# Leveraging Artificial Intelligence to Predict Health Belief Model and COVID-19 Vaccine Uptake Using Survey Text from US Nurses

Samaneh Omranian *, Alireza Khoddam, Celeste Campos-Castillo, Sajjad Fouladvand, Susan McRoy, Janet Rich-Edwards

Posted Date: 25 December 2023

doi: 10.20944/preprints202312.1877.v1

Keywords: COVID-19 vaccination; healthcare providers; Nurses’ Health Study; vaccine hesitancy; Health Belief Model; artificial intelligence; natural language processing; text classification

*Article*

# Leveraging Artificial Intelligence to Predict Health Belief Model and COVID-19 Vaccine Uptake Using Survey Text from US Nurses

**Samaneh Omranian [1,2,*], Alireza Khoddam [1], Celeste Campos-Castillo [3], Sajjad Fouladvand [4], Susan McRoy [2] and Janet Rich-Edwards [1]**

[1] Division of Women's Health, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School; Boston, MA 02115
[2] Department of Computer Science Department, University of Wisconsin-Milwaukee; Milwaukee, WI 53211
[3] Department of Media and Information, Michigan State University, East Lansing, MI 48824
[4] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305
* Correspondence: omranian@uwm.edu

**Abstract:** We investigated how Artificial Intelligence (AI) reveals factors shaping COVID-19 vaccine hesitancy among healthcare providers by examining their open-text comments. We conducted a longitudinal survey starting in spring of 2020 with 38,788 current and former female nurses in three national cohorts to assess how the pandemic has affected their livelihood. In January and March-April 2021 surveys, participants were invited to contribute open-text comments and answer specific questions about COVID-19 vaccine uptake. A closed-ended question in the survey identified vaccine-hesitant (VH) participants who either had no intention or were unsure of receiving a COVID-19 vaccine. We collected 1,970 comments from VH participants and trained two Machine Learning (ML) algorithms to identify behavioral factors related to VH. The first predictive model classified each comment into one of three Health Belief Model (HBM) constructs (barriers, severity, and susceptibility) related to adopting disease prevention activities. The second predictive model used the words in January comments to predict the vaccine status of VH in March-April 2021; vaccine status was correctly predicted 89% of the time. Our results showed that 35% of VH participants cited barriers, 17% severity, and 7% susceptibility to receiving a COVID-19 vaccine. Out of the HBM constructs, the VH participants citing a barrier, such as allergic reactions and side effects, had the most associated change in vaccine status from VH to later receiving a vaccine.

**Keywords:** COVID-19 vaccination; healthcare providers; Nurses' Health Study; vaccine hesitancy; Health Belief Model; artificial intelligence; natural language processing; text classification

## 1. Introduction

Healthcare providers worldwide became the frontline workers in battling COVID-19 by treating infected patients. Healthcare providers encountered numerous challenges increasing their risk of infection, such as inadequate Personal Protective Equipment (PPE), high workloads, and extended shifts [1]. In December 2020, US healthcare personnel began to be offered mRNA vaccines (Pfizer-BioNTech and Moderna) under emergency use authorizations (EUA) [2], a critical step toward protecting healthcare providers [3]. Several surveys conducted before the EUA documented that 8-18% of healthcare personnel expressed hesitancy about the safety and efficacy of the new vaccines [4]. Vaccine hesitancy dropped after the EUA and the vaccine rollout to healthcare personnel (HCP), yet a small minority of nurses remained vaccine hesitant even into spring 2021 [5–7]. Determining why hesitancy lingered is a challenge. Heyerdahl et al. coined 'unspoken vaccine hesitancy' in describing when nurses may be uncomfortable expressing their vaccine-related concerns and beliefs due to the social and institutional pressure to get vaccinated [8].

Uncovering and understanding the factors behind vaccine hesitancy is crucial to increasing vaccine confidence in healthcare institutions and mitigating the spread of COVID-19. Previous studies have shown that open-ended questions in healthcare surveys are valuable resources to elicit

 doi:10.20944/preprints202312.1877.v1



respondents' concerns [9] and can reduce response biases stemming from respondent beliefs about desired outcomes [10]. Therefore, to examine what underlies vaccine hesitancy among nurses and what factors influenced uptake, as expressed in their own words, we leveraged a national longitudinal survey of US nurses and applied Machine Learning (ML) methods to the open-text comments regarding vaccines in January 2021 (winter 2021 survey) as predictors of survey respondents' vaccination status in March-April 2021 (spring 2021 survey). Among a cohort of initially vaccine hesitant nurses who included comments, 40% received a COVID-19 vaccine by spring 2021, but 60% remained unvaccinated. To help understand why, we employed constructs from the Health Belief Model to help interpret the nurses' rationales in their responses [11].

### *1.1. Gaining insight from open-field comments*

Qualitative research can help understand the concerns and opinions of healthcare providers concerning COVID-19 vaccines. By analyzing respondents' words, qualitative research enables us to gain insights into their beliefs, experiences, attitudes, behavior, and interactions [12]. Traditional qualitative analysis involves human coders reading respondents' words and categorizing similar responses into multiple codes or themes through an iterative consensus process [13,14].

Manually analyzing high-volume, multi-class data collected during the qualitative research process is tedious and time-consuming. Over the past decades, ML has incorporated ideas from psychology, sociology, statistics, and mathematics to enable computers to predict outcomes based on specific predictors [15]. The core functionality of machine learning is to train computers to automatically solve problems like classification using data or prior experience [16]. We used ML in two ways: 1) assist with the qualitative research process by classifying text into multiple codes; 2) use the multiple codes to predict changes in vaccine status.

### *1.2. Health Belief Model Constructs*

The codes we used to analyze open-field comments were developed from the Health Belief Model (HBM), which can be applied to understand why healthcare providers receive or decline the COVID-19 vaccine. HBM is a psychosocial model that researchers use to understand people's health-related behavior, including their decision to adopt or decline disease prevention measures [11]. Previous research supports the utility of HBM constructs in predicting the uptake of vaccines and measures to prevent being infected with the virus that causes COVID-19 [17].

According to HBM, whether an individual adopts a health-promoting behavior depends on five beliefs: (i) the individual's perception of themselves as prone to disease or health risks (perceived susceptibility), (ii) their feeling of the severity of the disease (perceived severity), (iii) their challenge of taking preventative actions (perceived barriers), (iv) their view of the benefits of taking those actions (perceived benefits), and (v) their view of how well they can successfully implement the recommended health behavior (self-efficacy). Since VH individuals are hesitant about receiving a vaccine, we reasoned that the benefit and self-efficacy HBM constructs, which focus on the positives of vaccination, would not capture the health behavior of VH individuals [11]. Therefore, we focused on detecting severity, susceptibility, and barriers within the comments from nurses.

## 2. Materials and Methods

We developed two ML models using open-text comments provided by VH nurses in two consecutive surveys in winter-spring 2021. The first model identifies the presence of HBM constructs within the comments, and the second uses this information to predict change in VH between the two surveys.

### *2.1. Study Population and Data Collection*

We first designed our study population to investigate the change in COVID-19 vaccine hesitancy among nurses who had participated in one of several ongoing large-scale studies. In the spring of 2020, we launched a series of surveys regarding the COVID-19 exposures and experiences of

 doi:10.20944/preprints202312.1877.v1



participants in the Nurses' Health Study II and Nurses' Health Study 3. The Nurses' Health Study II (NHSII) was initiated in 1989 with 116,429 female registered nurses (RNs) aged 25–42 resident in 14 states. The Nurses' Health Study 3 (NHS3) is an open cohort launched in 2010 that continues to enroll nurses and nursing students aged 18 and older, born since January 1, 1965. This cohort includes RNs, licensed practical and vocational nurses, specialized RNs, and nursing students. The NHS3 cohort was expanded to include male nurses in 2015. Altogether, the NHS studies cover more than 280,000 participants.

In April-May 2020, we invited participants who had returned the most recent primary cohort questionnaires to complete a supplementary COVID-19 survey. Exclusions to the baseline invitation, such as the lack of a valid email address, are detailed in a previous study [1]. Of 105,662 invited participants, 58,606 (55%) completed the baseline survey. Respondents were surveyed again 1, 2, 3, 6, 9, and 12 months after the initial survey. Data collection for each survey was rolled out over three weeks. We restricted the current analysis to female credentialed current, active and former nurses living in the United States who responded to the spring 2021 survey [20]. This restriction left us with 38,788 participants.

We defined vaccine hesitant (VH) as a participant who answered 'no' or 'unsure' to the winter 2021 survey question, 'Do you plan to receive a COVID-19 vaccine?' According to our definition, of 38,788 participants, we excluded 18,395 participants who had already been vaccinated at the time the winter survey, 337 participants who had missing data on vaccination status, and 15,814 participants who indicated they had a plan to get a COVID-19 vaccine soon, leaving a population of 4,242 VH individuals. The winter 2021 survey also provided two opportunities for respondents to express their thoughts: 1) an unprompted open-text box after the vaccine questions and 2) an open-text box at the end of the survey, prompting, 'We are interested in learning more about your experiences during this pandemic. Please add anything else you would like to tell us here.' Of the 4,242 VH individuals in winter 2021, we excluded 2,741 participants who didn't write any comments in the open-text boxes, leaving a study population of 1,501 of VH participants with comments. Figure 1 depicts the flowchart of our study population.

We used this dataset of 1,501 comments as input to train an HBM ML model that predicts whether each comment belonged to any of the three HBM constructs (barriers, severity, and susceptibility) or was a non-HBM comment.

To monitor the vaccine status of survey participants three months after winter 2021, we asked the same question about vaccine status in spring 2021 survey. From the initial pool of 1,501 VH individuals surveyed in winter 2021, 96 participants who didn't specify their vaccine status were excluded, resulting in 1,405 VH participants. Our focus for subsequent model development was narrowed down to these 1,405 individuals who participated in both our winter and spring 2021 surveys. As delineated in Figure 1, our second model is designed to predict the vaccine status of these VH individuals in spring 2021.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



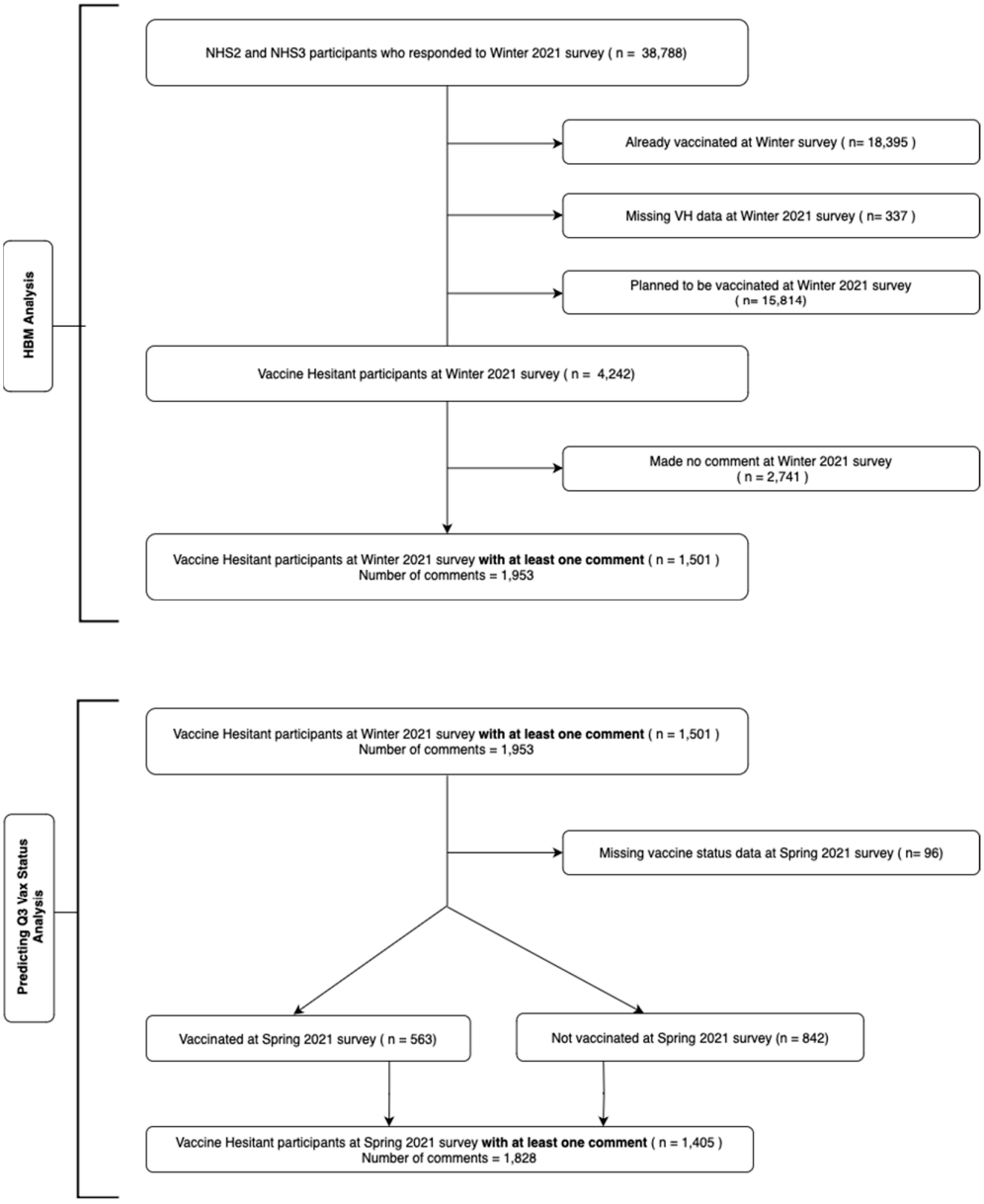


**Figure 1.** The flowchart diagram of the population lection for training ML models. Model 1 categorizes comments from winter 2021 into HBM constructs. Model 2 uses comments from VH individuals in winter 2021 to predict vaccine status in spring 2021.

*2.2. Machine Learning Models*

In this study, we trained two ML models: 1) a model to predict an HBM construct for each comment of VH participants in winter 2021 and 2) a model to predict a change in vaccine status of VH individuals from winter 2021 to spring 2021.

To develop the HBM model, four expert annotators created a training sample set by manually annotating 300 (16% of all VH comments) comments as one of the three mentioned HBM constructs or non-HBM-related comments. Table 1 shows the guidelines for classifying the three HBM constructs, definitions, and example comments from the winter 2021 survey.

 doi:10.20944/preprints202312.1877.v1



**Table 1.** Health Belief Model constructs, definitions, and example comments from the VH comments in the winter 2021 survey.

| HBM Construct | Definition | Example Comment |
|---|---|---|
| **Perceived Barriers** | Belief about the tangible and psychological costs of the advised action. | "Not fully tested or approved and is unnecessary for someone low risk like me. Also has not been studied for its effects on fertility and future pregnancies. I also know someone person that died 2 days after receiving the vaccine who had no medical conditions other than being overweight." |
| **Perceived Severity** | Feelings about the seriousness of contracting an illness or of leaving it untreated include evaluations of both medical and clinical consequences (for example, death, disability, and pain) and possible social consequences (such as the effects of the conditions on work, family life, and social relations). | "There is a 99.7% survival rate for someone my age anyway." |
| **Perceived Susceptibility** | Belief about the chances of getting a condition or disease. | "95-98% cure rate I am (in) good health." |
| **Non-HBM** | Comments that do not fall into any of the above categories. | "I needed to travel in mid January as my dad had major surgery and needed someone to be with him." |

We utilized the Python programming language and its built-in scikit-learn free software machine learning library for developing and analyzing our ML models. Before training our two models, we first preprocessed the data. This involved data cleaning, segmenting each comment into words, eliminating low-variance terms (words) related to the output variables and converting words to numerical values. The details of the data preprocessing and feature selection can be found in Appendix 1.

For each training task, we trained five ML models: Random Forest (RF) [18], Multinomial Naïve Bayes (Multinomial NB), Logistic Regression, Scholastic Gradient Descent (SGD) [19], and a multi-layer perceptron Neural Network (NN) [20], and then selected the model with the best performance. The details of the data preprocessing and the features used in the ML models can be found in Appendix A.

## 3. Results

Our analysis of input feature reduction dimensionality by variance threshold method revealed that a subset of 705 high-variance input features is the best for training an HBM model for comments related to COVID-19 vaccination, and a subset of 430 input features is the optimal size for training a change in the vaccine hesitancy model. To provide a more comprehensive understanding of both models performance, we used weighted average recall (sensitivity), precision, F1-measure, accuracy, and Area Under the Curve for the Receiver Operating Characteristic curve (AUC-ROC).

Recall, also known as sensitivity and true positive rate, is a metric that measures the proportion of accurate positive predictions among all possible positive predictions. Recall ranges from 0 to 1 and high recall indicates that the model can identify most of the positive instances and is calculated using the equation below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

We explain the details of each model in the following sections. A precision measure calculates how many correctly positive predictions were made and is calculated using the equation below:

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

F1-measure is a harmonic mean of precision and recall and ranges from 0 (prediction failure) to 1 (perfect prediction). A high F-measure indicates both good precision and recall, meaning the model has a good balance between minimizing false positives and false negatives.

$$F1 - measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy is a common metric used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances out of the total instances in the dataset.

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + False\ Negative + True\ Negative}$$

The Area Under the Curve (AUC) of the ROC curve is a scalar value that quantifies the overall performance of a classification model. A model with no predictive power would have an AUC of 0.5. This is equivalent to a random guess. A model with perfect predictive power would have an AUC of 1. A steeper ROC curve generally indicates better model performance.

### *3.1. HBM Prediction*

Table 2 depicts the performance of the models we trained to predict the label for the unlabeled dataset (the rest of data that we didn't use for HBM annotation). The label was any of the three constructs of HBM (barrier, susceptibility, and severity) or a non-HBM-related construct. The NN model outperformed other models with 82% accuracy. As Figure 7 (b) shows the NN model demonstrated robust performance with an area under the ROC curve (AUC-ROC) of 91%, indicates its high discriminatory ability. Figures 2–5 show the most influential meaningful words in detecting HBM constructs in all comments from VH individuals.

**Table 2.** Machine Learning performance results for HBM Prediction model using different combinations of vectorization methods (column) and classifiers (rows).

| Machine Learning Algorithm | Recall (Sensitivity) | Precision | F1-Score | Accuracy | AUC-ROC |
|---|---|---|---|---|---|
| Random Forest | 0.60 | 0.61 | 0.59 | 0.60 | 0.84 |
| Multinomial NB | 0.60 | 0.57 | 0.56 | 0.72 | 0.86 |
| Logistic Regression | 0.58 | 0.54 | 0.55 | 0.68 | 0.86 |
| SGD | 0.61 | 0.63 | 0.61 | 0.67 | 0.87 |
| Neural network | 0.82 | 0.85 | 0.79 | 0.82 | 0.91 |

The results of our subsequent data analysis of the words in VH comments for each HBM construct are shown in Figure 6. We found that 35% of the comments from the VH nurses indicated perceived barriers to vaccination and that this was the most frequently expressed belief type. Our ML analysis on HBM constructs also suggested more specific details about these barriers; the most significant barriers to receiving the vaccine for the VH participants appear to be side effects and allergic reactions, with other concerns regarding pregnancy and fertility that were potentially related to the lack of research on the vaccine side effects within these domains.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



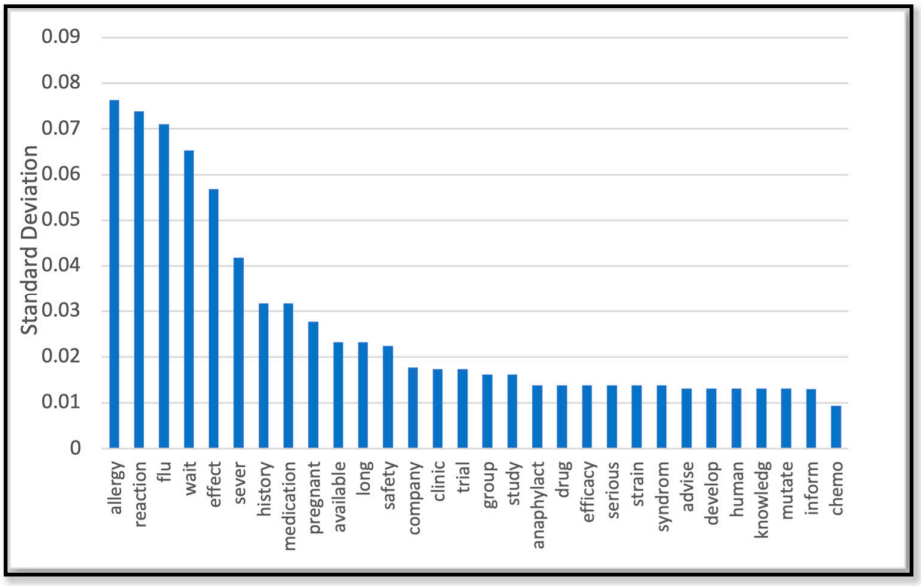


**Figure 2.** Top 30 most effective words in detecting the barrier construct.

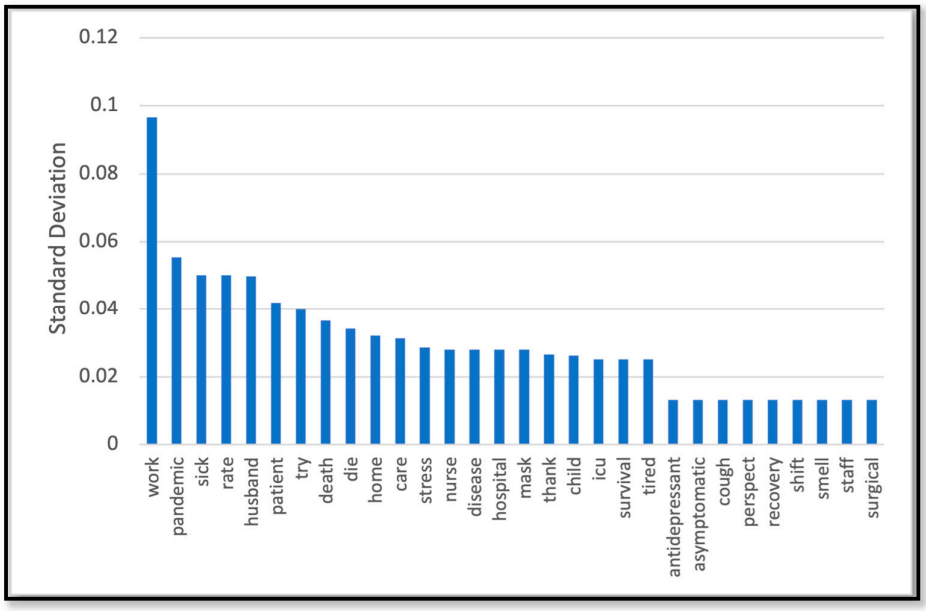


**Figure 3.** Top 30 most effective words in detecting the severity construct.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



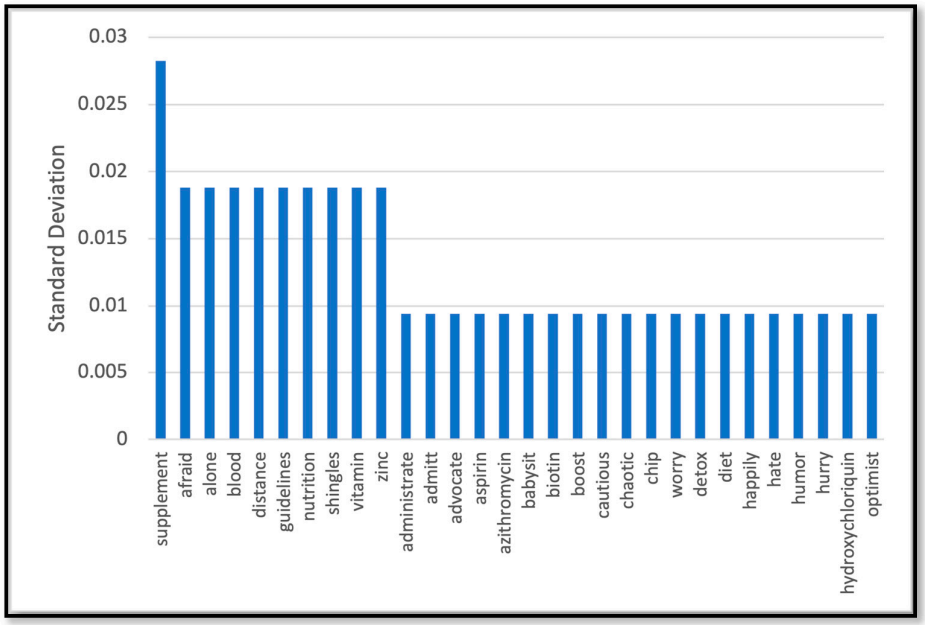


**Figure 4.** Top 30 most effective words in detecting the susceptibly construct.

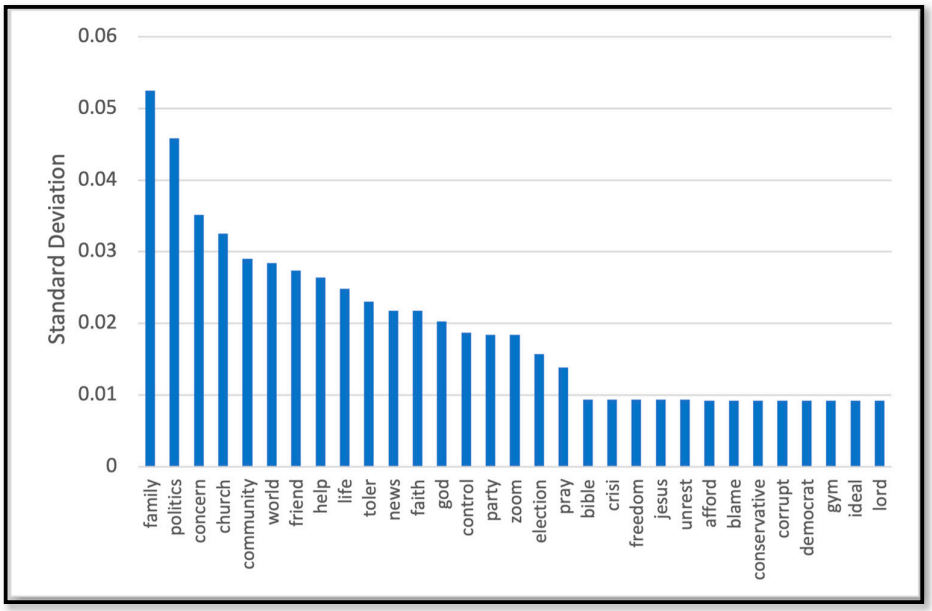


**Figure 5.** Top 30 most effective words in detecting the non-HBM group.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



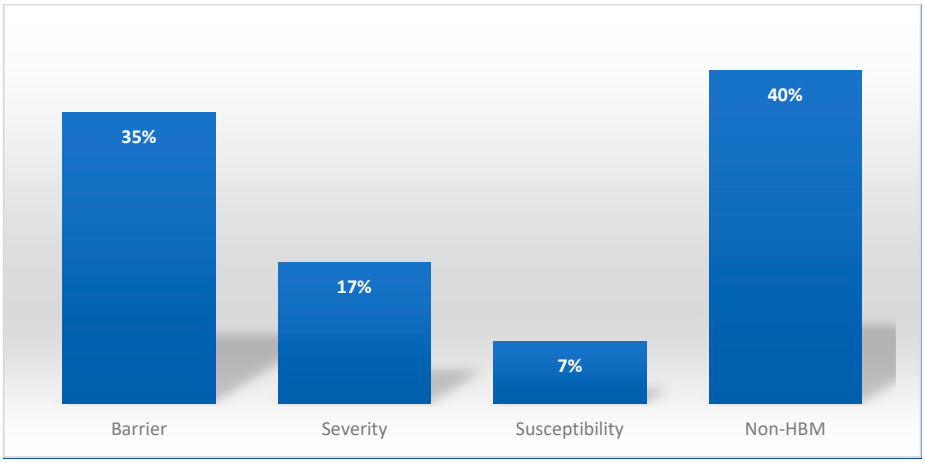


**Figure 6.** Distribution of the health belief model (HBM) classification on 2,424 comments of nurses on the winter 2021 survey.

The HBM classification also revealed that 17% of VH comments mentioned a perceived lack of severity of COVID-19. The most related topics under this category tended to mention high recovery rates from contracting COVID-19. An additional 16% of VH comments mentioned low susceptibility to contracting COVID-19. This category mentioned words suggesting respondents felt they already had a robust immune system, such as 'supplements,' 'boost,' and 'diet.'

*3.2. Results for Vaccine Status Prediction*

The model used the comments from VH individuals in winter 2021 to predict vaccine status of the VH in spring 2021. After expressing vaccine hesitancy in the winter 2021 survey, 40% of nurses received a COVID-19 vaccine by spring 2021, while 60% remained unvaccinated. Table 3 shows the performance of the trained models for predicting the vaccine status from VH comments. The best accuracy achieved was 89% using a NN. Figure 7 (a) shows the AUC-ROC curve of the model performance.

**Table 3.** Machine Learning accuracy results for Vaccine Status Prediction model using different combinations of vectorization methods (column) and classifiers (rows).

| Machine Learning Algorithm | Recall (Sensitivity) | Precision | F1-Score | Accuracy | AUC-ROC |
|---|---|---|---|---|---|
| Random Forest | 0.71 | 0.76 | 0.72 | 0.78 | 0.83 |
| Multinomial NB | 0.84 | 0.89 | 0.86 | 0.88 | 0.86 |
| Logistic Regression | 0.69 | 0.86 | 0.71 | 0.79 | 0.85 |
| SGD | 0.78 | 0.81 | 0.79 | 0.82 | 0.83 |
| Neural network | 0.89 | 0.90 | 0.89 | 0.89 | 0.88 |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 December 2023 doi:10.20944/preprints202312.1877.v1



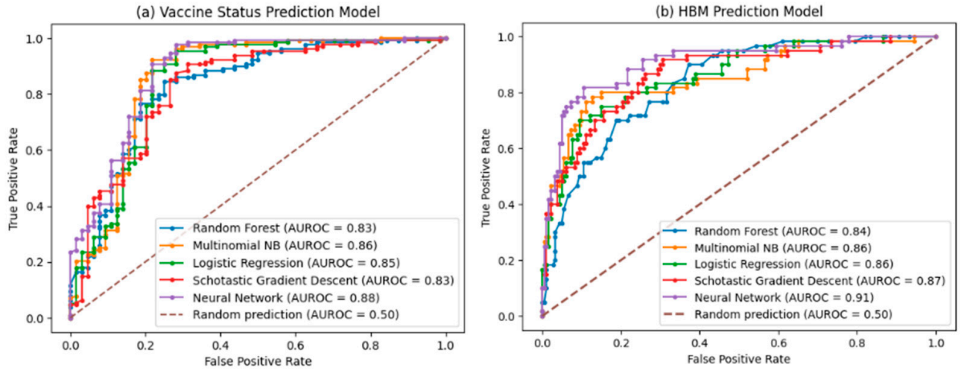


**Figure 7. (a)** Shows the AUC-ROC curve of the vaccine status prediction model; the Neural Network model outperformed other models with the AUC of 0.88. (b) shows the AUC-ROC curve of predicting HBM constructs. As shown the Neural Network model achieved the highest AUC = 0.91 among other models. .

## 4. Discussion

This study used open-text comments collected in winter 2021 to predict vaccine status and hesitancy three months later. Throughout this time, the vaccines were accessible only under an Emergency Use Authorization and not yet mandated for nurses to be vaccinated. The winter 2021 survey results showed that 11% of the participating nurses expressed vaccine hesitancy and 4% wrote at least one comment. In spring 2021, 60% of the VH in winter 2021 remained hesitant three months later. Our results suggest that while most early VH may be due to perceived barriers, there is a small cohort for whom a belief in low COVID-19 severity or low personal susceptibility will linger. Our findings extend previous studies using qualitative coding to analyze COVID-19-related surveys [24–26].

We found the most commonly occurring HBM construct within the comments was perceived barriers, suggesting concerns over the tangible and intangible costs associated with getting the vaccine were prevalent among VH nurses. Specific costs mentioned included side effects and allergic reactions, indicating opportunities to develop messaging to allay these concerns. Pregnancy and fertility concerns were also frequently mentioned among comments assigned to this construct, likely owing to the lack of evidence regarding the vaccine's safety among pregnant women and the messaging uncertainty among obstetricians [27].

This study takes the unique approach of using ML to categorize open-text comments into HBM constructs. Both HBM classifier's 82% accuracy rate to predict HBM constructs and 89% accuracy rate for predicting vaccine status from text reveal that implementing ML has the potential to automate the qualitative research project. In addition, the area under ROC curve for both models exceeding 80% indicates robust model performance, signifying models' effectiveness in predicting the targets compared to random guessing.

Our methods would also be helpful for future research augmenting ML with health behavior models. Such theoretical models of human behavior boast clear demarcations between constructs, necessitating corresponding ML methods to identify distinguishing features (e.g., words). Our analysis of feature selection revealed that a subset of high-variance words from each comment, rather than the entire comment, leads to improved performance in machine learning models. Additionally, we observed that excluding words representing two classes equally enhanced model performance. Similar approaches will likely yield better performance of ML methods that adopt a theoretical framework.

In our study, among several classical ML models we tested, the neural network model surpassed all other models, achieving the highest performance in both the HBM model and vaccine status prediction models. We also discovered that applying variance threshold for input feature dimensionality reduction helped to optimize the ML models training.

 doi:10.20944/preprints202312.1877.v1



Readers should bear in mind the study's limitations when drawing conclusions. First, our findings only apply to English-speaking nurses in the US, as only English text was used to develop the model. Additionally, the cohort primarily consisted of individuals of white ethnicity hence our ML model has not been trained to recognize how other racial groups feel about the COVID-19 vaccines. Moreover, not every VH nurse responded to the open-text questions used in the analysis. Lastly, in considering susceptibility, this category could potentially carry a valence. For instance, a comment could mention that they are choosing to not receive a vaccine because they are not susceptible to COVID. Opposingly, a comment could mention that they are choosing to receive a vaccine because they are susceptible to COVID. Since our study population is VH individuals, we chose to focus our analysis on the former valency of the susceptibility category. By expanding our study population to all individuals, future studies can consider both valences. Nonetheless, we were able to obtain a high accuracy rate. Despite these limitations, we believe that ML has the ability to streamline qualitative research. As the field of ML advances with newer models, researchers in other fields have an opportunity to automate tedious tasks and extract useful information from their data.

**Author Contributions:** Conceptualization, S.O. and A.K.; methodology, S.O. and S.F.; resources, J.RE.; data curation, S.O.; writing—original draft preparation, S.O.; writing—review and editing, J.RE., C.CC., S.M., S.F, and A.K.; visualization, S.O. and A.K.; supervision, J.RE. and S.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study protocol was approved by the institutional review boards (IRBs) of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health (protocol: 2020P001020).

**Informed Consent Statement:** The IRBs allowed participants' completion of questionnaires to be considered as implied consent.

**Data Availability Statement:** Our public website (www.nurseshealthstudy.org) includes a brief description of the Nurses' Health Study cohorts, all questionnaires, and a description of resource sharing procedures. An automated online form requests the applicant to briefly describe the hypothesis and aims, variables needed, etc. Requests are presented to the cohort investigator meetings every other week, and replies are provided within 2 weeks. After appropriate institutional IRB approvals, data access occurs in one of three ways: 1) the external investigator, with a password-protected login to our system, securely accesses and analyzes cohort data/specimen results; 2) the external investigator requests collaboration with or support of an internal investigator and/or programmer who conducts the analyses on the external investigator's behalf; or 3) a specific dataset and data dictionary are created to send to the external collaborator. Most often, investigators are provided with direct access to the cohort computing system; the third option is usually reserved for consortia projects pooling data from cohorts. Access is provided with evidence of human subjects training (required by our IRB) and a standard data use agreement. Login to the computing system is easily done from anywhere in the world, with a password-protected logon, which provides access to data systems and intranet site, with educational materials and documentation. These are the same online data, materials, and tools accessed by internal investigators.

**Acknowledgments:** We express our gratitude to Iris Ayse Becene for her invaluable contribution to this research project. Additionally, we extend our thanks to Sayan Banerjee for his insightful ideas on Machine learning models. We would also like to express our gratitude to Mehrzad Khoddam for his statistical expertise, which enriched the quality of our research. .

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A: Data Preprocessing and Feature Selection

Prior to applying the trained model, we preprocessed the data, including data cleaning, splitting each comment into words, and removing low-variance terms concerning the output variables (sometimes described as feature selection or dimensionality reduction). Data cleaning generally involves organizing the data according to its structure and removing irrelevant subtexts [29].

Feature selection is necessary because working with this volume of words without any filtering may result in higher memory requirements, longer execution times, yet lower performance of ML algorithms [29,30]. Based on common practice, we used variance thresholding, an NLP preprocessing

 doi:10.20944/preprints202312.1877.v1



technique, to identify the most discriminatory words. Low-variance words can either be highly frequent or infrequent in text data, are specific to a particular dataset and are not discriminatory [29]. For instance, since the word 'covid' appears frequently across different classifications in our surveys, it is not an informative word for making a prediction.

In this study, we initially computed a measure representing the percentage of how many times a word appears in each class, denoted as p(w,c). Subsequently, the variance of p(w,c) values was calculated as:

$$v(w)=p(w,i), \quad \text{for } i=1,2,\ldots,N$$

Here, N denotes the total number of classes, which is four in our case. To determine the cutoff percentage of occurrence, which denotes significance, we conducted a grid search to identify the optimal value that maximizes the accuracy. Words with v(w) greater than $\delta^2 = 0.001$ thresholds were retained.

Additionally, we performed a second round of feature selection. During this step, words representing two classes equally were removed. This was achieved by evaluating the ratio of p(w,i) for the top two classes associated with each word. A ratio of 1.0 indicates equal representation, while a ratio of 0.0 signifies occurrence in only one class. We conducted another grid search to determine the ratio of association with two classes and found a set cutoff value.

## References


1. Dooling, K.; Marin, M.; Wallace, M.; McClung, N.; Chamberland, M.; Lee, G. M.; Talbot, H. K.; Romero, J. R.; Bell, B. P.; Oliver, S. E. The Advisory Committee on Immunization Practices' Updated Interim Recommendation for Allocation of COVID-19 Vaccine — United States, December 2020. *MMWR Morb. Mortal. Wkly. Rep.* **2021**, *69* (5152), 1657–1660. https://doi.org /10.15585/mmwr.mm695152e2.
2. Rich-Edwards, J. W.; Ding, M.; Rocheleau, C. M.; Boiano, J. M.; Kang, J. H.; Becene, I.; Nguyen, L. H.; Chan, A. T.; Hart, J. E.; Chavarro, J. E.; Lawson, C. C. American Frontline Healthcare Personnel's Access to and Use of Personal Protective Equipment Early in the COVID-19 Pandemic. *Journal of Occupational & Environmental Medicine* **2021**, *63* (11), 913–920. https://doi.org/10.1097/JOM.0000000000002308.
3. US Food and Drug Administration. (URL: https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines (April 29 2023).
4. Biswas, N.; Mustapha, T.; Khubchandani, J.; Price, J. H. The Nature and Extent of COVID-19 Vaccination Hesitancy in Healthcare Workers. *J Community Health* **2021**, *46* (6), 1244–1251. https://doi.org/10.1007/s10900-021-00984-3.
5. Toth-Manikowski, S. M.; Swirsky, E. S.; Gandhi, R.; Piscitello, G. COVID-19 Vaccination Hesitancy among Health Care Workers, Communication, and Policy-Making. *American Journal of Infection Control* **2022**, *50* (1), 20–25. https://doi.org/10.1016/j.ajic.2021.10.004.
6. Billings, J.; Ching, B. C. F.; Gkofa, V.; Greene, T.; Bloomfield, M. Experiences of Frontline Healthcare Workers and Their Views about Support during COVID-19 and Previous Pandemics: A Systematic Review and Qualitative Meta-Synthesis. BMC Health Serv Res **2021**, 21 (1), 923. https://doi.org/10.1186/s12913-021-06917-z.
7. Gharpure, R.; Guo, A.; Bishnoi, C. K.; Patel, U.; Gifford, D.; Tippins, A.; Jaffe, A.; Shulman, E.; Stone, N.; Mungai, E.; Bagchi, S.; Bell, J.; Srinivasan, A.; Patel, A.; Link-Gelles, R. Early COVID-19 First-Dose Vaccination Coverage Among Residents and Staff Members of Skilled Nursing Facilities Participating in the Pharmacy Partnership for Long-Term Care Program — United States, December 2020–January 2021. *MMWR Morb. Mortal. Wkly. Rep.* **2021**, *70* (5), 178–182. https://doi.org/10.15585/mmwr.mm7005e2.
8. Heyerdahl, L. W.; Dielen, S.; Nguyen, T.; Van Riet, C.; Kattumana, T.; Simas, C.; Vandaele, N.; Vandamme, A.-M.; Vandermeulen, C.; Giles-Vernick, T.; Larson, H.; Grietens, K. P.; Gryseels, C. Doubt at the Core: Unspoken Vaccine Hesitancy among Healthcare Workers. *The Lancet Regional Health - Europe* **2022**, *12*, 100289. https://doi.org/10.1016/j.lanepe.2021.100289.
9. The Nurses' Health Study and Nurses' Health Study II are among the largest investigations into the risk factors for major chronic diseases in women. URL: https://nurseshealthstudy.org (April 8 2023).
10. Riiskjaer, E.; Ammentorp, J.; Kofoed, P.-E. The Value of Open-Ended Questions in Surveys on Patient Experience: Number of Comments and Perceived Usefulness from a Hospital Perspective. *International Journal for Quality in Health Care* **2012**, *24* (5), 509–516. https://doi.org/10.1093/intqhc/mzs039.
11. Rosenstock, I. M. Historical Origins of the Health Belief Model. *Health Education Monographs* **1974**, *2* (4), 328–335. https://doi.org/10.1177/109019817400200403.
12. Pathak, V.; Jena, B.; Kalra, S. Qualitative Research. *Perspect Clin Res* **2013**, *4* (3), 192. https://doi.org/10.4103/2229-3485.115389.

13. Charmaz, K. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis; SAGE Publications Ltd, 2006.
14. Nguyen, K. T. N. H.; Stuart, J. J.; Shah, A. H.; Becene, I. A.; West, M. G.; Berrill, J.; Gelaye, B.; Borba, C. P. C.; Rich-Edwards, J. W. Novel Methods for Leveraging Large Cohort Studies for Qualitative and Mixed-Methods Research. *American Journal of Epidemiology* **2023**, *192* (5), 821–829. https://doi.org/10.1093/aje/kwad030.
15. Marsland, S. *Machine Learning: An Algorithmic Perspective*, 2nd ed.; Chapman and Hall/CRC, 2014. https://doi.org/10.1201/b17476.
16. 
17. Muhammad, I.; Yan, Z.; Southwest Jiaotong University, China. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *IJSC* **2015**, *05* (03), 946–952. https://doi.org/10.21917/ijsc.2015.0133.
18. Lu, J.; Luo, M.; Yee, A. Z. H.; Sheldenkar, A.; Lau, J.; Lwin, M. O. Do Superstitious Beliefs Affect Influenza Vaccine Uptake through Shaping Health Beliefs? *Vaccine* **2019**, *37* (8), 1046–1052. https://doi.org/10.1016/j.vaccine.2019.01.017.
19. Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
20. Ruder, S. (2016). An overview of gradient descent optimization algorithms. *ArXiv, abs/1609.04747*.
21. *Neural Network Models*. https://scikit-learn.org/stable/modules/neural_networks_supervised.html.
22. Uther, W.; Mladenić, D.; Ciaramita, M.; Berendt, B.; Kołcz, A.; Grobelnik, M.; Mladenić, D.; Witbrock, M.; Risch, J.; Bohn, S.; Poteet, S.; Kao, A.; Quach, L.; Wu, J.; Keogh, E.; Miikkulainen, R.; Flener, P.; Schmid, U.; Zheng, F.; Webb, G. I.; Nijssen, S. TF–IDF. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2011; pp 986–987. https://doi.org/10.1007/978-0-387-30164-8_832.
23. Le, Q. V.; Mikolov, T. Distributed Representations of Sentences and Documents. **2014**. https://doi.org/10.48550/ARXIV.1405.4053.
24. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp 1532–1543. https://doi.org/10.3115/v1/D14-1162.
25. Shmueli, L. Predicting Intention to Receive COVID-19 Vaccine among the General Population Using the Health Belief Model and the Theory of Planned Behavior Model. *BMC Public Health* **2021**, *21* (1), 804. https://doi.org/10.1186/s12889-021-10816-7.
26. Lin, Y.; Hu, Z.; Zhao, Q.; Alias, H.; Danaee, M.; Wong, L. P. Understanding COVID-19 Vaccine Demand and Hesitancy: A Nationwide Online Survey in China. *PLoS Negl Trop Dis* **2020**, *14* (12), e0008961. https://doi.org/10.1371/journal.pntd.0008961.
27. Mercadante, A. R.; Law, A. V. Will They, or Won't They? Examining Patients' Vaccine Intention for Flu and COVID-19 Using the Health Belief Model. *Research in Social and Administrative Pharmacy* **2021**, *17* (9), 1596–1605. https://doi.org/10.1016/j.sapharm.2020.12.012.
28. Rasmussen, S. A.; Kelley, C. F.; Horton, J. P.; Jamieson, D. J. Coronavirus Disease 2019 (COVID-19) Vaccines and Pregnancy: What Obstetricians Need to Know. *Obstet Gynecol* **2021**, *137* (3), 408–414. https://doi.org/10.1097/AOG.0000000000004290.
29. Dogra, V.; Verma, S.; Kavita; Chatterjee, P.; Shafi, J.; Choi, J.; Ijaz, M. F. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience* **2022**, *2022*, 1–26. https://doi.org/10.1155/2022/1883698.
30. Jun Yan; Benyu Zhang; Ning Liu; Shuicheng Yan; Qiansheng Cheng; Fan, W.; Qiang Yang; Xi, W.; Zheng Chen. Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing. *IEEE Trans. Knowl. Data Eng.* **2006**, *18* (3), 320–333. https://doi.org/10.1109/TKDE.2006.45.
31. Xu, X.; Liang, T.; Zhu, J.; Zheng, D.; Sun, T. Review of Classical Dimensionality Reduction and Sample Selection Methods for Large-Scale Data Processing. *Neurocomputing* **2019**, *328*, 5–15. https://doi.org/10.1016/j.neucom.2018.02.100.