HONG KONG MEDICAL JOURNAL 香港醫學雜誌

Supplementary material

Supplement to: JT Tan, R Zhang, KH Chan, et al. Machine learning model for prediction of coronavirus disease 2019 within 6 months after three doses of BNT162b2 in Hong Kong. Hong Kong Med J 2025;Epub 23 Jun 2025. https://doi.org/10.12809/hkmj2411879.

Appendix. Recruitment checklist

Covid19_FU_Vaccine Study

Participant's initials: _____ Sex/Age: _____

Please check eligibility:

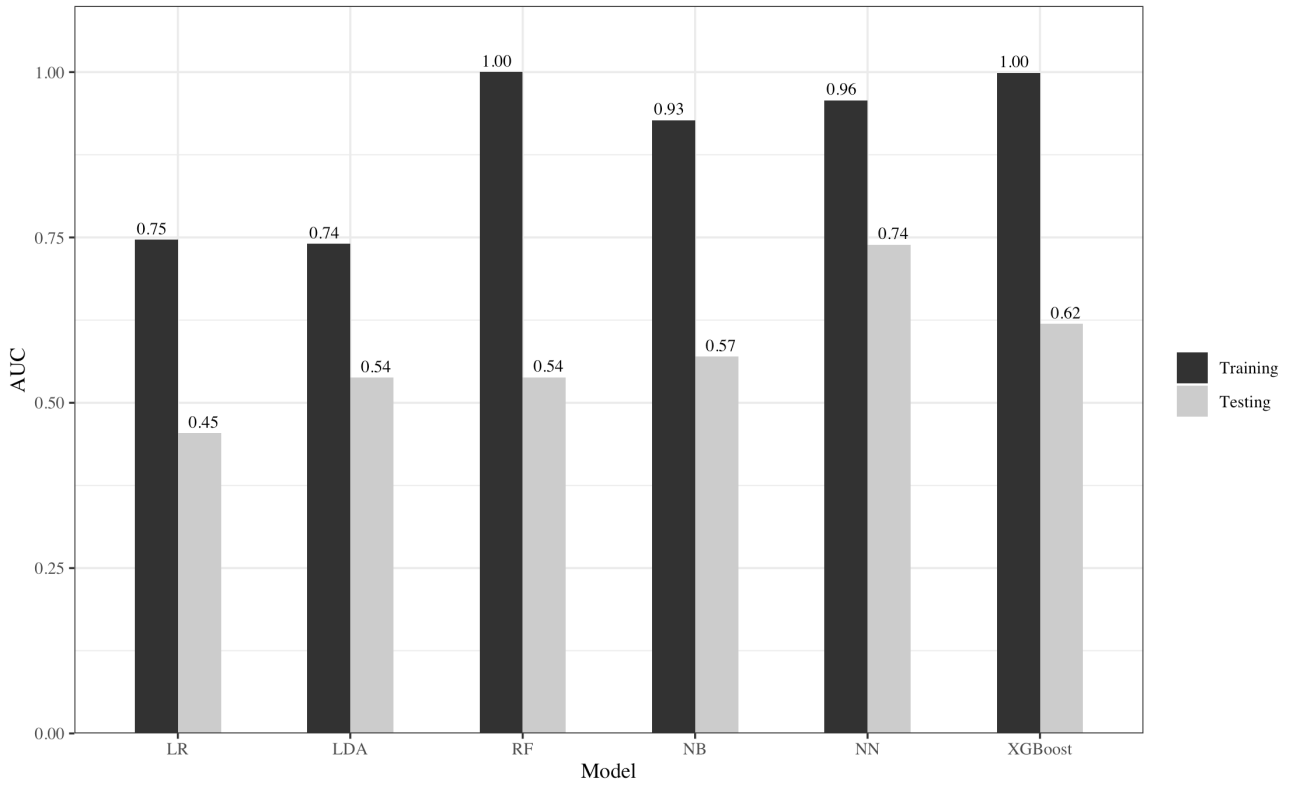
Inclusion criteria	Meet	Not meet
1. Recruited individuals must be adults aged ≥18 years.		
2. All participants must provide written informed consent.		
3. Participants must be available to complete the study and comply with study procedures. Willingness to allow serum samples to be stored beyond the study period for potential additional future testing to better characterise immune response is required.		

Exclusion criteria	Meet	Not meet
1. Inability to comprehend and follow all required study procedures.		
2. Recent (documented, confirmed, or suspected) flu-like illness within 1 week of vaccination.		
3. Known allergy to polyethylene glycol (PEG) or other components of the study vaccines, or a history of anaphylaxis or serious vaccine reactions to any excipients.		
4. History of receiving immunoglobulin or other blood products within 3 months prior to vaccination in this study.		
5. Known active human immunodeficiency virus (HIV) infection.		
6. Receipt of an experimental agent (vaccine, drug, biologic, device, blood product, or medication) within 1 month prior to vaccination in this study, or expectation of receiving an experimental agent during the study. Unwillingness to refrain from participation in another clinical study through the end of this study.		
7. Tympanic temperature ≥38°C within 3 days of the intended study vaccination.		
8. History of alcohol or drug abuse in the past 5 years.		
9. History of Guillain-Barré Syndrome, transverse myelitis, or Bell's palsy.		
10. Female participants planning pregnancy from study enrolment to 28 days after receiving the second dose of vaccine.		
11. Any condition that the investigator believes may interfere with successful completion of the study (e.g., **post–COVID-19 status**).		

Participant eligible for study: ☐ Yes, Study No.: _____ ☐ Exclude from study

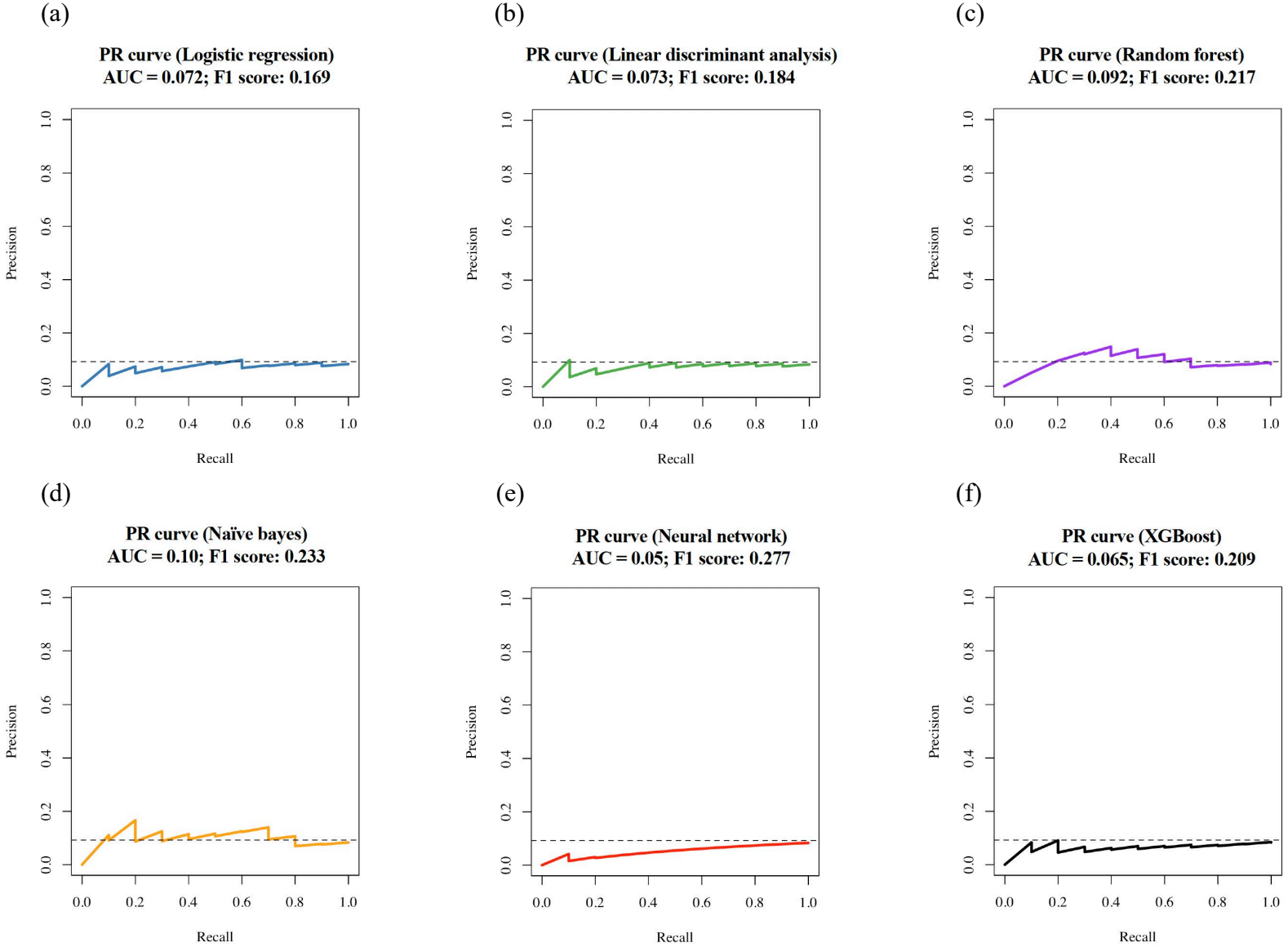
Checked by _____ Date: _____

Supplementary Figure 1. Comparison of areas under the receiver operating characteristic curve between the training and testing sets across the six models



Abbreviations: AUC = area under the receiver operating characteristic curve; LDA = linear discriminant analysis; LR = logistic regression; NB = naïve Bayes; NN = neural network; RF = random forest; XGBoost = extreme gradient boosting

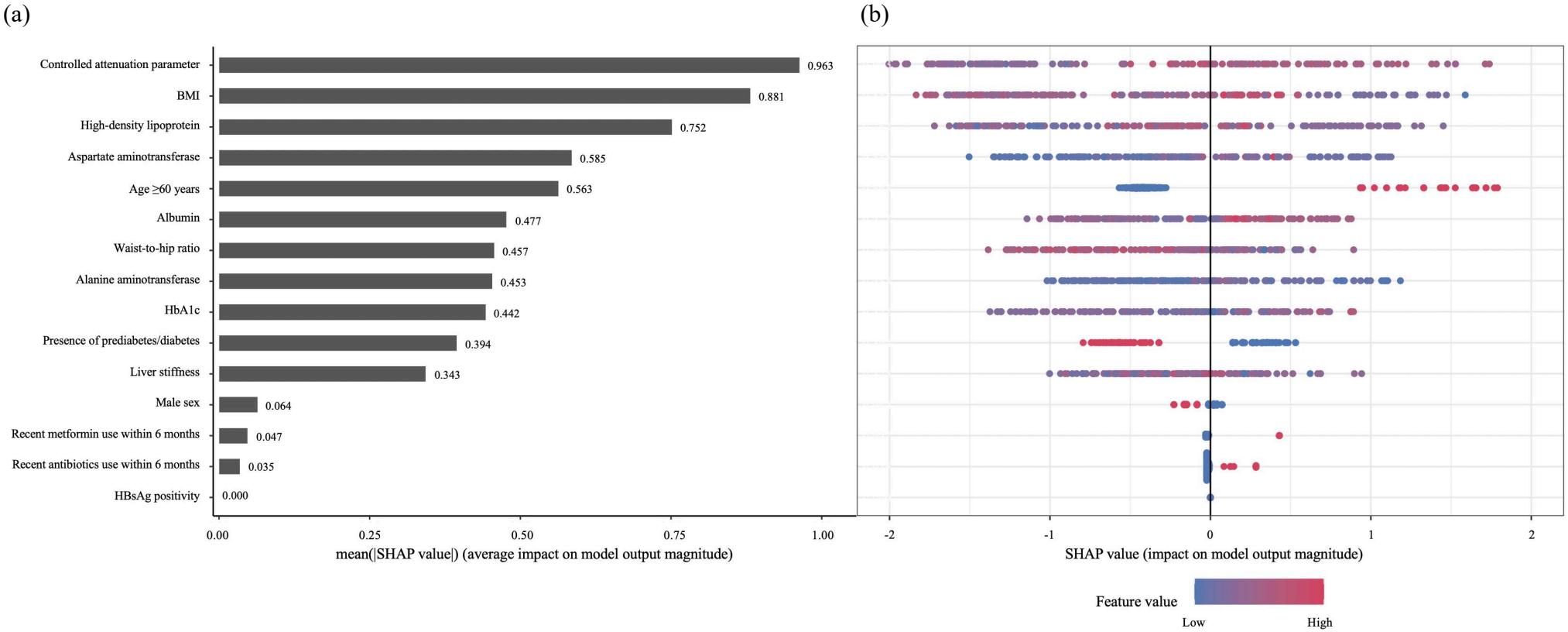
Supplementary Figure 2. Precision-recall curves and F1 scores of different machine learning models using the testing set. (a) Logistic regression model. (b) Linear discriminant analysis model. (c) Random forest model. (d) Naïve Bayes model. (e) Neural network model. (f) Extreme gradient boost model*



Abbreviations: AUC = area under the receiver operating characteristic curve; PR curve = precision-recall curve; XGBoost = extreme gradient boosting

* Dotted horizontal line indicates baseline of the PR curve (0.092), which corresponds to the number of positive cases divided by the total number of training data

Supplementary Figure 3. SHapley Additive exPlanations (SHAP) analysis representing risk factors in predicting the risk of coronavirus disease 2019 using the extreme gradient boosting model. (a) Mean absolute SHAP value* of each variable. (b) Value summary dot plot of the respective variable†

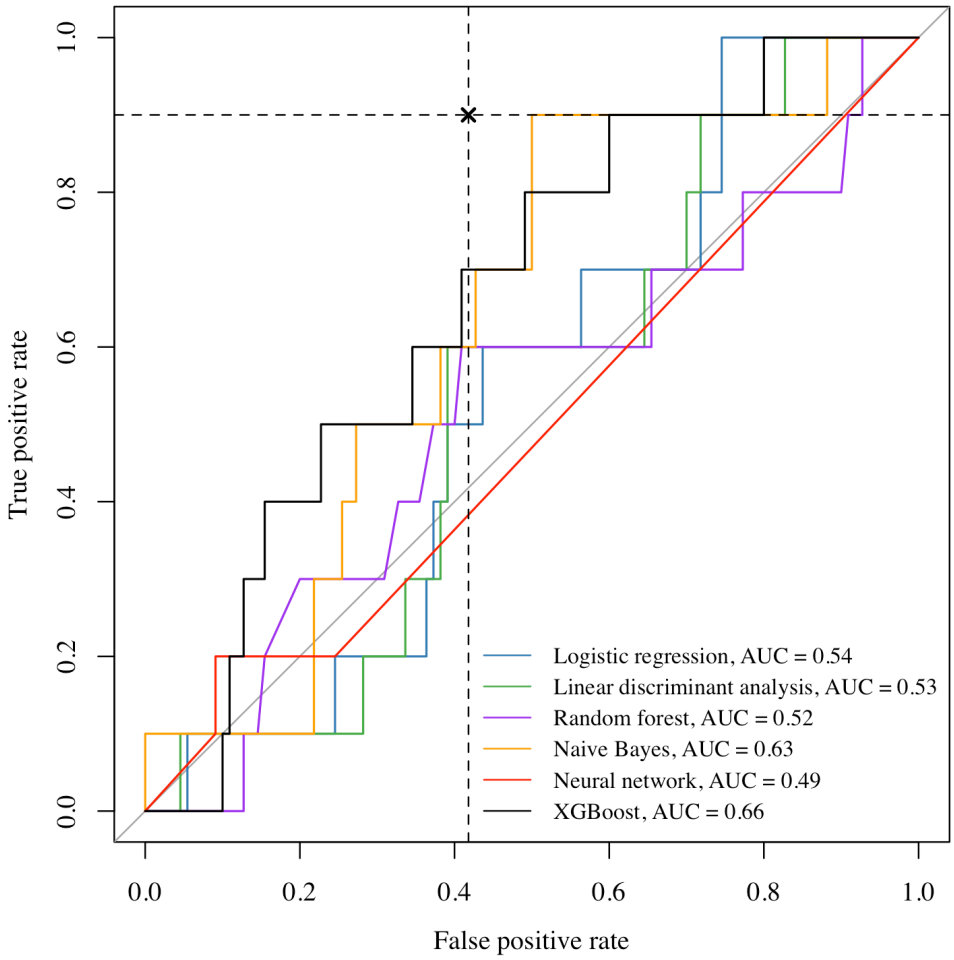


Abbreviations: BMI = body mass index; HbA1c = glycated haemoglobin; HBsAg = hepatitis B surface antigen; SHAP = SHapley Additive exPlanations

* Representing the average impact on model output magnitude

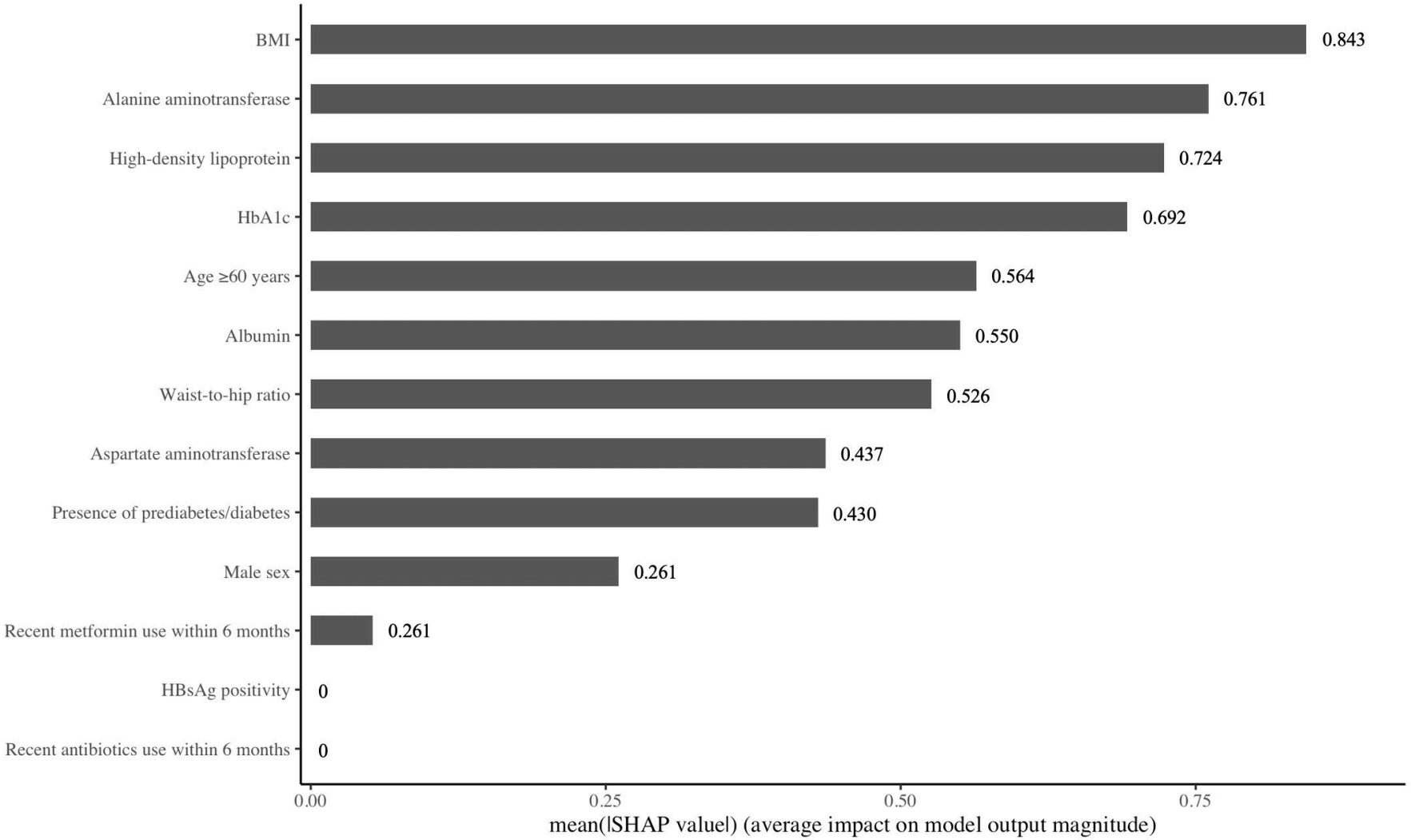
† Each dot represents a feature. Its position on the x-axis corresponds to its SHAP value; positive and negative values indicate positive or negative impact on the model output, respectively. Dot colour represents feature value; blue and red indicate low and high values, respectively. Dot density reflects the distribution or pattern of feature values in the dataset, such that denser regions indicate more common values and sparse regions indicate less common values. Features with high mean absolute SHAP values are important for model interpretation

Supplementary Figure 4. Receiver operating characteristic curves of different machine learning models using the testing set, after exclusion of controlled attenuation parameter and liver stiffness during model training



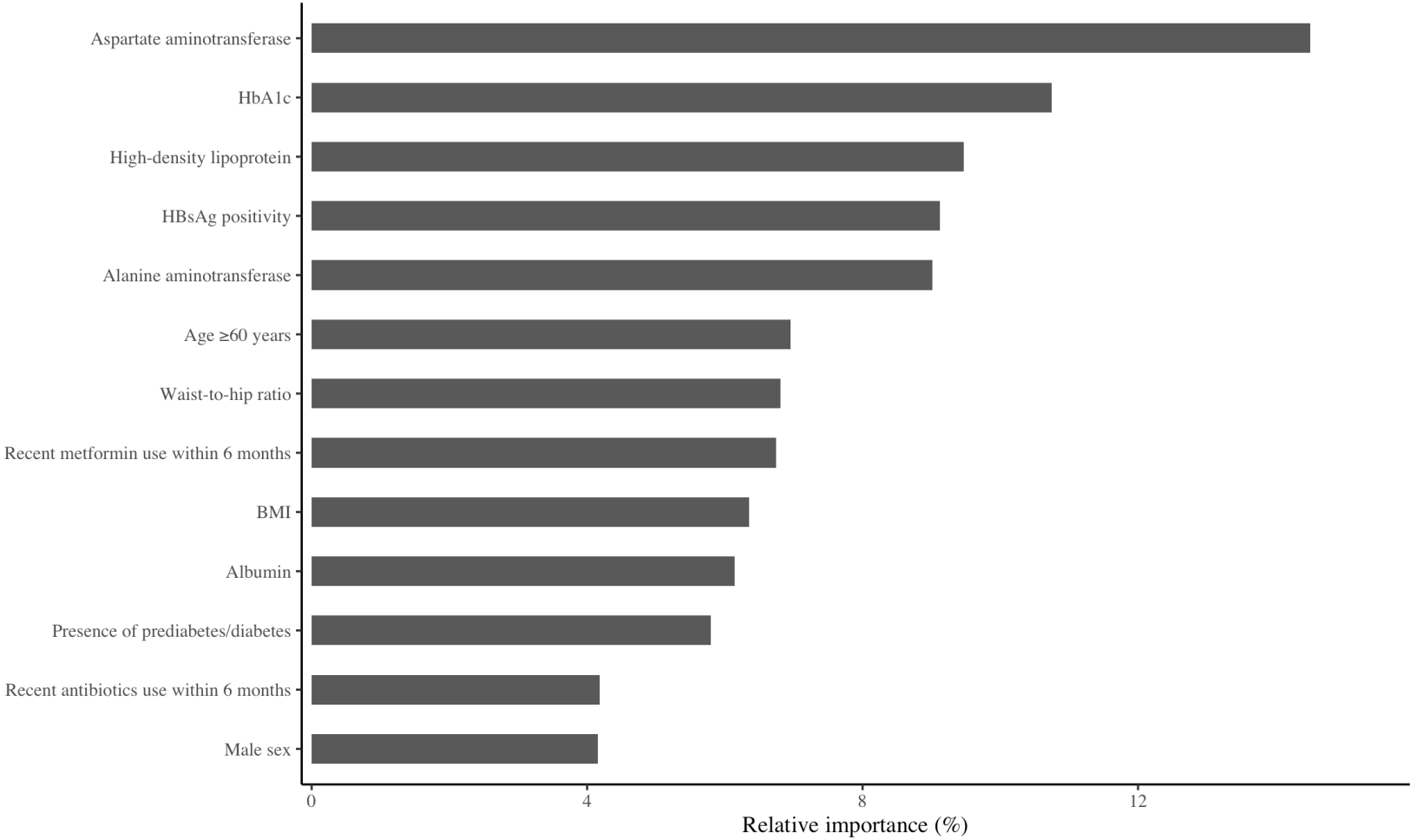
Abbreviations: AUC = area under the receiver operating characteristic curve; XGBoost = extreme gradient boosting

Supplementary Figure 5. Relative importance of risk factors in predicting the risk of coronavirus disease 2019 based on the extreme gradient boosting model, after exclusion of controlled attenuation parameter and liver stiffness during model training



Abbreviations: BMI = body mass index; HbA1c = haemoglobin A1c; HBsAg = hepatitis B surface antigen; SHAP = SHapley Additive exPlanations

Supplementary Figure 6. Relative importance of risk factors in predicting the risk of coronavirus disease 2019 by the neural network model, after exclusion of controlled attenuation parameter and liver stiffness during model training



Abbreviations: BMI = body mass index; HbA1c = haemoglobin A1c; HBsAg = hepatitis B surface antigen

Supplementary Table 1. Summary of variables used for model development and testing

Variable category	Selected variables
Demographics	Age, sex, BMI, waist-to-hip ratio, smoking, alcohol history
Medical data/co-morbidities	CAP, liver stiffness, hypertension, prediabetes/diabetes, GI surgery
Blood tests	Haemoglobin, white blood cells, platelet count, neutrophil count, lymphocyte count, creatinine, eGFR using CKD-EPI, albumin, globulin, bilirubin, alkaline phosphatase, alanine aminotransferase, gamma-glutamyl transferase, fasting glucose, glycated haemoglobin, triglycerides, total cholesterol, high-density lipoprotein, low-density lipoprotein, hepatitis B surface antigen test
Recent 6-month medication use	Proton pump inhibitors, antibiotics, probiotics and prebiotics, statins, metformin, antidepressants

Abbreviations: BMI = body mass index; CAP = controlled attenuation parameter; eGFR using CKD-EPI = estimated glomerular filtration rate using the creatinine equation from the Chronic Kidney Disease Epidemiology Collaboration; GI surgery = gastrointestinal surgery

Supplementary Table 2. Summary descriptions of machine learning algorithms

	Characteristics	Important hyperparameters
LR	Models the probability of a binary outcome Simple and interpretable but may not capture complex feature interactions	None
LDA	Finds the best linear combination of features for each class, assuming normality and equal covariance matrices	None
RF	Ensemble of decision trees Robust to noise and can capture non-linear interactions	'mtry'
NB	Assumes independence between features Simple and fast but may not capture complex feature interactions	'fL', 'usekernel', 'adjust'
NN	Non-linear model that learns complex feature interactions	'size', 'decay'
XGBoost	Ensemble of decision trees with gradient boosting Effective in handling imbalanced data and feature interactions	'eta', 'max_depth', 'gamma', 'colsample_bytree', 'min_child_weight', 'subsample', 'nrounds'

Abbreviations: LDA = linear discriminant analysis; LR = logistic regression; NB = naïve Bayes; NN = neural network; RF = random forest; XGBoost = extreme gradient boosting

Supplementary Table 3. Baseline characteristics of participants based on train-test dataset before multiple imputation (n=304)[*]

	Training set (n=184)	Testing set (n=120)	P value[†]
Demographics			
Age ≥60 y	28 (15.2%)	23 (19.2%)	0.368
Male sex	62 (33.7%)	32 (26.7%)	0.195
BMI, kg/m^2	23.3 (20.8-25.5)	22.6 (20.8-24.7)	0.267
Unknown	4 (2.2%)	2 (1.7%)	
Waist-to-hip ratio	0.9 (0.8-0.9)	0.9 (0.8-0.9)	0.216
Unknown	5 (2.7%)	6 (5.0%)	
Smoking			0.769
Non-smoker	163 (88.6%)	102 (85.0%)	
Current smoker	10 (5.4%)	7 (5.8%)	
Ex-smoker	9 (4.9%)	8 (6.7%)	
Unknown	2 (1.1%)	3 (2.5%)	
Alcohol use			0.597
Non-drinker	162 (88.0%)	108 (90.0%)	
Current drinker	18 (9.8%)	8 (6.7%)	
Ex-drinker	1 (0.5%)	1 (0.8%)	
Unknown	3 (1.6%)	3 (2.5%)	
Medical data/co-morbidities			
CAP, dB/m	233.5 (203.8-270.2)	224.5 (200.2-254.2)	0.047
Unknown	4 (2.2%)	4 (3.3%)	
Liver stiffness, kPa	4.3 (3.6-5.2)	4.3 (3.6-5.4)	0.988
Unknown	4 (2.2%)	4 (3.3%)	
Hypertension	29 (15.8%)	20 (16.7%)	0.834
Pre-diabetes/diabetes	75 (40.8%)	46 (38.3%)	0.673
GI surgery	5 (2.7%)	5 (4.2%)	0.523
Blood tests			
Haemoglobin, g/dL	13.7 (12.8-14.5)	13.4 (12.7-14.3)	0.209
White blood cells, ×10^9/L	5.6 (4.8-6.8)	5.8 (4.8-6.9)	0.732
Platelets, ×10^9/L	241.5 (216.8-287.2)	259.0 (229.0-299.2)	0.111
Neutrophils, absolute, ×10^9/L	3.1 (2.4-4.0)	3.2 (2.5-4.1)	0.592
Lymphocytes, absolute, ×10^9/L	1.8 (1.5-2.1)	1.9 (1.5-2.1)	0.457
Creatinine, µmol/L	66.0 (59.0-80.0)	67.0 (58.0-78.0)	0.880
eGFR using CKD-EPI, unit	90.0 (88.0-90.0)	90.0 (81.0-90.0)	0.051
Unknown	1 (0.5%)	0	
Albumin, g/L	46.0 (44.0-47.0)	46.0 (44.0-47.0)	0.275
Globulin, g/L	30.0 (28.0-32.0)	30.0 (28.0-32.0)	0.712
Bilirubin, µmol/L	10.0 (7.0-12.0)	9.0 (7.0-12.0)	0.218
Alkaline phosphatase, total, U/L	63.5 (54.0-79.0)	66.5 (54.0-78.0)	0.850
Alanine aminotransferase, U/L	19.5 (15.0-28.0)	19.0 (15.0-24.0)	0.143

Aspartate aminotransferase, U/L	23.0 (19.0-26.0)	22.0 (19.0-26.0)	0.390
Unknown	1 (0.5%)	1 (0.8%)	
Gamma-glutamyl transferase, U/L	21.0 (16.0-34.5)	20.0 (16.0-28.0)	0.310
Unknown	1 (0.5%)	1 (0.8%)	
Fasting glucose, mmol/L	5.1 (4.7-5.5)	5.1 (4.7-5.4)	0.483
HbA1c, %	5.5 (5.3-5.8)	5.6 (5.3-5.7)	0.783
Triglycerides, mmol/L	0.9 (0.7-1.3)	0.9 (0.7-1.2)	0.510
Total cholesterol, mmol/L	5.0 (4.4-5.6)	4.8 (4.1-5.4)	0.299
Cholesterol, HDL, mmol/L	1.6 (1.4-1.9)	1.7 (1.4-1.9)	0.129
Cholesterol, LDL, mmol/L	2.8 (2.3-3.2)	2.7 (2.2-3.1)	0.106
HBsAg-positive	7 (3.8%)	13 (10.8%)	0.016
Medications‡			
Proton pump inhibitor	20 (10.9%)	19 (15.8%)	0.206
Antibiotics	13 (7.1%)	16 (13.3%)	0.069
Probiotics and prebiotics	4 (2.2%)	4 (3.3%)	0.717
Statin	23 (12.5%)	17 (14.2%)	0.674
Metformin	9 (4.9%)	7 (5.8%)	0.719
Antidepressant	8 (4.3%)	3 (2.5%)	0.536

Abbreviations: BMI = body mass index; CAP = controlled attenuation parameter; eGFR using CKD-EPI = estimated glomerular filtration rate using the creatinine equation from the Chronic Kidney Disease Epidemiology Collaboration; GI surgery = gastrointestinal surgery; HbA1c = glycated haemoglobin; HBsAg = hepatitis B virus antigen; HDL = high-density lipoprotein; LDL = low-density lipoprotein

* Data are shown as No. (%) or median (interquartile range)

† Fisher's exact test, Pearson's Chi squared test and Wilcoxon rank-sum test

‡ Recent drug usage within 6 months prior to vaccination

Supplementary Table 4. Confusion matrix for different machine learning models predicting coronavirus disease 2019

LR	
6 True positive	55 False positive
4 False negative	55 True negative

LDA	
8 True positive	69 False positive
2 False negative	41 True negative

RF	
5 True positive	31 False positive
5 False negative	79 True negative

NB	
7 True positive	43 False positive
3 False negative	67 True negative

NN	
9 True positive	46 False positive
1 False negative	64 True negative

XGBoost	
7 True positive	50 False positive
3 False negative	60 True negative

Abbreviations: LDA = linear discriminant analysis; LR = logistic regression; NB = naïve Bayes; NN = neural network; RF = random forest; XGBoost = extreme gradient boosting

Supplementary Table 5. P values for multiple comparisons between the neural network model and other machine learning algorithms in terms of performance metrics

	LR	LDA	RF	NB	XGBoost
AUC*					
NN	0.01	0.08	0.06	0.15	0.32
Sensitivity†					
NN	0.08	0.32	0.05	0.16	0.16
Specificity†					
NN	0.003	<0.001	<0.001	0.08	0.05
PPV‡					
NN	0.02	<0.001	0.55	0.32	0.07
NPV‡					
NN	0.07	0.20	0.06	0.17	0.15
PLR§					
NN	0.02	<0.001	0.54	0.31	0.08
NLR§					
NN	0.08	0.11	0.12	0.18	0.15

Abbreviations: AUC = area under the receiver operating characteristic curve; LDA = linear discriminant analysis; LR = logistic regression; NB = naïve Bayes; NLR = negative likelihood ratio; NN = neural network; NPV = negative predictive value; PLR = positive likelihood ratio; PPV = positive predictive value; RF = random forest; XGBoost = extreme gradient boosting

* DeLong's test was used to compare AUCs of two models

† McNemar's test was used to compare sensitivity and specificity of two models

‡ Comparison of differences in PPV and NPV between two models was based on the method proposed by Moskowitz and Pepe (2006) [see reference 1]

§ A regression model approach proposed by Gu and Pepe (2009) was used to test for differences in PLR and NLR (see reference 2)

References

1. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials 2006;3:272-9.
2. Gu W, Pepe MS. Estimating the capacity for improvement in risk prediction with a marker. Biostatistics 2009;10:172-86.