Efficient Flu-Shot Prediction using Data Modelling and Analysis

Aditya Kumar Sasmal
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxas24@nottingham.ac.uk

Shruthi Kogileru
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
psxsk9@nottingham.ac.uk

***Abstract*— This study aims to predict whether people will choose to get vaccinated in the future using machine learning techniques. It's crucial to predict vaccination behavior to effectively promote immunization and prevent the spread of diseases. To achieve this, various machine learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting were utilized. These algorithms analyzed a comprehensive dataset comprising demographic information, health indicators, and behavioral patterns.**

Previous studies have extensively researched the prediction of public behavior towards taking H1N1 and seasonal flu vaccinations in the future. In our report, we focused on two approaches, critiquing existing methods and developing efficient and effective machine learning models. The results demonstrated that machine learning models are effective in accurately predicting future vaccination behavior. By considering factors like demographics, health beliefs, doctor recommendations, and socioeconomic characteristics, these models provide valuable insights into individuals' likelihood of getting vaccinated.

The use of machine learning techniques offers valuable insights into predicting future vaccination behavior for H1N1 and seasonal flu. These insights can guide proactive strategies to increase immunization rates, improve public health outcomes, and shape targeted interventions tailored to specific populations.

Keywords—Decision Tree classifier, Gaussian Naïve Bayes, Machine learning models, AUC, Hyperparameter tuning.

I. INTRODUCTION

In recent years, predicting individuals' vaccination behaviour has become an important area of research in public health. The ability to accurately determine the likelihood of individuals receiving H1N1 and seasonal flu vaccines can greatly assist in planning and implementing effective vaccination campaigns. Machine learning techniques have proven to be highly effective in analysing large datasets and making accurate predictions for complex outcomes, making them an increasingly popular choice in data analysis. This project aims to work on critical thinking. We have developed machine learning algorithms to predict the probability of individuals receiving their H1N1 and seasonal flu vaccines using data collected from the National 2009 H1N1 Flu survey.

DATASET

The dataset used for this study was obtained from a telephone survey conducted on over 26,000 individuals residing in the United States. The survey aimed to investigate influenza immunization coverage during the 2009-2010 season. The dataset comprises of 36 columns. The first column, ‘respondent_id’, is the unique identifier. The remaining 35 features consist of demographic features such as age group, education, race, sex, etc, and features relating to the people’s concern and knowledge about the H1N1 flu, effectiveness of the vaccines, worry about the after-effects of the vaccines, etc.

The goal of this report is to predict the chances of people receiving H1N1 and seasonal flu vaccines. To find the two requisite probabilities, there are two target variables: one for the H1N1 vaccine and one for the seasonal flu vaccine. These variables are binary, with 0 indicating no and 1 indicating yes. As a result, this is a multilabel problem.

Several machine learning algorithms predicted which individuals would receive the vaccinations. Among the models tested, the Gradient Boosting Classifier performed exceptionally well. Because of the model's ensemble structure, it was able to correct errors caused by previous models, resulting in highly accurate forecasts. This model was beneficial in finding crucial factors influencing vaccination acceptability.

A few research questions to keep in mind while modelling are -

- How important do people think it is to take the vaccine for both H1N1 and seasonal flu to avoid being infected?
- Do people believe in the effectiveness of the vaccines?
- What are some of the people’s behaviours that impacted their decision of taking the vaccines?
- What are some of the demographic information related to their status of taking the vaccines? • What are the demographics of people - age-group, education, race, sex, marital status?
- Have people avoided contact with people exhibiting flu-like symptoms?

LITERATURE REVIEW

Machine Learning (ML) models have shown their expertise in characterizing the hidden patterns of data and therefore, have been employed in various complex classification tasks [1] Classification Problems: Data sets that contain fixed output (yes/no or true/false).[2] H1N1 according to virology is a subtype of Influenza. It is a virus that is also written as (A/H1N1). The influenza ‘A’ virus was also a common reason for Influenza (flu) in 2009- 2010 and is associated with the great plague of the 20th century that devastated Spain (1918-1920). 2009 Health.[3]With 50 million cases in the United States and 812,000 deaths as of December 7, 2021, the

need for COVID 19 vaccination is evident.[3] This shows the importance of taking vaccines. The analysis showed that the perception of the vaccine as an effective strategy to prevent both levels of risk and vaccine follow-up posed by the 2009 pandemic flu outbreak has increased. [4] The missing values must be handled as they reduce the quality of any of our performance metrics. Unless the data is pre-processed to the extent that an analyst will encounter nonessential values such as NaN it can also lead to incorrect prediction or classification and can also cause a high bias for any given model. Missing values can be represented as a question mark (?) or a zero (0) or minus one (-1) or as a blank. Fillna() function is used to fill NaN values.[4] Gradient boosting algorithms just like random forests are supervised machine learning algorithms used for classification and regression. It is an ensemble learning model Tree like algorithms are widely used. This method comes with high accuracy and accuracy.[4] Accuracy can be increased by many units hidden within a level with regularization techniques using Hyper parameter tuning. Hyper parameter Tuning is done to find the most optimal parameter for the model on which the model gives the best results. Various Hyperparameter tuning methods such as GridseacrchCV, RandomSearchCV for machine learning models to obtain better results. K fold cross Validation method has is used to tune hyperparameters for the Artificial Neural Network.[2], [4]

[2]The study includes 1467 patient data (70% from H1N1 and 30% from COVID-19) with 42 attributes used in classification. Experimental results show that the Bayes network gives 86.57% accuracy, the naive Bayes classifier gives 82.34% accuracy, the multilayer perception algorithm gives 99.31% accuracy, the locally-weighted learning algorithm gives 88.89% accuracy, and random forest gives 83.16% accuracy for the same data set.

Modern technologies cannot completely replace traditional epidemiological monitoring networks; they can only play a complementary role. In this study, a combination of CDC and GFT data is used to construct a novel combination forecasting model.[5]

OBJECTIVE

Our primary goal in this project was to create machine learning models for classifying the 2009 H1N1 flu survey dataset obtained from a Kaggle competition. To foster critical thinking, we took distinct approaches by exploring two different techniques at each stage of the project.

Our project is focused on identifying the optimal approach between two methodologies for constructing a machine learning model that can accurately predict an individual's likelihood to receive vaccinations in the future.

II. METHODOLOGY

We began by thoroughly examining the dataset and preparing the data for modelling, ensuring its reliability and appropriateness. Subsequently, we developed machine learning models for classification. By comparing and analysing the results obtained from both approaches, we aim to gain valuable insights and determine the most effective machine learning models and techniques suitable for the given dataset. We now discuss the methodologies used for each step in detail.

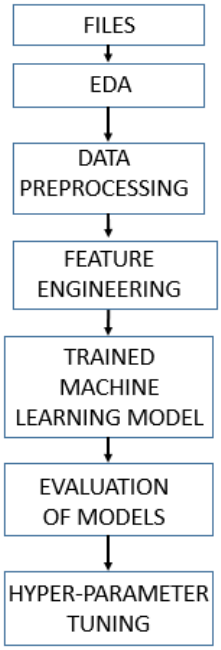

Figure 1: Flow of the project

EXPLORATORY DATA ANALYSIS

We first import the necessary libraries. Then, after loading the data, we employed various data visualisation techniques to gain insights on the dataset. We have used plotly from matplotlib and seaborn libraries for our visualisations. These visualisations enabled us to get a comprehensive understanding of vaccination counts for H1N1 vaccine and seasonal vaccine respectively, explore the relationship between geographic regions and vaccination and aided us to analyse the distribution patterns of different categorical variables in the dataset.

Moving forward, we adopted two distinct approaches to developing the machine learning models, which will be discussed further. These two approaches will be addressed as approach 1 and approach 2.

Approach 1 – Adopted by Aditya Kumar Sasmal

Approach 2 – Adopted by Shruthi Kogileru

DATA PRE-PROCESSING

In our project, both approaches dealt with missing values by removing columns that have a large percentage of missing data and filling them with alternate values. Approach 1 involved inserting the median value of the column into the missing values of the numerical columns and the mode value of the column into the missing values of the categorical columns. Approach 2 involved inserting the rounded mean value of the column into the missing values of the numerical columns and 'None' into the missing values of the categorical columns. Both approaches chose label encoding to convert categorical columns to numeric columns.

FEAUTURE ENGINEERING

The feature engineering steps adopted in approach 1(Aditya) are -

1. A new feature called "hygiene" was created by combining columns related to personal hygiene practices.
   The "hygiene" feature was plotted using a count plot to visualize the distribution of hygiene practices.
   The "hygiene" feature was then analyzed against the target variables ("h1n1_vaccine" and "seasonal vaccine") using bar plots to observe the relationships.
2. Another feature called "opinion" was created by combining columns related to people's opinions on vaccines.
   The "opinion" feature was plotted using a count plot to visualize the distribution of opinions.
   The "opinion" feature was analyzed against the target variables using bar plots to examine the relationships.
3. Separate features were created for opinions related to the h1n1 vaccine and the seasonal vaccine.
4. The correlation between different columns and the "h1n1_vaccine" and "seasonal vaccine" columns was computed and displayed.
5. Three addition features were created –
   - 'concerned' (1 if the corresponding h1n1_concern value >2, else 0)
   - 'danger' (1 if the corresponding 'opinion_h1n1_risk' >3, else 0)
   - 'well_aware_h1n1' (1 if the corresponding 'h1n1_knowledge' >2, else 0)
6. After all these steps, all the redundant columns were dropped.
7. After feature engineering, there remained 19 columns and 26,707 rows.

The feature engineering steps adopted in approach 2 (Shruthi) are -

1. The correlation matrix of the dataset was visualized using a heatmap.
   The correlation of each column with the "h1n1_vaccine" and "seasonal_vaccine" columns was calculated and analyzed.
2. The columns with negative or almost 0 correlation to with the target variables were dropped from the dataset.
3. Z-scores were calculated for each numeric column to detect outliers, and the rows with outlier values were removed.
4. After feature engineering, there remained 27 columns and 22,943 rows.

IMPLEMENTING MACHINE LEARNING MODELS

We have implemented one common machine learning model (Gradient Boosting) to compare the two approaches. K-fold cross validation with 10 splits has been used to split the data.
The machine learning models implemented in approach 1 (Aditya) are:

- Gradient Boosting (Ensemble model) – It creates a strong predictive model by combining multiple weak learners. It can handle non-linear relationships, noisy observations.
- Bernoulli Naïve Bayes (Naïve Bayes model) - It assumes that the features follow a Bernoulli distribution and works well for binary classification.
- Logistic Regression (Generelised linear model)– It is reliable and easy to interpret and is well suited for a binary classification problem.
- AdaBoost (Ensemble model) – It creates a strong predictive model by combining multiple weak learners. It effectively handles strong relationships and improves the accuracy of the predictions.

The machine learning models implemented in approach 2 (Shruthi) are:

- Gradient Boosting (Ensemble model) – It creates a strong predictive model by combining multiple weak learners. It can handle non-linear relationships, noisy observations.[6]
- K-Nearest Neighbours (Nearest Neighbour model) – It classifies data based on its neighbors and is effective when instances with similar features tend to have the same class, as in the vaccination dataset in concern. [7]
- Random Forest (Ensemble model) – It effectively handles the dataset complexity by building an ensemble of decision trees as well as aggregating predictions. This approach generates robust and accurate classifiers, successfully predicting individuals' flu vaccination decisions.
- XGBoost classifier (Gradient Boosting model) – It is a powerful gradient boosting algorithm and predicts with a high accuracy. It handles non-linear relationships and complex interactions well.

HYPERPARAMETER OPTIMISATION

Hyperparameter optimisation is done to select the optimal values of hyper parameters for the machine learning model.

In our project we have used GridSearchCV and RandomizedSearchCV for hyper parameter optimisation. GridSearchCV provides an efficient evaluation of all possible combinations within the well-defined search space and provides interpretable results. RandomizedSearchCV has freedom to explore different combinations and has the chances of finding unexpected hyperparameters that might lead to a better model.

Process	Approach 1 (Aditya)	Approach 2 (Shruthi)
EDA	Provides visualisation	
Data pre-processing	- Imputing with median (for numeric columns) and mode (for categorical columns). - Label encoder used	- Imputing with rounded mean (for numeric columns) and 'None' (for

		categorical columns). - Label encoder used.
Feature engineering	- Creating five new columns: ‘hygiene’, ‘opinion’, ‘danger’, ‘concerned’, ‘well_aware_h1n1’ - Dropping all redundant columns	- Dropping columns with negative or 0 correlation - Dropping rows with outliers
Machine Learning Models	- Gradient Boosting - Bernoulli Naïve Bayes - Logistic Regression - AdaBoost	- Gradient Boosting - k-Nearest Neighbours - Random Forest - XGBoost
Hyperparameter optimisation	- GridSearchCV - RandomizedSearchCV	

Table 1: The different approaches to solve the problem in concern.

III. RESULTS

After data pre-processing and feature engineering, Aditya’s dataframe had 26707 rows and 19 columns, whereas Shruthi’s dataframe had 22,943 rows and 27 columns.

We took two distinct approaches and ran different models. After exploring a number of models using both the approaches, and accuracy and ROC AUC curve were used to evaluate the results.

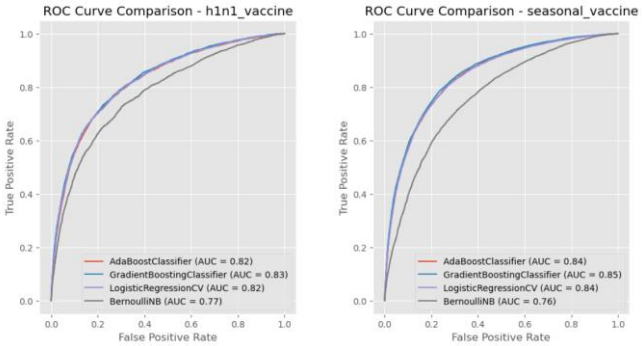

Fig 2: ROC AUC curve of Aditya Kumar Sasmal’s approach

Model Comparison:

	Model_Name	Test Accuracy (h1n1)	Test Accuracy (seasonal)
0	AdaBoostClassifier	0.828959	0.769087
1	GradientBoostingClassifier	0.835399	0.77328
2	LogisticRegressionCV	0.831168	0.768525
3	BernoulliNB	0.806268	0.705096

	Mean Accuracy	AUC
0	0.799023	0.843554
1	0.804340	0.849183
2	0.799847	0.842256
3	0.755682	0.764688

Table 2: Accuracy and AUC score of Aditya’s approach

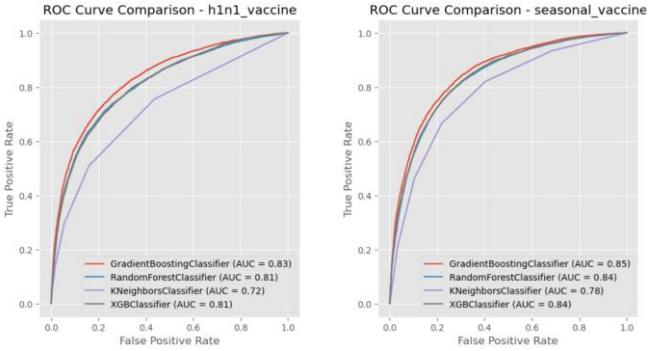

Fig 3: ROC AUC curve of Shruthi Kogileru’s approach

Model Comparison:

	Model_Name	Test Accuracy (h1n1)	Test Accuracy (seasonal)
0	GradientBoostingClassifier	0.846969	0.77745
1	RandomForestClassifier	0.840911	0.765376
2	KNeighborsClassifier	0.816676	0.728066
3	XGBClassifier	0.836464	0.765115

	Mean Accuracy	AUC
0	0.812209	0.851488
1	0.803143	0.835085
2	0.772371	0.780898
3	0.800789	0.837396

Table 3: Accuracy and AUC score of Shruthi’s approach

- We have also used the data pre-processed by Aditya and have applied Shruthi’s feature engineering approach to it. We have used this data on all models and have evaluated the results.

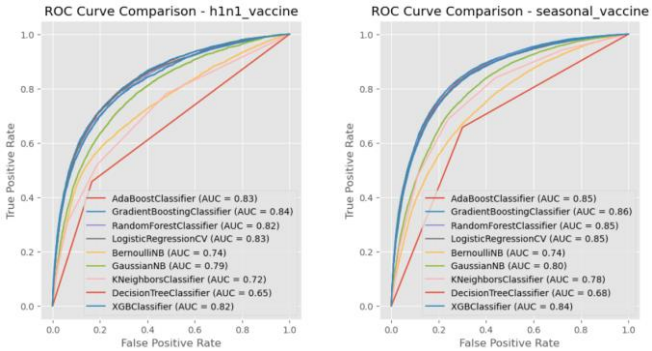

Fig 4: ROC AUC curve using the combination of the approaches used.

Model Comparison:

	Model_Name	Test Accuracy (h1n1)	Test Accuracy (seasonal)
0	AdaBoostClassifier	0.834800	0.776838
1	GradientBoostingClassifier	0.838207	0.780282
2	RandomForestClassifier	0.835324	0.773055
3	LogisticRegressionCV	0.835025	0.771895
4	BernoulliNB	0.801325	0.688396
5	GaussianNB	0.770884	0.734414
6	KNeighborsClassifier	0.802673	0.726327
7	DecisionTreeClassifier	0.753473	0.67956
8	XGBClassifier	0.832853	0.772232

	Mean Accuracy	AUC
0	0.805819	0.850750
1	0.809245	0.855184
2	0.804190	0.847511
3	0.803460	0.845577
4	0.744861	0.744557
5	0.752649	0.800910
6	0.764500	0.779478
7	0.716516	0.678440
8	0.802542	0.844774

Table 4: Accuracy and AUC score for the combination of the approaches used.

We have also performed hyperparameter optimisation and got the following results –

```
Performing hyperparameter optimization for GradientBoostingClassifier
Best parameters: {'learning_rate': 0.1, 'n_estimators': 200}
Best ROC AUC score: 0.8325244482785361
Cross-validated scores: [0.83260358 0.834483   0.8380816  0.84224877 0.8152053

Performing hyperparameter optimization for RandomForestClassifier
Best parameters: {'max_depth': 10, 'n_estimators': 200}
Best ROC AUC score: 0.8285699001469229
Cross-validated scores: [0.82675554 0.82722689 0.83651303 0.83754866 0.8139465
```

Fig 5: Results of HPO using GradientSearchCV

```
Performing hyperparameter optimization for GradientBoostingClassifier
Best parameters: {'n_estimators': 50, 'learning_rate': 0.1}
Best ROC AUC score: 0.829460412153329
Cross-validated scores: [0.82979451 0.82952922 0.83725326 0.83867913 0.81204594

Performing hyperparameter optimization for RandomForestClassifier
Best parameters: {'n_estimators': 50, 'max_depth': 10}
Best ROC AUC score: 0.82615903195786
Cross-validated scores: [0.82406749 0.82372961 0.83540328 0.83676171 0.80906194
```

Fig 6: Results of HPO using RandomizedSearchCV

We also performed feature engineering on our model.

Feature Importance:

	Feature	Importance
8	doctor_recc_h1n1	0.429721
13	opinion_h1n1_vacc_effective	0.075336
14	opinion_h1n1_risk	0.049524
12	health_worker	0.048318
17	opinion_seas_risk	0.028755
9	doctor_recc_seasonal	0.027553
18	age_group	0.022858
0	h1n1_concern	0.022809
1	h1n1_knowledge	0.021501
15	opinion_h1n1_sick_from_vacc	0.021054
19	race	0.020972
21	income_poverty	0.020123
10	chronic_med_condition	0.020084
7	behavioral_touch_face	0.019755
2	behavioral_avoidance	0.019657
16	opinion_seas_vacc_effective	0.019591
20	sex	0.019527
5	behavioral_large_gatherings	0.019329
6	behavioral_outside_home	0.019069
23	rent_or_own	0.018913
4	behavioral_wash_hands	0.018726
22	marital_status	0.018639
24	employment_status	0.018186
3	behavioral_face_mask	0.000000
11	child_under_6_months	0.000000

Fig 7: Feature importance results

IV.DISCUSSION

While comparing the approaches, we made the following observations –

- While Aditya concentrated more on dealing with the number of columns and finding the best combination of columns to use in the model, Shruthi concentrated more on the outlier rows and correlated columns.
- Advantages of Aditya's approach – It reduces the number of columns and hence reduces the processing time as well.
  Thorough investigation of the columns and their uses has been done, and the relevant columns have been created.
- Disadvantages of Aditya's approach – The basis on which the columns have been removed/altered does not have systematic reasoning and is based on general observation and knowledge.
- Advantages of Shruthi's approach –
  More number of columns have not been deleted, hence there is lesser chance for the loss of important information without concrete reasoning. Negatively correlated columns are removed.
- Disadvantages of Shruthi's approach – While outliers may reduce efficiency, many models are immune to it, and thus this step might not have a significant impact on the results.
  We can see from the feature importance computation that a few columns have no importance and can be dropped to improve the accuracy.
- By exploring the various methods and also the combinations, we have got the best results.
- While comparing to the approaches taken in the literature review, we can confirm that the ensemble methods have the highest accuracy.
  But the distance based algorithms, such as kNNs do not have high accuracy and are not suitable for problems such as the one in discussion.

We can observe from the above graphs and tables that –

- Both Aditya's and Shruthi's approaches worked best for the Gradient Boosting algorithm.
- We can see that the accuracy is generally higher for the h1n1_vaccine target variable as compared to the seasonal_vaccine target variable.
- In Aditya's approach, Bernoulli's Naïve Bayes has gotten the lowest accuracy out of the compared models. This is due to the following reasons –
  - Bernoulli's Naïve Bayes assumes a Bernoulli's distribution, which may not hold true always.
  - It assumes independence between features, which may limit its ability in the case of complex interactions.
  - It assumes binary features and can be affected using thresholds.
- In Shruthi's approach kNN has given the lowest accuracy of all the compared models. This is because –
  - kNN is sensitive to the scale of features since it depends on distance calculations for classification.
  - Hence it may struggle to capture complex relationship and non-linear decision boundaries present in the data.
- We can see that the AUC score after hyperparameter optimsation has not increased. This might be because the range of the hyperparameters explored during the tuning process might not be inclusive of the optimal values.
- The feature importance gives several features that are not important and hence can be dropped to improve the accuracy of the model.

We have answered several research questions –

- We can see that the number of people taking the flu shot is significantly higher than the ones who have taken the H1N1 vaccine.
- Doctor recommendation has a huge impact on the people's decision in taking the vaccine.
- We can see that the dataset has data relating majorly to the people of the white race.
- The dataset has more number of college students and less number of <12 children's data.

V. CONCLUSION

Logistic Regression, AdaBoost, Gradient Boosting, Bernoulli NB, K Nearest Neighbors, XGBoost and Random Forest - These models brought their unique strengths to the analysis, such as interpretable coefficients and probabilities (LogisticRegressionCV), handling of categorical and continuous data (BernoulliNB), leveraging instance similarity (KNeighborsClassifier), facilitating feature selection and rule extraction (RandomForestClassifier), and delivering high performance and scalability (XGBClassifier).

This collective utilization of diverse models provided comprehensive insights into the factors influencing vaccine acceptance and guided the development of targeted interventions and public health strategies to improve vaccination coverage.

In conclusion, this project aimed to develop efficient machine learning models and foster critical analytical thinking in predicting individuals' likelihood of taking vaccinations in the future. The process involved several crucial steps, including data cleaning, pre-processing, feature engineering, exploratory data analysis, machine learning, and evaluation.

The initial dataset, sourced from the National 2009 H1N1 Flu survey, consisted of responses from over 26,000 individuals residing in the United States. Our pre-processing efforts involved handling null data and converting categorical data into numerical format using a label encoder. Additionally, we implemented feature engineering techniques to create new columns, enhancing the efficiency of the machine learning models.

Extensive exploratory data analysis allowed us to uncover valuable insights about the dataset, shedding light on patterns and relationships within the data. KFold Cross validation has been done with 10 splits to ensure accurate model evaluation.

By implementing various machine learning models and evaluating their performance, we were able to make predictions regarding individuals' vaccination behavior. Through this project, we not only developed efficient machine learning models but also cultivated critical analytical thinking skills by questioning assumptions and making informed decisions based on the outcomes.

As improvements, Neural Networks can be analysed to work with this problem and also with similar ones.[8] Doctor recommendation has high importance and should be pushed more, whereas features with low importance should be excluded from the analysis.

Contributions

To the paper –

Shruthi – Results, Discussion

Aditya – Introduction, Conclusion

REFERENCES

[1] H. Gupta and O. P. Verma, 'Vaccine hesitancy in the post-vaccination COVID-19 era: a machine learning and statistical analysis driven study', *Evol. Intell.*, vol. 16, no. 3, pp. 739–757, Jun. 2023, doi: 10.1007/s12065-022-00704-3.

[2] E. Elbasi, A. Zreikat, S. Mathew, and A. E. Topcu, 'Classification of influenza H1N1 and COVID-19 patient data using machine learning', in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2021, pp. 278–282. doi: 10.1109/TSP52935.2021.9522591.

[3] P. Nair and D. P. Wales, 'Seasonal and 2009 Pandemic H1N1 Vaccine Acceptance as a Predictor for COVID-19 Vaccine Acceptance', *Cureus*, vol. 14, no. 1, 2022.

[4] 'Predicting H1N1 and Seasonal Flu : Vaccine Cases using Ensemble Learning approach | IEEE Conference Publication | IEEE Xplore'. https://ieeexplore.ieee.org/abstract/document/9362909?casa_token=kwPMEy20C38AAAAA:wMMkjviS29t5LlVFVmoMNsiyp97Ow-oPLcVLYSiRATzJ2XixbGpqTeRzhGAvMGqXT9vO0hVpjQ (accessed May 19, 2023).

[5] 'A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients - ScienceDirect'. https://www.sciencedirect.com/science/article/pii/S0957417420304851?via%3Dihub (accessed May 19, 2023).

[6] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, 'A comparative analysis of gradient boosting algorithms', *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[7] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, 'KNN Model-Based Approach in Classification', in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, R. Meersman, Z. Tari, and D. C. Schmidt, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.

[8] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, 'Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction', *Complex Intell. Syst.*, vol. 8, no. 4, pp. 3073–3087, Aug. 2022, doi: 10.1007/s40747-021-00398-7.