

Vaccine Hesitancy Prediction Based on Machine Learning Algorithms

Yike Zhao *

Department of Statistics, Ludwig Maximilian University, Munich, Germany

* Corresponding author: mburns66@liberty.edu

Abstract. Modern medical technology has led to the development and implementation of vaccines for various infectious or malignant diseases, offering active acquired immunity to those who receive them. Despite extensive research and validation confirming vaccine safety, a segment of the population exhibits reluctance or refusal to get vaccinated, known as vaccine hesitancy. This hesitancy posed challenges during the COVID-19 pandemic, hindering the attainment of herd immunity and contributing to a higher number of preventable deaths. It is imperative to accurately predict vaccine hesitancy by analyzing various population characteristics to distinguish between those with and without vaccine hesitancy. Subsequently, different communication strategies and incentives can be adopted. In this article, the author utilizes private data from a survey that includes demographic features about the participants, as well as their behaviors and beliefs about the COVID-19 pandemic and vaccination. To identify the most significant features, the author employed the feature importance method to eliminate features that were less influential and narrowed to the 17 most influential features. Machine learning techniques, including non-ensemble, ensemble, and neural networks, were employed to predict the data set and then calculate testing errors to compare each method on prediction accuracy. The authors concluded that neural networks have superior prediction accuracy and are mostly immune to overfitting. Although neural networks take a long time to train, the test error of less than 5.5% is significantly lower than the 6.5% average error of other methods.

Keywords: COVID-19, vaccine hesitancy, machine learning, deep learning.

1. Introduction

As the lethality of COVID-19 decreased, the pandemic that had the strongest impact in the age of the internet gradually transitioned from a global disease to a regional flu [1]. Once again, the medical knowledge, hospital infrastructure, dedication of healthcare workers, and advances in medical technology, including vaccines as the primary form, have saved innumerable lives, showing humanity's remarkable resilience. However, individual religious, cultural, and social identities can result in varying attitudes toward vaccination. Amid this pandemic, an earlier adoption of belief in and vaccination by a larger portion of the population could potentially have accelerated the achievement of the herd immunity turning point, leading to fewer COVID-related deaths. Therefore, accurately distinguishing between people's attitudes toward the new vaccine, and providing tailored popularization and stimulation, is a crucial concern at the close of this pandemic.

Due to the widespread use of the Internet and social media, obtaining information on people's attitudes towards vaccines in various nations and comparing them is now simpler. A study indicated that the difference in vaccine hesitancy among populations was due, in part, to the individuals' geographic location [2]. Primarily, populations in East and Southeast Asia demonstrate high acceptance rates of vaccines, at approximately 80%. Latin America and Mexico closely follow with vaccine acceptance rates also exceeding 70%. In the United States, the United Kingdom, and Canada, acceptance rates have consistently remained above 60%. However, in Europe, vaccine acceptance rates vary widely by country. Vaccine acceptance rates in Mediterranean countries, Eastern Europe, and Russia are below 60%. Conversely, Western European nations generally surpass the 60% threshold, with Nordic countries like Denmark exceeding an 80% acceptance rate. Data is scarce for African countries, but one study indicates that South Africa and Nigeria exceed an 80% acceptance rate, while another study shows that northern Nigeria falls below a 30% acceptance rate. This

demonstrates Africa's geographical diversity and complexity as a continent, characterized by highly unequal development, surpassing that of Europe.

Meanwhile, several studies have investigated the impact of demographic factors on vaccine hesitancy, with participants from multiple countries [3]. To isolate predictors of vaccine compliance, the studies use a multivariate logistic regression [4]. Some studies suggest that vaccine acceptance is low in younger age groups, while others show that the middle-aged group (35-60 years old) has lower vaccine acceptance compared to both younger (18-34 years old) and older (60 years old and above) age groups. There are also study says that vaccine hesitancy appears to be higher among people aged 75 years and above. In terms of race, vaccine acceptance rates were lower among Black individuals. Acceptance was lower among the unemployed with regards to work status. Those who had been vaccinated in the past were more likely to accept new vaccines based on their health beliefs. Women showed less acceptance in terms of gender. Acceptance among individuals with low education was lower based on their educational status. Furthermore, those with lower incomes also showed lower acceptance; albeit several studies also suggested that income was not related to vaccine acceptance. These studies can help identify vaccine distrust and improve vaccine promotion by governments. Cash incentives, for instance, may lead to higher vaccine acceptance among low-income individuals, but this may not be as effective among middle-aged groups.

In the field of medical diagnostics, the use of machine learning is now more widespread. As the paper states, not only the naive Bayesian classifier, neural networks or decision trees, there is also a clear shift from machine learning methods to deep learning methods to deal with more complex and abundant medical data [5].

In this paper, the author utilized non-public data comprising 30 European countries. These data were collected during the later stages of the epidemic to eliminate instances of non-vaccination due to inadequate vaccine availability. Based on this, it is possible to estimate the respondents' vaccine hesitancy as the target variable by examining their vaccination status. Because the questionnaire data had incomplete responses, some parameters had missing values. This study initially filtered the valid responses and then use features importance to reduce the dimensions and extract the most relevant variables. The data was then split into training and test groups, and this study employed K-Nearest Neighbors, Support Vector Classifier, Random Forest, Extremely Randomized Trees, AdaBoost, Neutral network and more to forecast vaccine hesitancy.

2. Method

2.1. Data Preparation

These are private data from a questionnaire response. The questionnaire was distributed in March 2022, and 107,286 valid responses were collected. The questionnaire comprised 67 questions that focused on two areas, namely demographic characteristics, and attitudes towards COVID-19 regarding behavioral and mental states. Although the focus is on vaccination hesitancy as the target variable, the questionnaire did not inquire about this directly, given the ambiguity in many people's attitudes towards vaccination. The author utilized three interconnected questions pertaining to respondents' vaccination status to map onto the targeted variables. These questions progressively relate to one another and are not parallel in structure. See the accompanying figure for a visual representation of this relationship (Fig. 1). Using these questions, the author classified the participants into two groups: those who believe in vaccines which in the green dotted box and those who do not which in the red dotted box.

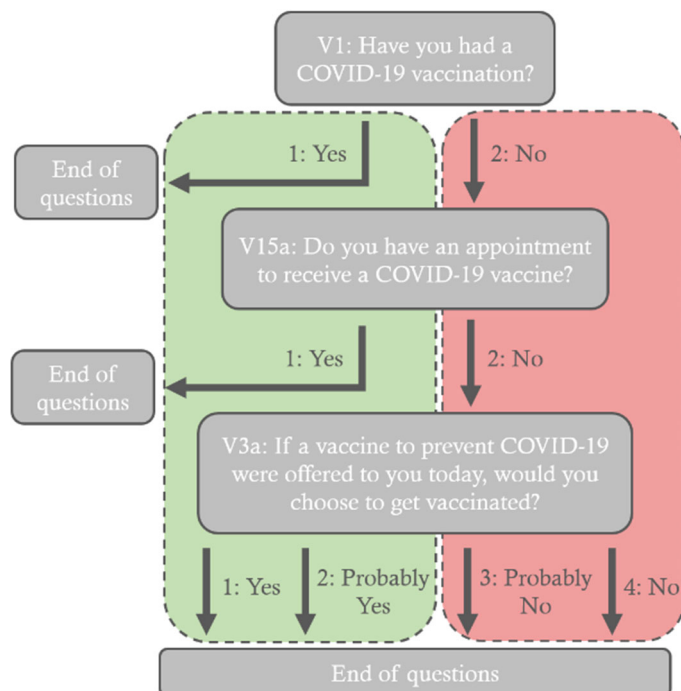


Figure 1. Workflow of the questions (Photo credit: Original)

2.2. Data Preprocessing

Because analyzing the 67-question questionnaire is challenging due to its large number and highly correlated questions that cannot be included in the model analysis concurrently, downscaling the data is necessary. To overcome this problem, the author experimented with different methods and eventually chose the feature importance method that yielded the best downscale results [6]. The importance values for each variable were computed, as shown below (Fig. 2). The author set a minimum threshold of 0.17 for all features, resulting in 18 remaining features. Five of the variables pertain to demographic characteristics, while the other 13 are associated with COVID-19 (Table 1). Fig. 3 displays variations in vaccine acceptance among the three demographic characteristics. The bar graph demonstrates the percentages, and the line graph indicates the absolute number of respondents. Among genders, female respondents had a higher vaccine acceptance than male respondents. Those who identified as "other" or "prefer not to answer" were significantly more distrustful of vaccines compared to "male" or "female" respondents. For respondents of different ages, vaccine hesitancy was slightly higher among middle-aged people (25-54) than among young people (18-24) and older (over 54). Additionally, vaccine hesitancy was lower among respondents who had just graduated from elementary school or had a college degree or higher. Fig. 4 displays a heat map of the correlation matrix of 3 series questions with the target variables using 'Kendall'.

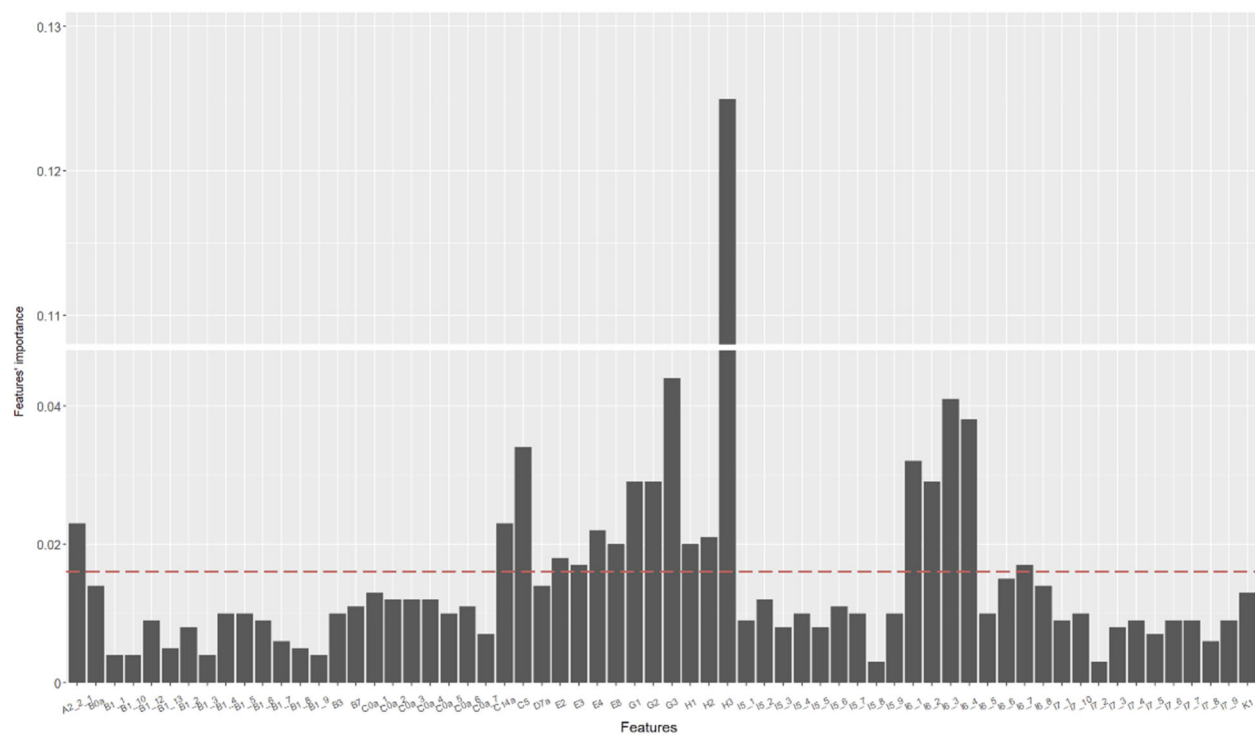


Figure 2. Importance of all features (Photo credit: Original)

Table 1. Description of the features

Variables	Type	Description
A2_2_1	Demographic	Country/Region
E2	Demographic	Staying area
E3	Demographic	Gender
E4	Demographic	Age
E8	Demographic	Education level
C5	Covid-related	Frequency of mask wearing
C14a	Covid-related	Frequency of avoiding others
G1	Covid-related	Severity of catching Covid
G2	Covid-related	Effective of social distance
G3	Covid-related	Effective of mask wearing
H1	Covid-related	Percentage of keeping social distance
H2	Covid-related	Percentage of mask wearing
H3	Covid-related	Percentage of family members vaccinated
I6_1	Covid-related	Attitude towards health workers
I6_2	Covid-related	Attitude towards scientists
I6_3	Covid-related	Attitude towards WHO
I6_4	Covid-related	Attitude towards government
I6_7	Covid-related	Attitude towards friends and family
Hesit.	Target variable	Vaccine hesitancy

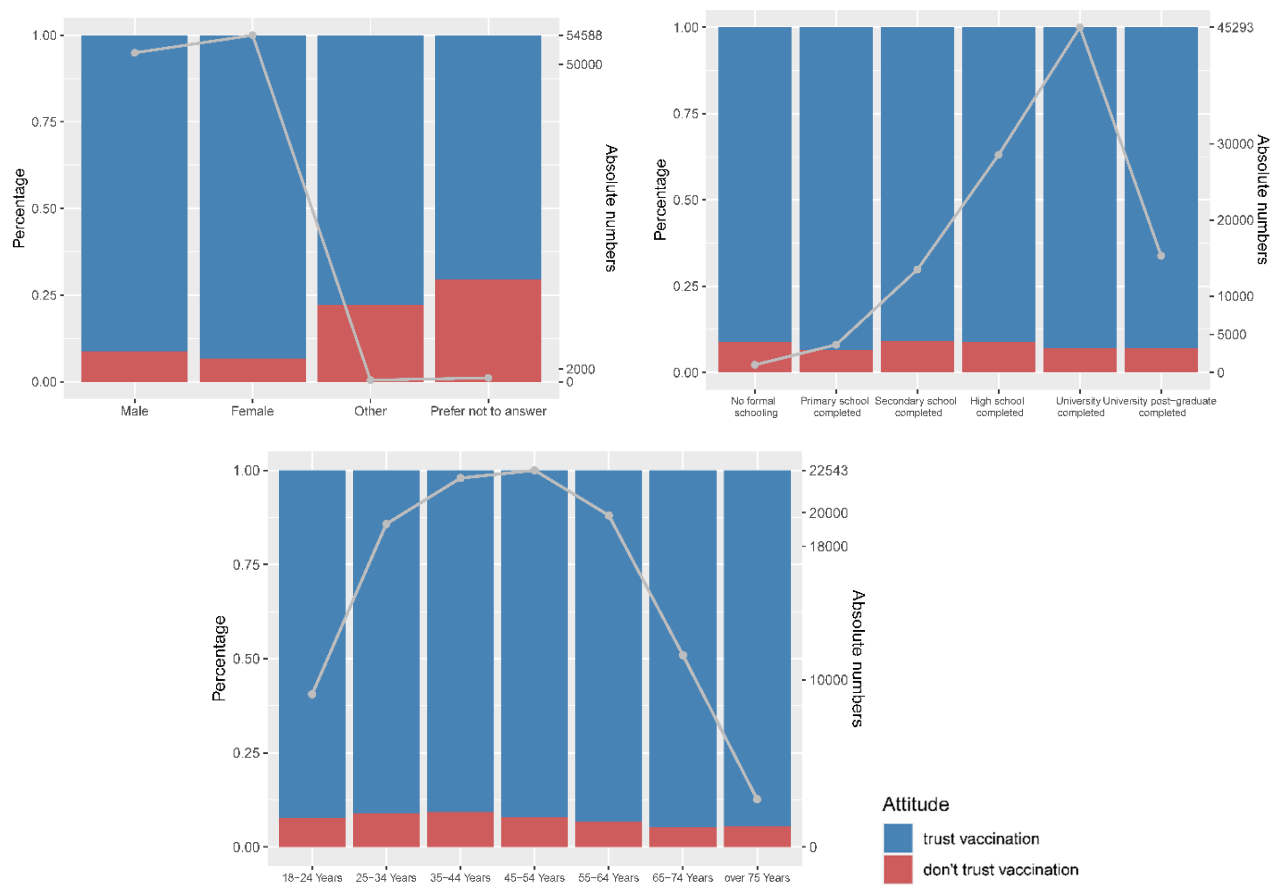


Figure 3. Vaccine hesitancy for three demographic features (Percentage and absolute numbers)
(Photo credit: Original)

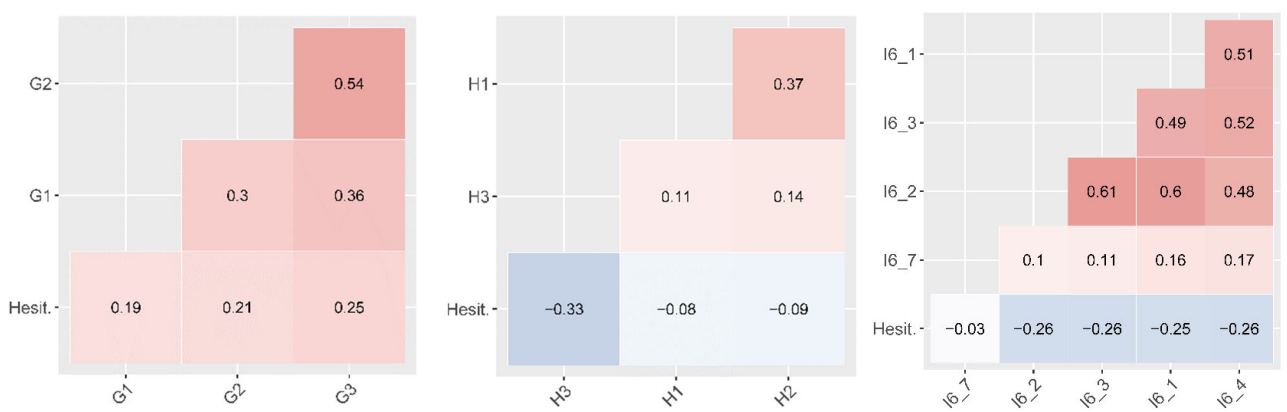


Figure 4. Heat map of the correlation matrix (Photo credit: Original)

2.3. Machine Learning Models

2.3.1. K-nearest Neighbors

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph). The K-Nearest Neighbors (KNN) is a versatile machine learning algorithm that falls into the supervised learning category which mainly used for classification and regression tasks.

KNN does not require a training phase, making it easy to understand and implement. However, it can be computationally intensive and struggle with high-dimensional datasets, as well as being sensitive to distance metrics. In this article, the value of the parameter 'n_neighbors' is defined as 10,

'weights' is defined as 'uniform' and 'algorithm' is defined as 'auto' which will automatically choose the most appropriate algorithm to fit model.

2.3.2. Support Vector Classifier

The Support Vector Classifier (SVC), also known as the Support Vector Machine (SVM), is a machine learning algorithm used for classifying data into multiple classes.

SVC performs well in high-dimensional feature spaces, making it suitable for complex data and outliers. It can handle both linear and nonlinear classification tasks using various kernel functions. However, training can be a time-intensive process. In this article, the kernel function is set to 'rbf' (radial basis function), 'degree' is defined as 3 and 'gamma' is defined as 'auto' which uses '1/n_features'.

2.3.3. Random Forest

Random Forest is a widely used ensemble learning method in machine learning. The main concept behind Random Forest is to create a forest of decision trees. Random Forest is widely recognized for enhancing predictive model accuracy and robustness. However, interpreting and visualizing the model poses challenges. In this article, the value of the parameter 'n_estimators' is defined as 100 and 'criterion' function is set to 'gini' to measure the quality of a split.

2.3.4. Extremely Randomized Trees

Extremely Randomized Trees (Extra-Trees) is an ensemble machine learning algorithm similar to Random Forest. The increased randomness in feature selection and splitting aids in reducing overfitting and improves algorithm stability. While Extra-Trees can be more efficient, they might not always outperform Random Forests, especially when the dataset is relatively clean and large. In this article, like the parameter in Random Forest, the value of the parameter 'n_estimators' is also defined as 100 and 'criterion' function is set to 'gini'.

2.3.5. Adaptive Boosting

Adaptive Boosting (AdaBoost) is a machine learning algorithm that belongs to the family of ensemble methods. It can significantly enhance the accuracy of weak learners, resulting in a powerful ensemble model. However, it is important to note that AdaBoost is sensitive to noisy data and outliers. In this article, the author uses the default estimator with 'n_estimators' set to 100. The 'learning_rate' parameter is defined as 1.0, and 'algorithm' is selected as 'SAMME.R', which converges at a faster rate than 'SAMME' to attain a lower test error with fewer boosting iterations.

2.3.6. Gradient Boosting

Gradient Boosting is an ensemble machine learning algorithm. The fundamental principle of Gradient Boosting involves building an ensemble of weak learners iteratively. Unlike AdaBoost, Gradient Boosting is robust to outliers, as it can assign lower weight to these data points, and it often generates highly accurate predictive models. In this article, the loss function is set to 'log_loss' which is appropriate to classification. The 'learning_rate' is set to 0.1, 'n_estimators' is set to 100 and 'criterion' function is defined as 'friedman_mse'.

2.3.7. Neutral Network

Neural Networks, also known as Artificial Neural Networks (ANNs), are a versatile set of machine learning models that simulate the human brain's neural structure. These networks comprise an input layer, one or more hidden layers, and an output layer [7, 8]. Neurons within these layers analyze data, and the connections between neurons have associated weights.

In this article, the author attempts to construct a with two-hidden layer model using a three-step process to predict vaccine hesitancy. Due to the complexity of the ANNs, the author reduced the training time by randomly selecting 30,000 and 60,000 participants from the entire sample to construct the model. The initial parameter 'solver' is set to 'lbfgs', 'max_iter' is defined as 1000 and the 'activation' function is set to 'logistic'. After setting up the model, the author first determines the

optimal regularization parameter alpha while assuming 50 neurons in the first hidden layer. Next the neurons number of the first hidden layer is calculated using the optimal alpha value obtained in step one. Finally, utilizing the optimal values of alpha and the optimal number of neurons in the first hidden layer, the number of neurons in the second hidden layer is determined. The resulting model can be employed to predict the test dataset [9].

3. Results and Discussion

In the article, the author classified errors into two categories: errors in training data and errors in test data. The former reflects the model's fitness to the given data, while the latter reflects its ability to predict unknown data. Table 2 illustrates that both Random Forest and Extra-Trees exhibit the best fit for the training data, with errors below 0.01%. In contrast, for the test data, Neural Networks provide the most accurate predictions, with only a 5.63% error. Although the training only uses part of the entire sample, it still achieves favorable results. Since the quality of training of neural networks depends on the amount of input data, with an increase in the training sample, a result that is better than the previous one can be achieved. Fig. 5 also shows that the larger the sample size, the more stable the error between the training and testing data sets which shows the advantages of neural networks for handling large amounts of data. Additionally, decreasing the error could also come about by enhancing the hidden layer.

Compared to non-ensemble learning methods, ensemble methods have lower testing errors due to improved generalization. By combining the predictions of multiple base models, ensemble methods become less sensitive to variations in the training data and capture more robust patterns, resulting in better performance on unseen data and more accurate predictions [10]. However, it should be noted that Random Forest and Extra-Trees also exhibit overfitting on the training data. Compared to single decision trees, multi-trees methods can generate a significant amount of decision trees, each of which can be extremely deep and complex. If not suitably constrained, they can expand until they exactly match the training data, finally resulting in overfitting [11].

Table 2. Results of the models

Method	Sample	Train/Test	Training error	Testing error
K-nearest neighbors	107, 286	9	5.92%	6.58%
Support Vector Classifier	107, 286	9	4.66%	6.38%
Random Forest	107, 286	9	<0.01%	6.11%
Extremely Randomized Trees	107, 286	9	<0.01%	6.17%
Adaptive Boosting	107, 286	9	6.07%	6.21%
Gradient Boosting	107, 286	9	5.86%	6.09%
Neutral network	30, 000	9	6.15%	5.63%
Neutral network	60, 000	9	5.16%	5.05%

Neural networks do not have such problems. They can be effectively regularized using techniques such as dropout, weight decay (L1 and L2 regularization), early stopping, and batch normalization. These techniques help prevent overfitting by controlling the complexity of the model and encourage it to generalize better to unseen data. But as noted in the introduction to the model above, ANN is like a black box and it's hard to interpret how or why the model came up with a certain output. Unlike algorithms such as decision trees or logistic regression, vaccine acceptance cannot be predicted in advance based on individual respondents' characteristics [12]. And the lengthy computation time makes it unsuitable for CPU training.

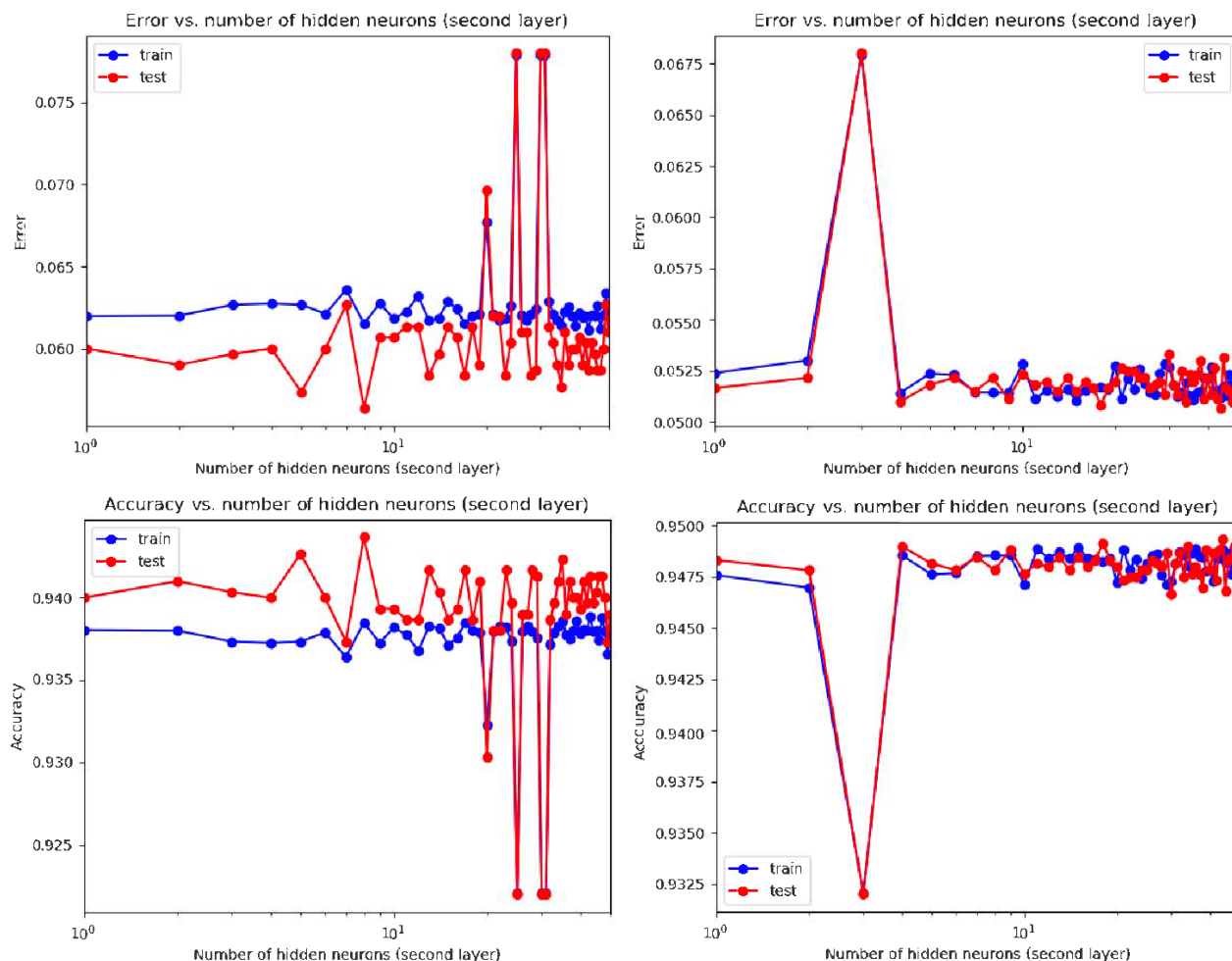


Figure 5. Error and accuracy vs. number of second layer
(left 2: 30,000 sample, right 2: 60, 000 sample) (Photo credit: Original)

4. Conclusion

The author utilized a technique for gauging feature importance to vet all questions in the questionnaire. The screened data was then subjected to non-ensemble, ensemble, and neural networks methods. KNN and SVC, the non-ensemble approaches, demonstrated the highest testing error, but their training error fell within an intermediate range. The ensemble methods, Random Forest, and Extra-Trees demonstrated overfitting in the training set. However, all ensemble methods showed smaller testing errors compared to non-ensemble methods. The neural network, a deep learning method, did not possess a significant advantage with regards to training error; however, its testing error was notably lower than the other methods, signifying that it was the best fit amongst the final models. In the future, deep learning could become more prevalent in medical diagnosis. For cases with extensive data sets, it produces better results than non-ensemble or ensemble methods. It is crucial to emphasize that the process of training neural networks is time-intensive, and in the future, there might be a need for accelerated training techniques. Employing additional feature extraction methods can enhance outcomes by independently filtering and fitting variables before incorporating them into the neural network model.

References

- [1] Feehan, Jack, and Vasso Apostolopoulos. Is COVID-19 the worst pandemic? *Maturitas* 149, 2021, 56 - 58.

- [2] Sallam, Malik. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines* 9.2, 2021, 160.
- [3] Troiano, Gianmarco, and Alessandra Nardi. Vaccine hesitancy in the era of COVID-19. *Public health* 194, 2021, 245 - 251.
- [4] Dror, Amiel A., et al. Vaccine hesitancy: the next challenge in the fight against COVID-19. *European journal of epidemiology* 35, 2020, 775 - 779.
- [5] Garg, Arunim, and Vijay Mago. Role of machine learning in medical research: A survey. *Computer science review* 40, 2021, 100370.
- [6] Venkatesh, B., and J. Anuradha. A review of feature selection and its methods. *Cybernetics and information technologies* 19.1, 2019, 3 - 26.
- [7] Yuhang, Q. et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 72, 2022, 103323.
- [8] Brause, Rüdiger W. Medical analysis and diagnosis by neural networks. *Medical Data Analysis: Second International Symposium, ISMDA 2001 Madrid, Spain, October 8 – 9, 2001 Proceedings 2*. Springer Berlin Heidelberg, 2001.
- [9] Raschka, Sebastian, and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.
- [10] Sagi, Omer, and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, 2018, e1249.
- [11] Tang, Cheng, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? *Advances in neural information processing systems* 31, 2018.
- [12] Mijwel, Maad M. Artificial neural networks advantages and disadvantages. Retrieved from LinkedIn <https://www.linkedin.com/pulse/artificial-neuralnet Work, 2018, 21>.