santos J (2014). Predicting increased blood pressure using machine learning. Journal of Obesity 23, 2014.

Guo SS, Wu W, Chumlea WC and Roche AF (2002). Predicting overweight and obesity in adulthood from body mass index values in childhood and adolescence. *The American Journal of Clinical Nutrition* **76**(3), 653–658.

https://doi.org/10.1017/S0021932024000063 Published online by Cambridge University Press

Hastie T, Tibshirani R, Friedman JH and Friedman JH (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hossain SM and Chetty G (2011). Next generation identity verification based on face-gait Biometrics. *In Proceedings of the International Conference on Biomedical Engineering and Technology* **11**, 142–148.

Hu M, Nohara Y, Wakata Y, Ahmed A, Nakashima N and Nakamura M (2018). Machine learning based prediction of non-communicable diseases to improving intervention program in Bangladesh. *European Journal of Biomedical Informatics* **14**(2), 20–28.

Islam MM, Rahman MJ, Roy DC, Tawabunnahar M, Jahan R, Ahmed NF and Maniruzzaman M (2021). Machine learning algorithm for characterizing risks of hypertension, at an early stage in Bangladesh. Diabetes & Metabolic Syndrome: Clinical Research & Reviews **15**(3), 877–884.

Islam SM, Purnat TD, Phuong NT, Mwingira U, Schacht K and Fröschl G (2014). Non-Communicable Diseases (NCDs) in developing countries: a symposium report. Globalization and Health **10**(1), 1–8.

Islam SM, Talukder A, Awal MA, Siddiqui MM, Ahamad MM, Ahammed B, Rawal LB, Alizadehsani R, Abawajy J, Laranjo L and Chow CK (2022). Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian countries. Frontiers in Cardiovascular Medicine 2022, 9.

James G, Witten D, Hastie T and Tibshirani R (2013). *An Introduction to Statistical Learning*. New York: Springer, Vol. 112, p. 18.

Khalequzzaman M, Chiang C, Choudhury SR, Yatsuya H, Al-Mamun MA, Al-Shoaibi AA, Hirakawa Y, Hoque BA, Islam SS, Matsuyama A and Iso H (2017). Prevalence of non-communicable disease risk factors among poor shantytown residents in Dhaka, Bangladesh: a community-based cross-sectional survey. *BMJ Open* **7**(11), e014710.

Liao Z, Ju Y and Zou Q (2016). Prediction of G protein-coupled receptors with SVM-prot features and random forest. *Scientifica* **2016**, 1–10.

Libbrecht MW and Noble WS (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16**(6), 321–332.

Liu H and Motoda H (2012). *Feature Selection for Knowledge Discovery and Data Mining*. New York: Springer Science & Business Media.

Ma D, Sakai H, Wakabayashi C, Kwon JS, Lee Y, Liu S, Wan Q, Sasao K, Ito K, Nishihara K and Wang P (2017). The prevalence and risk factor control associated with noncommunicable diseases in China, Japan, and Korea. *Journal of Epidemiology* **27**(12), 568–573.

Maniruzzaman M, Rahman MJ, Ahammed B and Abedin MM (2020). Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems* **8**, 1–4.

Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM, Suri HS, Biswas M, El-Baz A, Bangeas P, Tsoulfas G and Suri JS (2019). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. Computer Methods and Programs in Biomedicine 176, 173–193.

Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, Suri JS (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of Medical Systems 42, 1–7.

Maniruzzaman M, Shin J and Hasan MA (2022). Predicting children with ADHD using behavioral activity: a machine learning analysis. *Applied Sciences* **12**(5), 2737.

Matsuoka D (2021). Classification of imbalanced cloud image data using deep neural networks: performance improvement through a data science competition. *Progress in Earth and Planetary Science* **8**(1), 1.

Merlo J, Yang M, Chaix B, Lynch J and Råstam L (2005). A brief conceptual tutorial on multilevel analysis in social epidemiology: investigating contextual phenomena in different groups of people. *Journal of Epidemiology & Community Health* **59**(9), 729–736.

Montañez CA, Fergus P, Hussain A, Al-Jumeily D, Abdulaimma B, Hind J and Radi N (2017). Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Anchorage, AK, USA, pp. 2743–2750.

National Institute of Population Research and Training (NIPORT), and ICF (2020). Bangladesh Demographic and Health Survey 2017–18. Dhaka, Bangladesh, and Rockville, Maryland, USA: NIPORT and ICF.

Pranto B, Mehnaz SM, Mahid EB, Sadman IM, Rahman A and Momen S (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information* **11**(8), 374.

Quinlan JR (1986). Induction of decision trees. Machine Learning 1, 81–106.

Rabe-Hesketh S and Skrondal A (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**(4), 805–827.

Riaz BK, Islam MZ, Islam AS, Zaman MM, Hossain MA, Rahman MM, Khanam F, Amin KB and Noor IN (2020). Risk factors for non-communicable diseases in Bangladesh: findings of the population-based cross-sectional national survey 2018. *BMJ Open* **10**(11), e041334.

Russell S, Sturua L, Li C, Morgan J, Topuridze M, Blanton C, Hagan L and Salyer SJ (2019). The burden of non-communicable diseases and their related risk factors in the country of Georgia, 2015. *BMC Public Health* **19**, 1–9.

Saeed KM (2013). Prevalence of risk factors for non-communicable diseases in the adult population of urban areas in Kabul City, Afghanistan. *Central Asian Journal of Global Health* **2**(2), 1–20.

Shah S, Luo X, Kanakasabai S, Tuason R and Klopper G (2019). Neural networks for mining the associations between diseases and symptoms in clinical notes. *Health Information Science and Systems* **7**, 1–9.

Singh B and Tawfik H (2020). Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Springer International Publishing, Proceedings, Part IV 20, pp. 523–535.

Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M, Abbasi-Kangevari M, Abbastabar H, Abd-Allah F, Abdelalim A and Abdollahi M (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **396**(10258), 1204–1222.

Wang Q, Yang M, Pang B, Xue M, Zhang Y, Zhang Z and Niu W (2022). Predicting risk of overweight or obesity in Chinese preschool-aged children using artificial intelligence techniques. *Endocrine* **77**(1), 63–72.

World Health Organization (2013). *Global Action Plan for the Prevention and Control of NCDs 2013–2020*. Geneva: WHO.

World Health Organization (2016). *Global Report on Diabetes*. Geneva: World Health Organization (WHO).

World Health Organization (2020). Noncommunicable Diseases. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicablediseases (accessed 1st April 2020).

World Health Organization (2022). Noncommunicable Diseases. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (accessed 16th September 2022).

Yosef T (2020). Prevalence and associated factors of chronic non-communicable diseases among cross-country truck drivers in Ethiopia. *BMC Public Health* **20**(1), 1–7.

Zaman MM, Bhuiyan MR, Karim M, Rahman M, Akanda AW and Fernando T (2015). Clustering of non-communicable diseases risk factors in Bangladeshi adults: an analysis of STEPS survey 2013. *BMC Public Health* **15**(1), 1–9.

Zhang L, Yuan M, An Z, Zhao X, Wu H, Li H, Wang Y, Sun B, Li H, Ding S and Zeng X (2020). Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: a cross-sectional study of chronic diseases in central China. *PloS One* **15**(5), e0233166.

Cite this article: Al-Zubayer MA, Alam K, Shanto HH, Maniruzzaman M, Majumder UK, and Ahammed B (2024). Machine learning models for prediction of double and triple burdens of non-communicable diseases in Bangladesh. *Journal of Biosocial Science* **56**, 426–444. https://doi.org/10.1017/S0021932024000063