Supplementary Information for "Cultural-aware Machine Learning based Analysis of COVID-19 Vaccine Hesitancy"

Raed Alharbi[1,*], My T. Thai [1,*], Sylvia Chan-Olmsted[2,+], and Huan Chen[3,+]

[1]Computer and Information Science and Engineering Department, University of Florida, Gainesville, FL, 32611, USA.
[2]Department of Media Production, Management, and Technology, University of Florida, Gainesville, FL, 32611, USA.
[3]Department of Advertising, University of Florida, Gainesville, FL, 32611, USA.
[*]{r.alharbi, mythai}@ufl.edu
[+]{chanolmsted, huanchen}@jou.ufl.edu

Supplementary Table 1: List of final hyperparameters.

(A)

Hyperparameter	Value
Number of trees (n_estimators)	184
Maximum depth (max_depth)	12
Number of features considered when looking for best split (max_features)	Auto
The minimum number of samples required to split an internal node	4
The minimum number of samples required to be at a leaf node	4
Bootstrap sampling (bootstrap)	False

(B)

Hyperparameter	Value
Number of trees (n_estimators)	121
Maximum depth (max_depth)	5
Learning rate (learning_rate)	0.02
The minimum number of samples required to split an internal node	5
The minimum number of samples required to be at a leaf node	2
Bootstrap sampling (bootstrap)	True

(C)

Hyperparameter	Value
Number of neighbors (n_neighbors)	5
Algorithm used to compute the nearest neighbors	kd_tree
leaf size	20

(D)

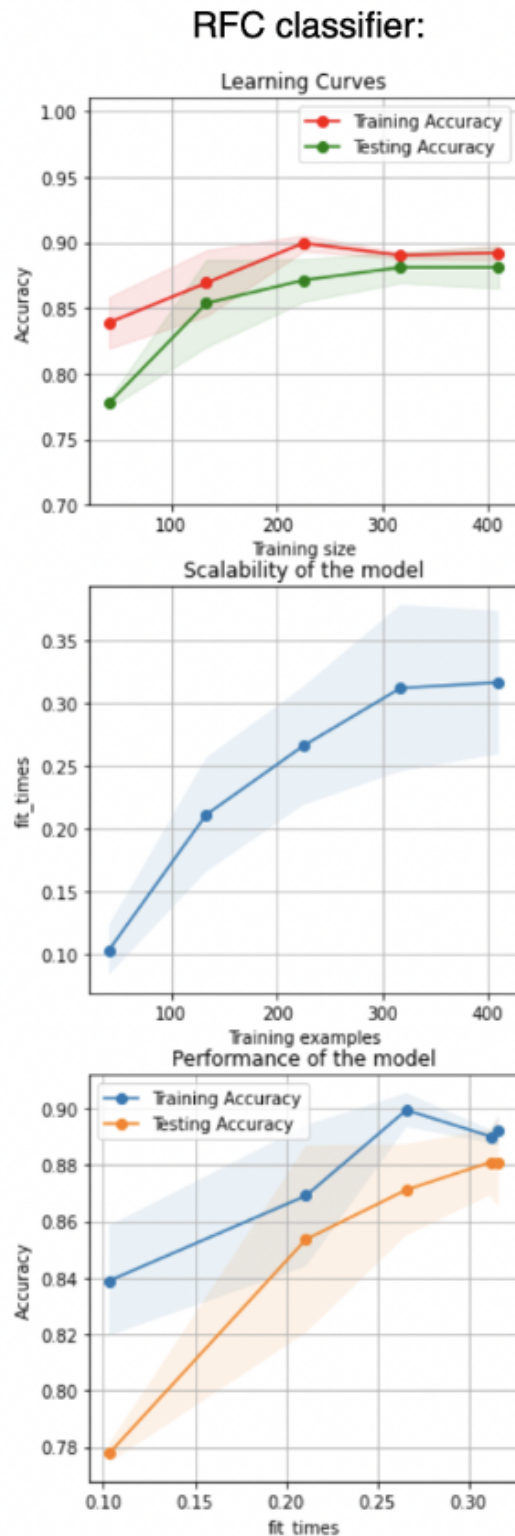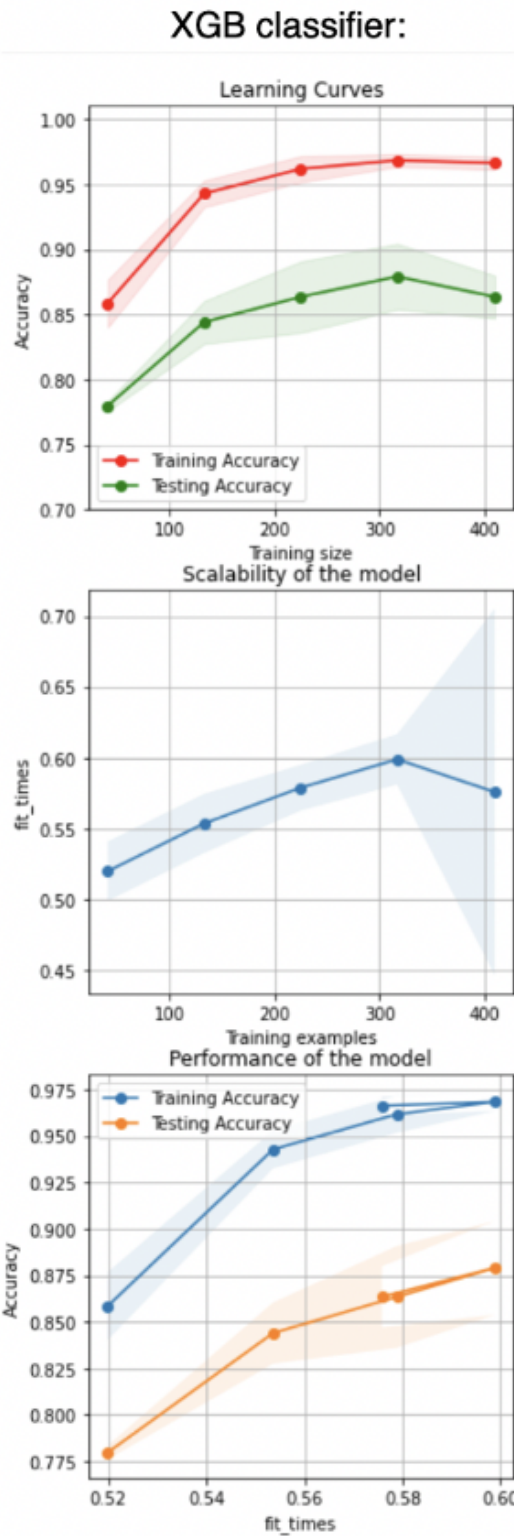| Hyperparameter	Value
kernel type	Linear
Kernel coefficient	Scale

(E)

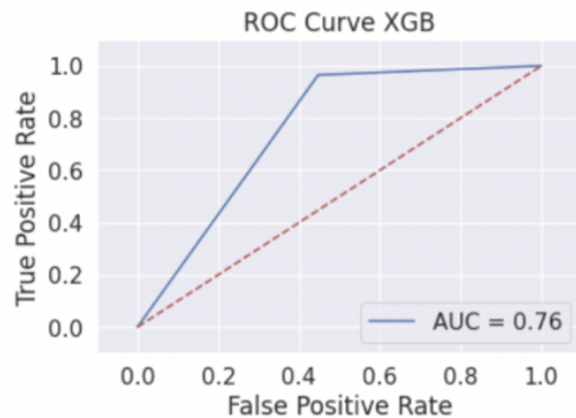| Hyperparameter	Value
Width of the decision prediction layer (n_d)	9
Width of the attention embedding for each mask (n_a)	10
Coefficient for feature reusage in the masks (gamma)	1.2

The hyperparameters for the top two models in terms of the accuracy performance. (A) The random forest classifier and (B) The extreme gradient Boosting model. (C) k-nearest neighbors. (D) Support vector machine. (E) Attentive Interpretable Tabular Learning.

Supplementary Grid matrix 1: correlation among features.

We construct a correlation matrix, a filtering strategy, to assess the relevance of the features to the target variable and then exclude the attributes that are poorly correlated with the target.
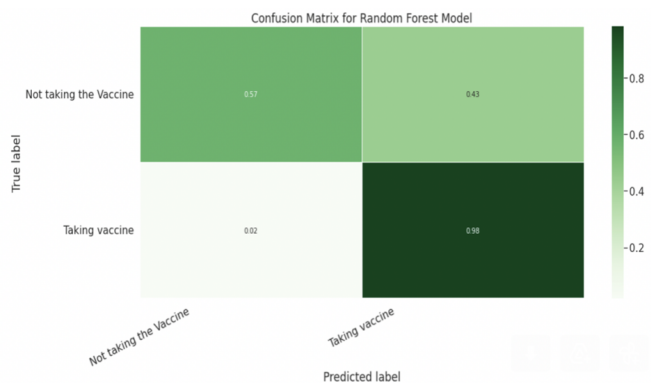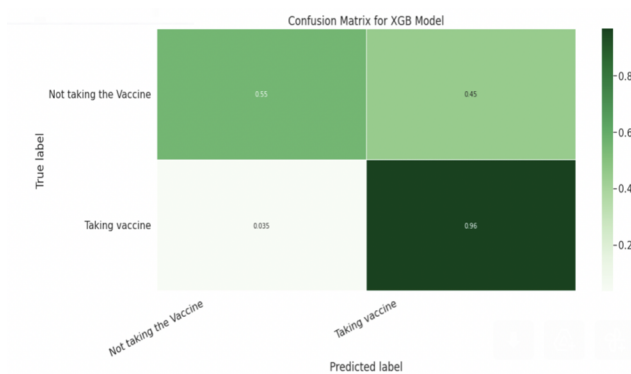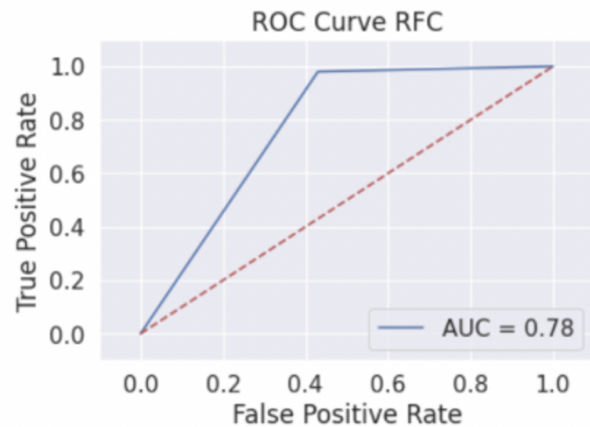
Supplementary Figure 1: Models Correctness.

XGB classifier:

ROC Curve XGB

AUC = 0.76

RFC classifier:

ROC Curve RFC

AUC = 0.78

Confusion Matrix for XGB Model

	Not taking the Vaccine	Taking vaccine
Not taking the Vaccine	0.55	0.45
Taking vaccine	0.035	0.96

Confusion Matrix for Random Forest Model

	Not taking the Vaccine	Taking vaccine
Not taking the Vaccine	0.57	0.43
Taking vaccine	0.02	0.98

The first row indicates that both XGB and RFC are in developmental areas, indicating that the score is still relatively high and that the validation score might be enhanced with further training samples. The charts in the second row illustrate how long the models require to train with varying amounts of training data. Next, the third-row plots illustrate the amount of time needed to train the models for various training sizes. Finally, the fourth row shows the receiver operating characteristic curve for both the XGB and RF models, where the RF AUC score (0.78) outperforms the XGB AUC score (0.76).

Supplementary Figure 2: XGB explanations .
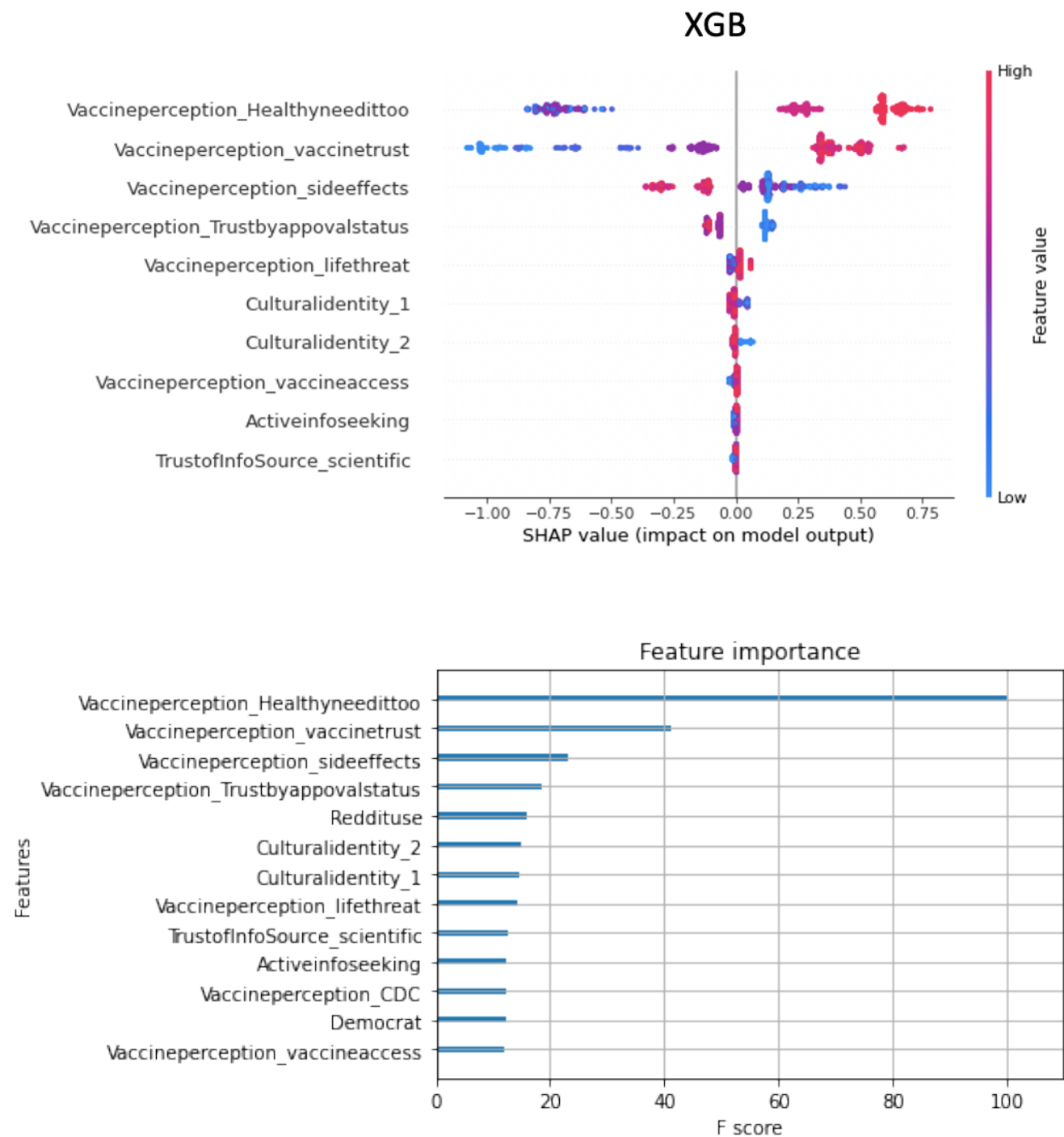
The interpretation of the XGB model using SHAP and the Scikit-learn technique for determining the feature significance. The first plot depicts the most important factors influencing a person's decision, with features listed according to their contribution to the final prediction. The second plot depicts the feature's average gain over all splits, with a high F score indicating the most significant values. We can observe overlap among the important features produced by the two methods.

Supplementary Table 2: Features Description in details.

Constructs/Variables	Scale	Constructs/Variables	Scale
Cultural factors		**Non-English language health/medical info source**	
Identity		Use of non-English health/medical information source	Do you use any health/medical information sources in a language that you are familiar with other than English? (Dichotomous Yes/No variable with value 1 and 2) 1- No 2- Yes
Social identity[1]	A cultural group is a collection of individuals who share a core set of beliefs, patterns of behavior, and values (e.g., Hispanic and Latino Americans, African Americans, Asian Americans, etc.). (1-5 scale from strongly disagree to strongly agree on six scale items below to be averaged. To what degree do you agree with following statements about your cultural group? 1- When someone criticizes this group, it feels like a personal insult. 2- I'm very interested in what others think about this group. 3- The limitations associated with this group apply to me also. 4- When I talk about this group, I usually say "we" rather than "they." 5- I have a number of qualities typical of members of this group. 6- When someone praises this group, it feels like a personal compliment.	Importance of non-English health/medical Information Source	If selected "Yes" for information source (5-point scale. See below for value codes) How important are these non-English health/medical sources in informing you? 1- Not at all important 2- Slightly important 3- Moderately important 4- Very important 5- Extremely important
		Accessibility of non-English health/medical information source	If selected "Yes" for information source (5-point scale. See below for value codes) How easy is it for you to access these non-English health/medical information sources? 1- Extremely difficult 2- Somewhat difficult 3- Neither easy nor difficult 4- Somewhat easy 5- Extremely easy
Cultural identity[2]	To what degree do you agree with following statements about your ethnicity? (1-5 scale from strongly disagree to strongly agree on six scale items below to be averaged) 1- I feel positively about my ethnicity 2- If I could choose, I would prefer to be of the same ethnicity 3- I know what my ethnicity means to me 4- I have experienced things that reflect my ethnicity such as eating food, listening to music, and watching movies 5- I have learned about my ethnicity by doing things such as reading (books, magazines, newspapers), searching the internet, or keeping up with current events 6- I have participated in activities that have taught me about my ethnicity	Quality of non-English health/medical information source	If selected "Yes" for information source (5-point scale. See below for value codes) In your opinion, the quality of the information from these non-English health/medical info sources is: 1- Very low 2- Somewhat low 3- Neither low nor high 4- Somewhat high 5- Very high
Social capital[3]		**Primary COVID information sources**	
Social capital	How much do you agree with following statements? (1-5 scale from strongly disagree to strongly agree on four scale times below to be averaged) 1- I have people I trust to help me 2- There is someone I can turn to for advice 3- I have people to turn to for resources if I need them 4- There are people who would support me emotionally	Top COVID information source (Top of Mind)[4]	Imagine that you had a strong need to get information about COVID-19 related topics. Where would you go first? (Choose one. Corresponding value codes below) 1- News websites 2- TV 3- Search engines (e.g., Google) 4- Social media 5- Health care provider 6- Friends/family/acquaintances 7- Scientific/medical websites or journals 8- Print media like newspapers/magazines 9- Radio 10 -Podcasts 11- Online forums 12- Government/health agency websites/posts 13- Religious leaders/organizations
Health-COVID information sources and trust		Trust of the top COVID information source[6]	How much do you agree with the following statements about (selected choice) as a COVID information source? (1-5 scale from strongly disagree to strongly agree on four scale times below to be averaged) 1- I trust it 2- I rely on it 3- It is honest 4- It is safe
Primary health/medical info sources			
Most recent top-of-mind health/ medical information source[4,5]	The most recent time you looked for information about health/medical related information, where did you go first? (Choose one, corresponding value codes below) 1- News websites 2- TV 3- Search engines (e.g., Google) 4- Social media 5- Health care provider 6- Friends/family/acquaintances 7- Scientific/medical websites or journals 8- Print media like newspapers/magazines 9- Radio 10 -Podcasts 11- Online forums 12- Government/health agency websites/posts 13- Religious leaders/organizations	Number of COVID information sources	Please check all the sources that you have used to gather COVID related information. (The total number of items checked would be the value for this variable) 1- News websites 2- TV 3- Search engines (e.g., Google) 4- Social media 5- Health care provider 6- Friends/family/acquaintances 7- Scientific/medical websites or journals 8- Print media like newspapers/magazines 9- Radio 10 -Podcasts 11- Online forums 12- Government/health agency websites/posts 13- Religious leaders/organizations
Trust of most recent top-of-mind primary health/medical Information source[6]	How much do you agree with the following statements about (the selected choice) as a health/medical information source? (1-5 scale from strongly disagree to strongly agree on four scale items below to be averaged) 1- I trust it 2- I rely on it 3- It is honest 4- It is safe		

Constructs/Variables	Scale
COVID info seeking behavior and trust/perceptions (continue)	
Trust of primary COVID information source (volume/experience)[7]	How much do you agree with the following statements about (selected choice) as a COVID-19 information source? (1-5 scale from strongly disagree to strongly agree on four scale times below to be averaged) 1- I trust it 2- I rely on it3- It is honest4- It is safe
Specific COVID information source	Please write down the "specific" source/media outlet that you use the most to gather COVID related information (e.g., Facebook, CNN, CDC website, county health website, my physician, etc.). (Text entry to be recoded and analyzed)
Non-English COVID information source usage (percentage)	Think about the COVID-19 related information you have received thus far, what percentage of it is in English/non-English? (Estimated percentage of English & non-English information source, totaling 100%) 1- English (actual %) 2- Non-English (actual %)
Value of received COVID Information[8]	How much do you agree with the following statements regarding the COVID-19 vaccine information that you have received thus far? (1-5 scale from strongly disagree to strongly agree on the following six scale items to be averaged) 1- It is useful 2- It is valuable 3- It helped me to learn about the vaccine 4- It influenced me to be more involved in self-health care 5- It helped me to learn about vaccination risks/benefits 6- It helped in asking my healthcare givers intelligent questions
Media Usage	
Media usage[9]	On average in any given week, how often do you use the following media/platforms? [5-point scale from None at all/very little (1) to A great deal (5)] 1- Broadcast radio 2- Online radio (streaming) 3- Broadcast/cable TV 4- Online TV (streaming) 5- Print newspaper 6- Online news sites 7- Social media in general 8- Messaging apps
Social media usage[9]	On average in any given week, how often do you use the following social media? [5-point scale from None at all/very little (1) to A great deal (5)] 1- Facebook 2- YouTube 3- Twitter 4- Instagram 5- Pinterest 6- TikTok 7- Reddit 9- Snapchat
COVID Vaccine status and hesitance	
Vaccination status	Have you received the COVID-19 vaccine (either one or two doses)? (Dichotomous variable with 1 indicating no and 2 yes) 1- No 2- Yes
Vaccine eagerness	(If yes to vaccination status) (Corresponding value codes below) When did you receive your COVID-19 vaccine? 1- As soon as it was available to me 2- I waited a little before getting it 3- I waited some before getting it 4- I waited a while just to be sure before getting it 5- I finally just got it recently

Constructs/Variables	Scale
COVID Vaccine status and hesitance (continue)	
Vaccine type	Which COVID-19 vaccine did you receive? (Corresponding value codes below) 1- Johnsons & Johnson / Janssen 2- Moderna 3- Pfizer-BioNtech
Vaccine willingness	(If no to vaccination status) Are you currently willing to receive the COVID-19 vaccine? 1- No 2- Yes 3- Not sure
Positive COVID test	Have you ever been tested positive for COVID-19? 1- No 2- Yes
Infect family	Has any member of your family been infected by COVID-19? 1-No 2-Yes
Family loss	Did you experience death of any family member/relative due to COVID-19? 1- No 2- Yes
Friend loss	Did you experience death of a friend due to COVID-19? 1- No 2- Yes
Infect colleague	Do you have any colleagues who have been infected by COVID-19? 1- No 2- Yes
Colleague loss	Did you experience death of a colleague due to COVID-19? 1- No 2- Yes
(various aspects)	Please indicate the degree to which you agree with following statements. (1-5 scale from strongly disagree to strongly agree. See specific variable names at the end of each statement) - I have easy access to COVID-19 vaccine (Vaccine Access) - I think the COVID-19 virus is a serious threat to my life (Life Threat) - The possible side effects of the COVID-19 vaccine prevent me from receiving the vaccine (Side Effect Concern) - I trust the COVID-19 vaccine (Vaccine Trust) - Many of the COVID-19 related conspiracy theories out there are worth considering (Conspiracy Theory Tendency) - I trust that those who have had the virus are immune to the COVID-19 virus (Immunity Perception) - I believe that the COVID-19 vaccine causes fertility issues (Fertility Hesitance) - I think that COVID-19 vaccine is necessary for healthy people as well (Healthy Vaccine Needs) - I do not trust the COVID-19 vaccine because of its current FDA approval status (Trust of Vaccine Approval) - I trust the FDA (FDA Trust) - I trust the CDC (CDC Trust) - My religious beliefs affect my decision to receive the COVID-19 vaccine (Religious Hesitance) - I believe there may be possible interaction between the vaccine and long-term health issues (Long-term Health Hesitance)
Trust of COVID Info Sources	
Trust of COVID information course (various sources)	How much do you trust the following sources in giving you accurate, reliable COVID related information? [1-5 point scale from Strongly distrust (1) to Strongly trust (5) of seven major sources below] 1- Doctors/nurses/medical professionals 2- Social media 3- News media 4- Friends and family members 5- Government/health agencies and their publications/websites 6- Scientific/medical research institutions and their publications/websites 7- Religious leaders/organizations

Constructs/Variables	Scale
Demographics	
Age	What is your age? (Value codes: 18=1, etc.) Single selection ranging from 18 – 60 over (1-43)
Marital status	What is your marital status? 1- Married 2- Widowed 3- Divorced 4- Separated 5- Never married 6- Prefer not to say
Gender	What is your gender? 1- Male 2- Female 3- Non-binary 4- Prefer not to say
Race	What is your race? 1- White/Caucasian 2- Black or African American 3- Hispanic/Latino 4- Asian 5- American Indian or Alaska Native 6- Native Hawaiian or Pacific Islander 7- Biracial 8- Other (please specify)
Multiracial	No = 1 Yes =2
Language	What is your primary spoken language? 1- English 2- Spanish 3- Portuguese 4- French and French Creole 5- Chinese (including Mandarin, Cantonese, and other varieties) 6- Tagalog/Ilocano 7- Vietnamese 8- Korean 9- Japanese 10- Bengali/Gujarti/Urdu/Hindi/Kannada/ Punjabi/Tamil 11- African languages 12- Other (please specify)
Religion	What do you identify as your religious affiliation? 1- Protestant 2- Catholic 3- Mormon 4- Jewish 5- Orthodox Christian 6- Unitarian/Universalist 7- Muslim 8- Hindu 9- Buddhist 10- Atheist or Agnostic 11- Other religion or unaffiliated
Religiousness	Please indicate the degree to which you consider yourself to be religious (5-point scale with the value codes below) 1- Not religious at all 2- Slightly religious 3- Moderately religious 4- Religious 5- Very religious

Constructs/Variables	Scale
Demographics (continue)	
Work location	(If unemployed is not selected from Employment question) Where is your primary work location? 1- Home 2- Office 3- Other public location 4- Hybrid 5- Other (please specify)
Travel	Have you traveled locally or internationally in the past 6 months? 1- Local only 2- International only 3- Both locally and internationally 4- I did not travel in the past 6 months
Household	Are there any members above the age of 60 residing in your household? 1- No 2- Yes
Politics	Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else? 1- Republican 2- Democrat 3- Independent 4- Something else
Income	Assessed categorically by following measures 1- Less than $10,000 2- $10,000 to $19,999 3- $20,000 to $29,999 4- $30,000 to $39,999 5- $40,000 to 49,999 6- $50,000 to $59,999 7- $60,000 to 69,999 8- $70,000 to $79,999 9- $80,000 to 89,999 10- $90,000 to $99,999 11- $100,000 to $149,999 12- $150,000 or more
Education	What is the highest level of school you have completed or the highest degree you have received? 1- Less than high school degree 2- High school graduate (high school diploma or equivalent including GED) 3- Some college but no degree 4- Associate degree in college (2-year) 5- Bachelor's degree in college (4-year) 6- Master's degree 7- Doctoral degree 8- Professional degree (JD, MD)
Employment	How would you best describe your current employment status? (Corresponding value codes below) 1- Employed 2- Unemployed (including stay-home homemakers) 3- Self-employed 4- Retired 5- Student 6- Disabled 7- Other (please specify)
Additional survey based variables	- Respondent ID - Survey completion time - Time spent completing survey - Respondent IP address - Respondent location (location latitude) - Respondent location (location longitude)

To fully comprehend the underlying analysis and the dataset, we present the above tables, which explain the features characteristics and their associated features value scales.

Supplementary Table 3: Statistical information in detail about the proposed dataset.

Demographic Type		Frequency	Perecent %
Gender	Male	589	28.7
	Female	1435	70.0
	Non-binary	20	1.0
	Prefer not to say	6	.3
Race	Black or African American	716	34.9
	Hispanic/Latino	991	48.3
	Asian	343	16.7
Age	18-24	559	27.3
	25-34	608	30.0
	35-44	363	17.7
	45-54	228	11.1
	55-59	101	4.9
	60 and over	191	9.3
Education	Less than high school degree	80	3.9
	High school diploma or equivalent	599	29.2
	Some college but no degree	511	24.9
	Associate degree in college (2-year)	241	11.8
	Bachelor's degree in college (4-year)	439	21.4
	Master's degree	133	6.5
	Doctoral degree	28	1.4
	Professional degree (JD,MD)	19	.9
Income	Less than $10,000	371	18.1
	$10,000 to $19,999	213	10.4
	$20,000 to $29,999	319	15.6
	$30,000 to $39,999	243	11.9
	$40,000 to $49,999	195	9.5
	$50,000 to $59,999	221	10.8
	$60,000 to $69,999	105	5.1
	$70,000 to $79,999	124	6.0
	$80,000 to $89,999	49	2.4
	$90,000 to $99,999	65	3.2
	$100,000 to $149,999	91	4.4
	$150,000 or more	54	2.6
Political affiliation	Republican	262	12.8
	Democrat	1117	54.5
	Independent	504	24.6
	Something else	167	8.1

References

1. Greene, S. Social identity theory and party identification. *Soc. science quarterly* **85**, 136–153 (2004).

2. Umaña-Taylor, A. J., Yazedjian, A. & Bámaca-Gómez, M. Developing the ethnic identity scale using eriksonian and social identity perspectives. *Identity: An international journal theory research* **4**, 9–38 (2004).

3. Williams, D. On and off the'net: Scales for social capital in an online era. *J. computer-mediated communication* **11**, 593–628 (2006).

4. Somera, L. P., Lee, H.-R., Badowski, G. & Cassel, K. Health information seeking, source trust, and culture: a comparative analysis of health information trends and needs between guam and the united states. *J. health communication* **21**, 469–478 (2016).

5. Rodrigues, U. M. & Xu, J. <? covid19?> regulation of covid-19 fake news infodemic in china and india. *Media Int. Aust.* **177**, 125–131 (2020).

6. Kim, T., Barasz, K. & John, L. K. Why am i seeing this ad? the effect of ad transparency on ad effectiveness. *J. Consumer Res.* **45**, 906–932 (2019).

7. Jung, M., Lin, L. & Viswanath, K. Associations between health communication behaviors, neighborhood social capital, vaccine knowledge, and parents' h1n1 vaccination of their children. *Vaccine* **31**, 4860–4866 (2013).

8. Huh, J., DeLorme, D. E. & Reid, L. N. The information utility of dtc prescription drug advertising. *Journalism & Mass Commun. Q.* **81**, 788–806 (2004).

9. Chan-Olmsted, S., Rim, H. & Zerba, A. Mobile news adoption among young adults: Examining the roles of perceptions, news consumption, and media usage. *Journalism & Mass Commun. Q.* **90**, 126–147 (2013).