Comparing Machine Learning Based Models: H1N1 and Seasonal Flu Vaccine Prediction

Clarine Tan Kaili
School of Computer Science
University of Nottingham
Nottingham, UK
hfyct2@nottingham.ac.uk

Carolyn Han En Qi
School of Computer Science
University of Nottingham
Nottingham, UK
hfych2@nottingham.ac.uk

***Abstract* — This paper investigates the factors influencing vaccination decisions, behaviour, and the likelihood of individuals receiving the H1N1 and seasonal flu vaccinations. EDA was performed to determine the data distribution of numerical features and to determine the top 10 features that have strong associations with the different vaccination uptake using the Chi-Square test. The pre-processing techniques used in this paper include dropping missing values, simple imputation (mean, median and mode), one-hot encoding, correcting data types, random oversampling, and SMOTEENN. Furthermore, two classification models, namely Random Forest and CatBoost, were used and compared using multiple evaluation metrics such as accuracy, hamming loss, F1 score (micro and weighted), and AUROC. The dataset used in this paper consists of responses from a phone call survey conducted by the NHFS in the U.S. and multilabel and imbalanced. This paper discusses several factors that could have influenced the population's vaccination behaviour. Besides that, the Random Forest outperforms CatBoost, while using mean and median imputation techniques did not have a notable impact on the model's performance. Random oversampling was more effective for Random Forest, whereas SMOTEENN resampling was better for CatBoost. The findings of this paper can be used to improve existing interventions aimed to increase the vaccination rate. Besides that, an accurate vaccination prediction rate can be determined and aid in decision-making regarding vaccine importation and manufacturing.**

Keywords* — *Random Forest, CatBoost, H1N1 vaccine, seasonal flu vaccine, data analysis, data pre-processing

I. Introduction

A. *Description of Dataset*

The dataset used in this paper is from the DrivenData competition "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines". The data was collected from a phone survey conducted by the National 2009 H1N1 Flu Survey (NHFS) in the US. This survey has collected social, economic, and demographic information from the respondents. The training dataset consists of 26,707 entries and 35 columns, excluding the index column. Each entry represents a respondent, and each column represents a feature. The index column, '*respondent_id'*, is a unique random identifier for each respondent. The dataset poses a multi-label problem to predict the probability of an individual receiving the H1N1 and seasonal flu vaccine. The training label dataset contains the labelled outcome for each respondent according to their ID. A test set is provided as part of the ongoing competition, however, no labelled outcome is available for the test set. Therefore, the test set will not be used to evaluate the models' performance in this paper. Instead, a train-test split dataset will be used. The dataset consists of 35 features, of which the *'respondent_id'* is discrete data. Only two are of discrete numerical data type: *'household_adults'* and *'household_children'*. These two features are top coded to 3. Thus, they have the same range and data normalisation is not required, especially when using ensemble algorithms. The other 33 features are categorical, more specifically, 15 are binary (including the 2 dichotomous features), 8 are ordinal, and 10 are nominal. No duplicates were found in the dataset. However, approximately 76% of the entries (20,270 instances) contain missing data, with 206 rows having at least 50% of their data missing. As shown in Fig. 1, only 6 features have complete data with *'employment_occupation'*, *'employment_industry'*, and *'health_insurrance'* having the highest percentage of missing values nearing 50%. As shown in the histogram in Fig. 2, the distribution of respondents who received the H1N1 and seasonal flu vaccines is imbalanced. Specifically, the distribution of the H1N1 cases is heavily skewed towards the negative instances, with 78.75% of the respondents reporting that they did not receive the H1N1 vaccine. In addition, 53.44% of the respondents did not receive the seasonal flu vaccine.

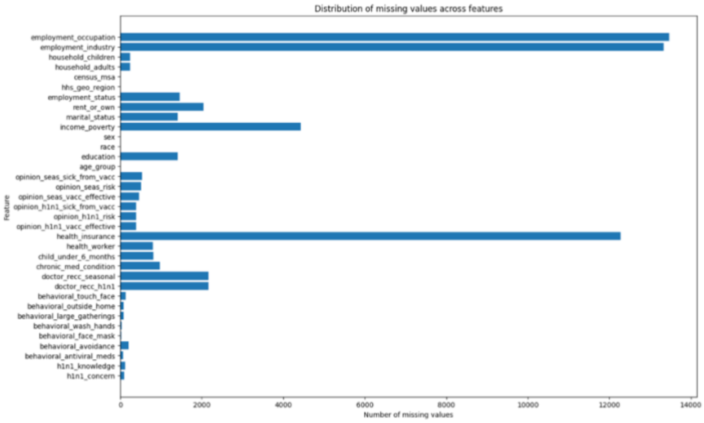

Fig. 1. Distribution of missing values across features

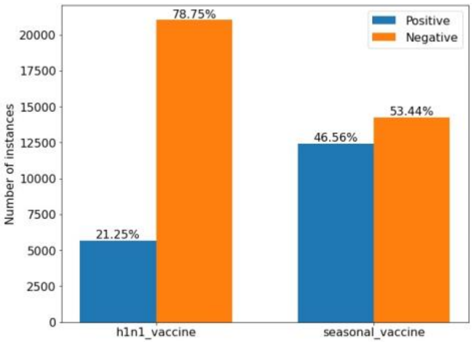

Fig. 2. Uptake of H1N1 and Seasonal vaccine

B. *Research Questions*

This paper aims to investigate and answer the following questions about vaccination behaviour: (1) What factors influence the population’s decision to receive vaccines? (2) What is the vaccination behaviour of the population? (3) What is the best machine learning model to predict the chances of a person receiving the H1N1 and seasonal flu vaccinations?

Question (1) can help authorities increase the rate of vaccination amongst the public through targeted intervention. Question (2) can shed light on the public’s vaccine awareness and acceptance that can be used to evaluate the effectiveness of their current interventions and campaigns. Lastly, question (3) can help healthcare professionals and relevant authorities can estimate the number of vaccines that should be manufactured and distributed by addressing.

II. Literature Review

The National 2009 Flu Survey (NHFS) was initiated by the Centres for Disease Control and Prevention (CDC) to monitor the coverage of 2009 influenza A (H1N1) and seasonal influenza vaccines [1]. A study was conducted to investigate the socio-demographic differences in opinions regarding the H1N1 and seasonal influenza vaccination and disease [1]. The researchers analysed the NHFS data and discovered that the opinions of the U.S. population differed by race, income, and educational level [1]. Further analysis using multivariable logistic regression revealed a strong association between H1N1 vaccination and both the perceived effectiveness of the vaccine and personal risk of contracting the disease [1].

Vaccination has been proven to be effective in mitigating the consequences of an influenza pandemic [2]. However, the actual effectiveness of the vaccination response heavily relies on the on-demand supply of vaccines [2]. It has been demonstrated that high-income country governments had advanced-purchase agreements resulting in high proportions of expected pandemic vaccines being reserved [2]. This result in high costs in storing large batches of vaccines in a safe environment and over-reserving vaccines leads to prolonged unavailability of such vaccines for developing countries. A study on the H1N1 vaccination program during the 2009 influenza season in the U.S. highlighted that the availability of vaccines was limited during the early stages of the epidemic when the impact of vaccination could have been the greatest [3]. Thus, this stresses the need to develop accurate prediction models for vaccination rates.

Although H1N1 was declared the first global influenza pandemic by the World Health Organisation (WHO) in over 40 years and was regarded as more severe than seasonal influenza, the vaccination rate among U.S. adults for the H1N1 vaccine was significantly lower compared to the seasonal vaccine [4]. This can be attributed to the common perception that the pandemic vaccine was less safe [4]. Among adults, those who relied on healthcare providers as their main source of information achieved the highest vaccination rate for both vaccines [4].

The objective of multilabel classification is to build a model that predicts a set of relevant labels for unseen data, where the labels may be reliant on each other [5, 6]. In a comparison of multilabel classification methods, it was discovered that having decision trees as the base classifier achieved the best performance, while ensemble methods had the highest overall performance [6]. This led to the idea of comparing tree-based ensemble models where an ensemble of *n* trees is built. Tree-based methods demonstrate the benefit of sub-linear to logarithmic prediction costs when the tree is balanced [7].

A study conducted on the same dataset investigated the performance of ensemble-based methods in predicting the vaccination probability of individuals [8]. The study revealed that the *CatBoost* model achieved the highest accuracy, while *Random Forest* was among the least-performing models [8]. It was suggested that this result might be caused by parameter tuning. This finding encourages further exploration of different methods and pipelines to fine-tune models and produce a more comprehensive comparison. Another study comparing machine learning models for predicting the probability of vaccination on the NFHS 2009 dataset [9] also demonstrated that the models' performance, including *Random Forest*, is sensitive to parameter settings and directly influences its prediction accuracy.

III. Methodology

A. Exploratory Data Analysis

1) Data Distribution of Numerical Features

The data distribution of the numerical (integer) features is visualised using histograms on the positive cases. This can provide insight into the influence of numerical features towards vaccination rates and justify the use and comparison of the mean and median imputation methods.

2) Chi-Square Test

The Chi-Square test is a statistical test to measure the association between categorical features by calculating a Chi-Square statistic value and a corresponding p-value. The p-value acts as an indicator of the measure of the association between two variables that is used as a deciding factor for rejecting the null hypothesis that assumes no association is between the variables. A smaller p-value indicates stronger evidence against the null hypothesis. This means that there is a relationship between the features. On the contrary, the Chi-Square statistic value is used to determine the significance of the relationship. A higher Chi-Square statistic value suggests a stronger association between the variables. It is crucial to use both the p-value and the Chi-Square statistic value to conclude the features. In this paper, the Chi-Square test is used to gain insight and understand the population’s vaccination behaviour. As the dataset consists of mainly categorical features, alternative correlation measures such as Pearson’s correlation cannot be utilised. Thus, the Chi-Square test is a suitable method to observe the relationship between the features in our dataset.

B. Data Pre-Processing

1) Dealing with Missing Values

Drop missing values. Rows with more than half of their features missing are removed. Deleting data with missing values is also known as *discarding instances* [10], while the drop of such rows is known as *list-wise* or *case deletion* [11]. However, discarding data can introduce bias into the results, especially when the frequency of data deletion is high. As the dataset is collected via a phone survey, it is reasonable to assume that the missing values are a result of denial in answering certain questions. To answer the research question, it was deemed necessary to exclude cases where respondents did not answer at least half of the survey.

These cases were the minority within the dataset, with only 206 entries out of the total 26,707 entries and are not representative of the entire population. In other words, it is a form of noise in the dataset and will not provide information gain towards the objectives of this paper.

Data imputation. Imputation techniques use estimated values computed based on the known feature values to then replace missing values [12]. In this paper, three types of imputation approaches are used: mean, median, and mode imputation. Mean imputation replaces the missing data with the computed mean of its corresponding collective; similarly, median imputation replaces the missing data with the computed median value of its corresponding collective. On the other hand, mode imputation uses the most frequent value in the corresponding feature to substitute the missing value. In this paper, mean and median imputation is the different pre-processing methods used and will be compared based on the performance of the models utilising these imputation techniques. These two imputation methods are applied only to the numerical features. To apply mean or median imputation to categorical features, the computed mean or median value may need to be rounded to the nearest integer. However, the choice of rounding up or down can directly influence the integrity of the dataset. Mode imputation is applied to all categorical features, including binary features. The reason to use mode imputation is primarily because the dataset is sufficiently large in comparison to its missing data.

2) *One-Hot Encoding*

One-hot encoding converts categorical features into numerical data, which is essential because most machine learning algorithms are developed to process numerical data. This creates new binary columns for each category. Label encoding was considered as an alternative. However, it can introduce bias into the model as an implicit order among the categories is created. This does not happen with one-hot encoding, but it increases the dataset dimensionality. Despite that, studies have shown that one-hot encoding achieved high accuracy for categorical datasets [13]. The dataset used in this study has only 35 features, which is relatively small and thus one-hot encoding was deemed the most appropriate choice.

3) *Correcting Data Types*

During exploratory data analysis, the dataset was found to have mismatched data types. Thus, the features were categorised into four groups (numerical, binary, ordinal, and nominal) before changing their data type. This increases the code readability and makes it more convenient for further data analysis. The nominal features were converted to *'int8'* to reduce memory usage, while binary and nominal features were corrected to have a *'bool'* data type. The ordinal features, which came label encoded, were changed to *'int'* data type to preserve their numerical representation.

4) *Resampling Techniques*

Random Oversampling. This technique reduces the potential bias towards the majority class by randomly duplicating instances from the minority class. This approach was chosen in this paper due to its simplicity and avoiding information loss from undersampling techniques. However, it may lead to overfitting.

Synthetic Minority Over-sampling Technique Edited Nearest Neighbors (SMOTEENN). SMOTEENN is a hybrid resampling technique that combines oversampling (SMOTE) and undersampling (ENN). SMOTE will increase the number of minority class instances, while ENN will reduce the overall dataset size by removing potential noise. This technique is used to overcome overfitting problems that arise from oversampling.

Both techniques improve the classification model accuracy by adding more entries of the minority class, which is the positive cases of H1N1 and seasonal vaccine uptake. Random oversampling is preferred when the class imbalance is not severe as SMOTEENN can increase the complexity of the dataset; whereas SMOTEENN is more appropriate when the dataset contains outliers. Research has suggested that as the level of noise in the dataset increases, the better the performance of the model when SMOTEENN resampling was used [14]. Thus, these two resampling techniques will be evaluated in this paper.

C. *Classification Model*

1) *Random Forest (RF)*

RF is an ensemble method that uses a majority voting system on the collection of decision trees it constructs. Each independent decision tree is built using a randomly selected subset of attributes while maintaining a consistent splitting rule [15]. The generalisation error of RF depends on the correlation between the trees in the forest and the performance of individual trees [16]. Increasing the number of individual trees can increase its classification forest until a pre-determined limit based on the underlying dataset [17], but there is a risk of overfitting if the forest becomes too large [18]. RF is known to handle large datasets and imbalanced classes in datasets [17]. It preserves the fine characteristics of tree-based models while providing higher prediction accuracy in classification problems [19].

The RF model is the *RandomForestClassifier* from the *sklearn.ensemble* library. It was noted that the rate of convergence could be manipulated by hyperparameter tuning [18], i.e., improving the model accuracy. Thus, the *GridSearchCV* function is used to find the optimal parameters combination that promises the best F1-micro score. This is because F1-micro is commonly used in multilabel classification problems, achieving a balance between micro recall and micro precision [20].

Due to time constraints, only four hyperparameters were tuned: *'criterion'*, *'max_depth'*, *'max_features'*, and *'n_estimators'*. *'criterion'* determines the split rule, and two values were considered: *'gini'* representing the gini impurity and *'entropy'* representing the impurity measure from a split. *'max_depth'* controls the maximum depth of each tree, with values 20 and 40 considered to investigate the effect of doubling the maximum depth while to possibly avoid overfitting as the default value is *'None'*, i.e., no limit on the number of splits in a tree, and the larger the depth the higher the variance and thus the higher the probability of overfitting the model. Next, *'max_features'* controls the number of features used in deciding the best split of a tree, and only *'sqrt'* and *'log2'* were considered. This hyperparameter was tuned due to its sensitivity towards overfitting. Lastly, *'n_estimators'* defines the number of constructible individual trees in a forest. The larger the forest, the higher the complexity. As mentioned above, increasing the number of trees does not increase the prediction accuracy once a certain limit is reached, which encourages the inclusion of this hyperparameter in the grid search. Only 2 values (400 and 500) were considered.

2) CatBoost

CatBoost was selected as a comparative algorithm as previous studies showed its effectiveness in predicting H1N1 and seasonal flu vaccination [8]. CatBoost is an ensemble machine learning algorithm, that utilises gradient boosting techniques. It constructs an ensemble of decision trees iteratively, where each subsequent tree improves on the previous ones by giving more weight to the misclassified data. This enhances the overall performance of the ensemble. CatBoost stands out from traditional gradient boosting algorithms as it uses a novel technique called ordered boosting. This technique is designed to tackle overfitting which is common in gradient boosting algorithms. Furthermore, by considering the order of objects in the dataset, the ordered boosting algorithm helps identify underlying patterns that may exist in the data which improves the CatBoost model performance.

CatBoost is a powerful algorithm that can handle structured and categorical data. Previous studies [21] have showcased its effectiveness in handling categorical data without requiring explicit encoding. This makes CatBoost particularly well-suited for datasets with many categorical features. However, to ensure a fair comparison with RF, our data was encoded instead of utilising CatBoost's built-in function. CatBoost also has a built-in mechanism to handle imbalanced datasets, but it does not apply to multilabel problems. Hence, resampling techniques are necessary to address the class imbalance issue in our paper. Additionally, CatBoost typically has a longer training time, making it computationally expensive. Moreover, its performance heavily depends on its hyperparameters, and finding the optimal combination can be extremely time-consuming [22].

The *CatBoostClassifier* was used for the classification problem. Hyperparameter tuning was performed using the *grid_search* function provided by the *catboost* library with 5-fold cross-validation. Cross-validation helps ensure that the model generalises well to different training sets. Four hyperparameters recommended by [23] were considered: *'iterations'*, *'learning_rate'*, *'depth'*, and *'L2_leaf_reg'*. *'iterations'* represents the number of trees and was tested with values 1000 and 1500 to balance the model performance and complexity. *'learning_rate'* controls the step size of the model and was experimented with values 0.05 and 0.1. *'depth'* was tuned to determine the best maximum depth for each decision tree, avoiding overfitting due to overly complicated trees. Lastly, *'L2_leaf_reg'* is the L2 regularisation hyperparameter, tested with values 0.3 and 0.5, to control overfitting and the model's sensitivity to noise.

D. Model Evaluation

1) Accuracy

This is the simplest performance measure that computes the percentage of correct predictions over all predictions made.

2) Hamming Loss

Hamming loss is a metric for multilabel problems. It takes both prediction errors and missing labels into account. The metric reflects the proportion of labels that are incorrect or missing compared to the total. A lower value indicates better accuracy, with 0 representing perfect prediction and 1 indicating complete misclassification or missing labels.

3) F1 Score

Micro F1 Score. The micro F1 score is useful to compare model performance in multilabel classification as it provides an overall understanding of the model performance by considering the true positives (TP), false positives (FP), and false negatives (FN).

Weighted F1 Score. The weighted F1 score is computed using precision and recall. This makes it suitable for imbalanced datasets as it assigns equal weight to each label's probability.

4) Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC plots the false positive rate (FPR) against the true positive rate (TPR). The graph provides a comprehensive examination of a model's performance across all class distributions. Thus, it is a reliable metric for imbalanced datasets.

IV. Results

Fig. 3 and Fig. 4 shows a similar downward trend despite having different y-axis range. Both figures depict that the vaccine uptakes tend to be higher when there are fewer children in a household. Specifically, the vaccine uptake is particularly high in households with no children. On the other hand, Fig. 5 and Fig. 6 demonstrate a right-skewed data distribution where the highest vaccine uptake is observed when there is only one adult in the household. Interestingly, the second highest number of vaccine uptakes is when there are no adults in the household.

The statistical analysis performed on the dataset revealed that the majority of features displayed a p-value below 0.05 as depicted in Fig. 7 and Fig. 8 where they are highlighted in red. The Chi-Square statistic values, on the other hand, showed much more variation among the features, with many exhibiting high values. To further understand the factors that may influence the vaccination rate of the population, the top five features with the highest Chi-Square statistic values with the target variables were identified and displayed in Table I. The 'doctor_recc_h1n1' feature has the strongest correlation to the 'h1n1_vaccine' target outcome, while 'opinion_seas_vacc_effective' obtained the highest Chi-Square statistics in relation to the 'seasonal_vaccine' target outcome. It should be noted that applying mean and median imputation to the dataset did not result in significant differences in the Chi-Square test. The variations observed between the two were minimal, with differences of less than 0.0001. However, in the heatmap, different colours can be observed for some boxes despite having the same value of 0.0000 shown. This is due to space limitation, which prevents the full value from being displayed. As the Chi-Square p-value and statistic value heatmap are very similar, only one is presented in this paper.

The results from the *GridSearchCV* for RF are shown in Table II. The resampling techniques improve the overall performance of the model significantly. The highest F1 micro score is 0.9191 for the combination of random oversampling and using mean imputation, while the lowest is 0.6885 for using mean imputation with no resampling. It can also be seen that all chosen *'max_depth'* values are consistent, having a value of 40. Additionally, only the best model has a value of 400 for its *'n_estimators'* parameter. Another observation is that only when random oversampling is used, the *'max_features'* parameter will be having the value *'log2'* after hyperparameter tuning with *GridSearchCV*.

As observed in Fig 9, the performance of the CatBoost algorithm was significantly worst without resampling. Notably, the algorithm showed the best performance when the dataset underwent SMOTEENN resampling. Additionally, no significant difference was observed when the CatBoost algorithm was applied to the dataset that underwent mean or median imputation.

Using the best pre-processing pipeline and hyperparameter combination found using grid search, two new models, hereby referred to as "the best RF model" and "the best CatBoost model", were trained and evaluated with their results presented in Table III. The accuracy and AUROC of each label are measured to provide a thorough evaluation of its multilabel prediction accuracy. The best RF model was able to achieve a relatively satisfactory combined AUROC value of 0.8393, whereas the best CatBoost model achieved a slightly lower AUROC value of 0.8271.

V. Discussion

The findings illustrated in Fig. 3 and 4 suggest that median imputation may be much suited for the *'household_children'* feature as it is not a normal distribution. On the contrary, Fig. 5 and 6 depict a near-normal data distribution. Therefore, mean imputation can be done for the *'household_adults'* feature. Using the appropriate imputation technique would ensure that the overall shape of the distribution is preserved. This is crucial to ensure that the subsequent data analyses will be reliable.

The negative correlation observed in Fig. 3 and 4 is in line with the findings of [24], which revealed that adults within the same household are concerned about vaccine safety is one of the main reasons that contributes to their unwillingness to be vaccinated. Therefore, relevant healthcare authorities should prioritise educating adults, especially those residing in households with children, about vaccines. On the contrary, in single-adult households, the decision-making process regarding vaccination is solely dependent on one adult. This eliminates the possibility of conflicting opinions, and the decision can be made quickly which may explain the observed trend seen in Fig. 5 and 6.

The extremely low p-value strongly indicates a high chance of meaningful dependencies between all the features observed in Fig. 7 and Fig. 8, indicating that the survey had only collected the necessary information. This finding implies that all features are important to make accurate predictions, thereby eliminating the need for feature selection.

Doctors' recommendations (*'doctor_recc_h1n1'* and *'doctor_recc_seasonal'*) are a critical factor that influences individuals' vaccination decisions as indicated in Table I. This finding is supported by previous research [4] where the U.S. population vaccination rate is highly dependent on the advice given by healthcare workers. The strong trust in doctors suggests that the effectiveness of awareness campaigns can be maximised by having doctors present or lead them. The perception of the vaccine effectiveness and the perceived risks of not being vaccinated are strongly associated with vaccine uptake which is also in line with the other studies [1]. This is self-evident as individuals that perceive vaccines as effective and recognise the potential consequences of not taking the vaccine are more likely to be vaccinated for protective measures.

The vaccination behaviour of the population can be determined using Fig. 2. A much higher percentage of respondents have received the seasonal flu vaccination compared to the H1N1 vaccine. This finding suggests that there is a higher acceptance of the seasonal flu vaccine, which may be due to familiarity as it is a routine vaccination that is recommended annually. In contrast, the population is less accepting of the H1N1 vaccine as they are unfamiliar with it. Therefore, this highlights the importance of awareness campaigns to increase public familiarity with the H1N1 vaccine.

By comparing the highest and lowest F1 micro score in Table II, the large imbalance in the dataset and parameter tuning are dominant factors in affecting the performance of a model, as both the best scoring and lowest scoring are using mean imputation but differ in the presence of resampling techniques and the value of the '*max_features*' hyperparameter. Although RF is noted to handle imbalanced datasets, the use of resampling techniques in Table II is undeniably beneficial. This may be because RF has a certain limit for handling imbalances, i.e., if the imbalance is over a certain ratio, it will not be able to perform well, however, this could also be due to the limitations of the hyperparameter tuning presented in this paper.

It is also obvious that the resampling technique applied, and the 'c*riterion*' are the main contributors towards the increase or decrease in performance. The consistent choice of '*max_depth*' having a value of 40 suggests that the value could have been increased. It can also be observed that increasing the number of individual trees in the forest ('*n_estimators*') does not always increase the performance and that hyperparameters may have some co-dependency relationship. Another observation from Table II is that except for random oversampling, all other combinations have the exact grid search results where this could be caused by the sensitivity of RF towards noise or bias that may have been produced by the imputation methods.

As mentioned previously, CatBoost can handle imbalanced data by adjusting the class weight. However, this approach does not apply to multilabel classification. Therefore, without using any resampling technique to address the imbalanced data, the performance of the CatBoost algorithm is greatly affected. This is consistent with the findings depicted in Fig. 11, where the algorithm showed extremely poor performance when no resampling technique was used. The CatBoost algorithm achieved the best performance with SMOTEENN resampling as seen in Fig. 9. SMOTEENN is a hybrid resampling technique that can overcome the limitations of random oversampling. The combination of SMOTE and ENN has greatly improved the model's performance by removing irrelevant instances while increasing the samples of the minority class. Furthermore, the observed result suggests that the CatBoost algorithm may have been slightly overfitted with random oversampling as compared to SMOTEENN.

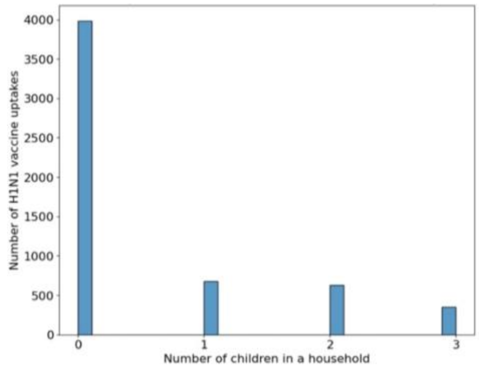

Fig. 3. Relationship between the number of children and H1N1 vaccine uptakes

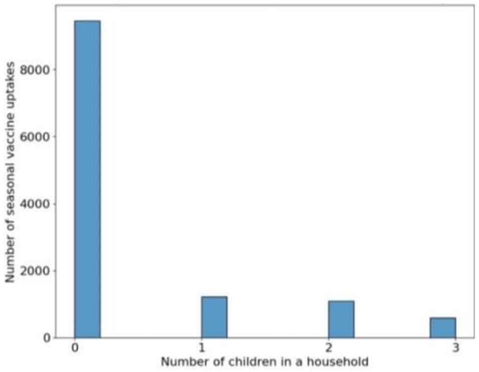

Fig. 4. Relationship between the number of children and seasonal vaccine uptakes

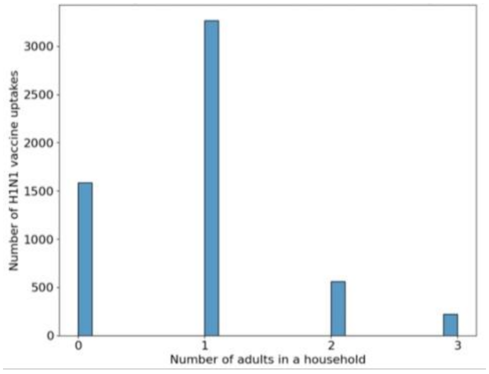

Fig. 5. Relationship between the number of adults and H1N1 vaccine uptakes

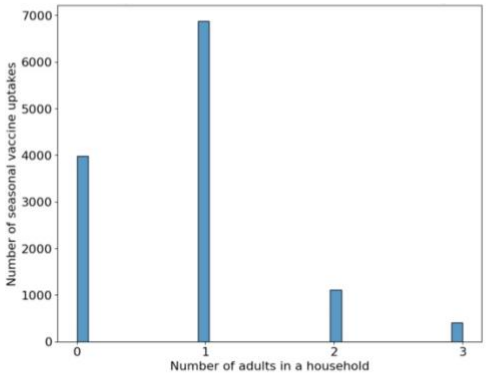

Fig. 6. Relationship between the number of adults and seasonal vaccine uptakes

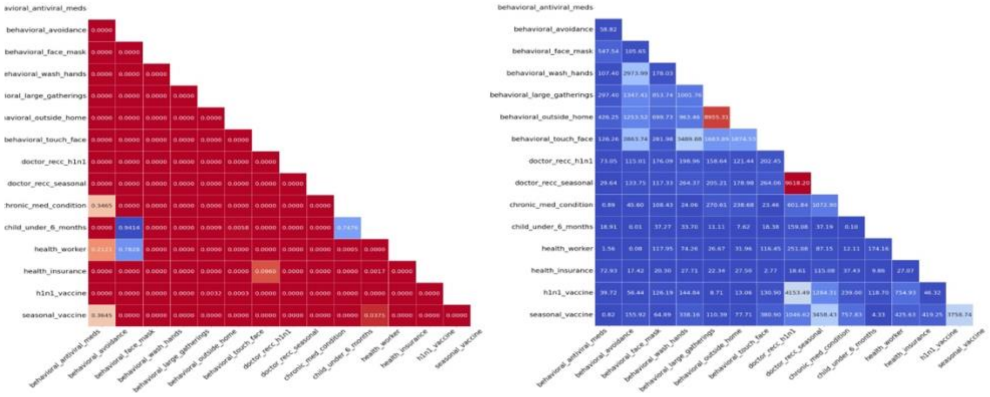

Fig. 7. Chi-Square P-Value and Statistics Heatmap for Binary Features

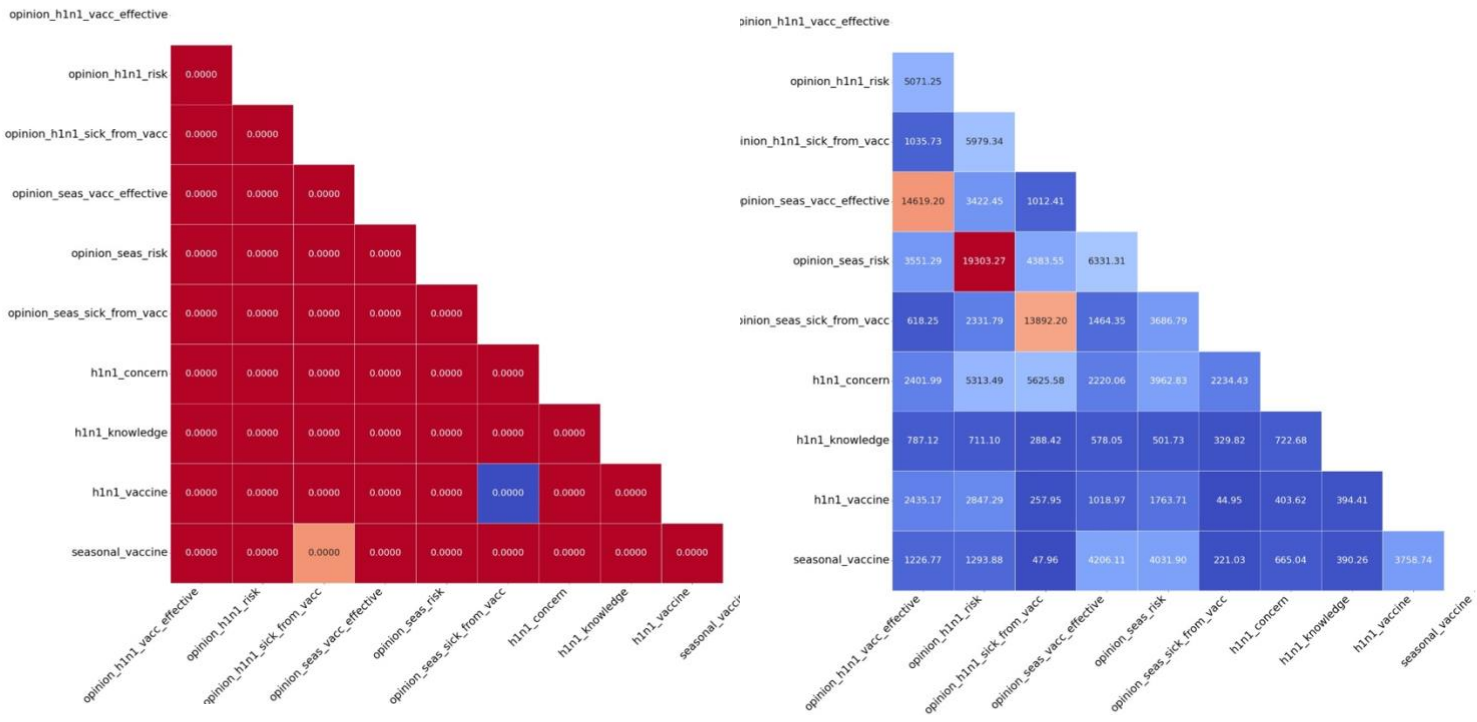

Fig. 8. Chi-Square P-Value and Statistics Heatmap for Ordinal Features

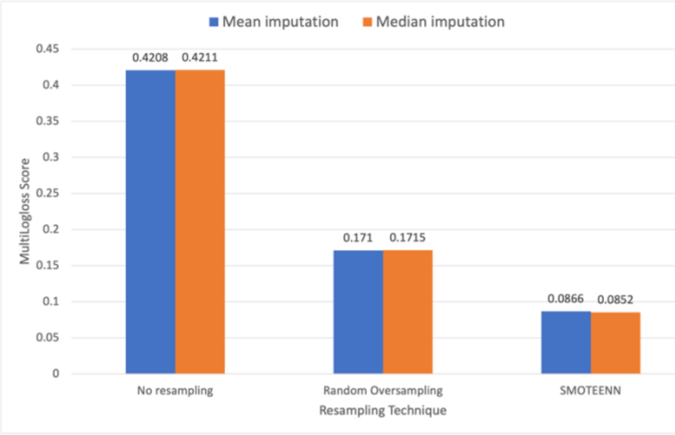

Fig. 9. GridSearchCV Result (MultiLogloss) for CatBoost

TABLE I.

FEATURES WITH TOP 5 CHI-SQUARE STATISTICS

H1N1 Vaccine	Seasonal Vaccine
doctor_recc_h1n1	opinion_seas_vacc_effective
seasonal_vaccine	opinion_seas_risk
opinion_h1n1_risk	h1n1_vaccine
opinion_h1n1_vacc_effective	doctor_recc_seasonal
opinion_seas_risk	age_group

TABLE II.

RESULTS OF GRIDSEARCHCV FOR RANDOM FOREST

Resampling Technique	Imputation Method	Hyperparameters				F1 Micro Score
		criterion	*max_depth*	*max_features*	*n_estimators*	
No resampling	Mean	gini	40	log2	400	0. 9191
	Median	entropy	40	log2	500	0.9187
Random Oversampling	Mean	entropy	40	sqrt	500	0. 8869
	Median	entropy	40	sqrt	500	0.8876
SMOTEENN	Mean	gini	40	sqrt	500	0.6885
	Median	gini	40	sqrt	500	0.6892

TABLE III.

BEST MODELS PERFORMANCE

Model	Evaluation Metrics								
	Overall Accuracy	*Accuracy (H1N1 Vaccine)*	*Accuracy (Seasonal Vaccine)*	*Hamming Loss*	*F1 Micro Score*	*F1 Weighted Score*	*AUROC (H1N1 Vaccine)*	*AUROC (Seasonal Vaccine)*	*Combined AUROC*
Random Forest	0.6588	0.8362	0.7756	0.1941	0.7010	0.6863	0.8276	0.8509	0.8393
CatBoost	0.5585	0.7337	0.7213	0.2725	0.6789	0.6859	0.8133	0.8410	0.8271

Applying either mean or median imputation did not have a significant impact on both the models' performance. This finding is depicted in Fig. 9, which showed no noticeable difference in the model's performance between the mean and median imputation technique applied to the dataset. Therefore, it can be inferred that the primary factor influencing the algorithm's performance lies in the choice of resampling technique. Besides that, the mean and median imputation were only done on two numerical features. The small proportion of missing values in those two features may have reduced the impact of the imputation technique.

Overall, the best RF model has a higher combined AUROC value than the best CatBoost model, with a small difference of 0.0122. Both models have a higher accuracy in predicting the H1N1 vaccine while having a higher AUROC for seasonal vaccines (Table III). However, the difference between the prediction performance of both vaccines in accuracy and AUROC have a smaller difference in CatBoost as opposed to RF. Therefore, it can be said that CatBoost is more dependable when considering each type of vaccine separately, while RF is better performing when considering vaccination rates in general. A possible hypothesis from the results is that for certain resampling techniques, RF performs better than CatBoost, to conclude if this hypothesis is dependent on the dataset will require further research. Tree-based models can be seen to have relatively well performance on imbalanced datasets when combined with resampling techniques.

VI. CONCLUSION

The visualisation of the chi-square test can be used to answer both question 1 and question 2, indicating the effect of personal assumptions on vaccination. The influence of having information about vaccination from a trustable and reliable source also contributes towards the positive increase in vaccination rate for both vaccines.

The best model evaluated using the evaluation measures in this paper is the Random Forest model with hyperparameters of: 'criterion=gini', 'max_depth=40', 'max_features=log2', 'n_estimators=400'; that also applies random oversampling and mean imputation. Although satisfactory results on the performance measures were achieved by each model, the hyperparameter tuning was minimal due to time restrictions. Further experiments on the hyperparameters can be conducted by varying the number of hyperparameters tuned to determine the effect on the models. Additionally, future research can consider other resampling techniques that could potentially improve the RF and CatBoost accuracy. Although the performance of the models presented in this paper was satisfactory for large-scale productions relating to the health and safety of a population, models with higher accuracy will be able to address question 3 more confidently.

The findings of this paper are reliable as the dataset is from a reliable source but is constrained by the fact that it only applies to the U.S. population. Furthermore, the dataset may not be representative of the current U.S. population given that another pandemic has taken place in the year 2019: the infamous Coronavirus Disease (COVID-19) which is also a respiratory disease, where views and opinions of the population may have experienced a drastic change. Therefore, the survey should be re-conducted and analysed to obtain more dependable results, providing insights on possible interventions to improve the vaccination rate for H1N1 vaccines, and optimistically for all respiratory vaccines.

Authors' Contributions

CHEQ and CTK both contributed to the design and objectives of the paper. The abstract and introduction were written by CTK. On the other hand, CHEQ wrote the literature review and conclusion. As for the methodology, CHEQ wrote the sections on data distribution of numerical features, dealing with missing values, correcting data types, and model evaluation; while CTK wrote the description of the Chi-Square test, one-hot encoding, and the resampling techniques. CHEQ described the RF model and CTK described the CatBoost model. Both authors contributed to the results and discussions. Additionally, both authors contributed to the revision, read, and approved the submitted version of this paper.

References

[1] T. A. Santibanez, J. A. Singleton, S. S. Santibanez, P. Wortley and B. P. Bell, "Socio-demographic differences in opinions about 2009 pandemic influenza A (H1N1) and seasonal influenza vaccination and disease among adults during the 2009–2010 influenza season," in *Annual meeting of the American Public Health Association*, Denver, 2010.

[2] J. Partridge and M. P. Kieny, "Global production of seasonal and pandemic (H1N1) influenza vaccines in 2009–2010 and comparison with previous estimates and global action plan targets," *Vaccine,* vol. 28, no. 30, pp. 4709-4712, 2010.

[3] R. H. Borse, S. S. Shrestha, A. E. Fiore, C. Y. Atkins, J. A. Singleton, C. Furlow and M. I. Meltzer, "Effects of Vaccine Program against Pandemic Influenza A(H1N1) Virus, United States, 2009–2010," *Emerging Infectious Diseases,* vol. 19, no. 3, pp. 439-448, 2013.

[4] J. Maurer, L. Uscher-Pines and K. M. Harris, "Perceived seriousness of seasonal and A(H1N1) influenzas, attitudes toward vaccination, and vaccine uptake among U.S. adults: Does the source of information matter?," *Preventive Medicine,* vol. 51, no. 2, pp. 185-187, 2010.

[5] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *International Journal of Data Warehousing and Mining,* vol. 3, no. 3, pp. 1-13, 2009.

[6] P. E. Kafrawy, A. Mausad and H. Esmail, "Experimental Comparison of Methods for Multi-Label Classification in Different Application Domains," *International Journal of Computer Applications,* vol. 114, no. 19, p. 0975 – 8887, 2015.

[7] W. Liu, H. Wang, X. Shen and I. W. Tsang, "The Emerging Trends of Multi-Label Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[8] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia and M. Dixit, "Predicting H1N1 and Seasonal Flu : Vaccine Cases using Ensemble Learning approach," in *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN),* 2020.

[9] S. Inampudi, G. D. Johnson, J. Jhaveri, S. Niranjan, K. Chaurasia, and M. K. Dixit, "Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination," in *Communications in computer and information science*, Springer Science+Business Media, 2020. doi: 10.1007/978-981-16-0401-0_11.

[10] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence,* vol. 17, no. 5-6, pp. 519-533, 2003.

[11] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data,* vol. 8, 2021.

[12] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence,* vol. 17, no. 5-6, pp. 519-533, 2003.

[13] K. Potdar, T. S. Pardawala, and C. D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7–9, 2017, doi: 10.5120/ijca2017915495.

[14] A. Puri and M. Gupta, "Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data," *The Computer Journal*, vol. 65, no. 1, pp. 124–138, May 2021, doi: 10.1093/comjnl/bxab039.

[15] T. M. Oshiro, P. S. Perez and J.´. A. Baranauskas, "How Many Trees in a Random Forest?," 2012.

[16] L. Breiman, "Random Forests," 2001.

[17] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha and S. Kundu, "Improved Random Forest for Classification," *IEEE TRANSACTIONS ON IMAGE PROCESSING,* vol. 27, no. 8, 2018.

[18] P. Probst and A.-L. Boulesteix, "To Tune or Not to Tune the Number of Trees in Random Forest," *Journal of Machine Learning Research,* vol. 18, pp. 1-18, 2018.

[19] J. L. Speiser, M. E. Miller, J. Tooze and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Systems with Applications,* vol. 134, pp. 93-101, 2019.

[20] N. Rastin, M. z. Jahromi and M. Taheri, "Multi-label Classification Systems by the Use of Supervised Clustering," in *Artificial Intelligence and Signal Processing Conference (AISP)*, 2017.

[21] J. Hancock and T. M. Khoshgoftaar, "Medicare Fraud Detection using CatBoost," *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, Las Vegas, NV, USA, 2020, pp. 97-103, doi: 10.1109/IRI49571.2020.00022.

[22] J. Hancock and T. M. Khoshgoftaar, "Impact of Hyperparameter Tuning in Classifying Highly Imbalanced Big Data," *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, Las Vegas, NV, USA, 2021, pp. 348-354, doi: 10.1109/IRI51335.2021.00054.

[23] C. Bentéjac, A. Csörgő and G. Martínez-M, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, p. 1937–1967, 2020, doi: 10.1007/s10462-020-09896-5.

[24] C. McKee and K. K. Bohannon, "Exploring the Reasons Behind Parental Refusal of Vaccines," *The Journal of Pediatric Pharmacology and Therapeutics*, vol. 21, no. 2, pp. 104–109, Apr. 2016, doi: 10.5863/1551-6776-21.2.104.