Predicting Nodes for Effectively Spreading Vaccine Awareness: An Influence Maximization Approach

Deboleena Bhattacharyya
Department of Computer Science Engineering
SRM Institute of Science and Technology, Ramapuram Campus
Chennai, India
Email: deboleenab46@gmail.com

Khavin Shankar G
Department of Computer Science Engineering
SRM Institute of Science and Technology, Ramapuram Campus
Chennai, India
Email: khavinshankar@gmail.com

P Aaranan
Department of Computer Science Engineering
SRM Institute of Science and Technology, Ramapuram Campus
Chennai, India
Email: aarananprabaharan@gmail.com

K Raja
Department of Computer Science Eng
SRMIST, Ramapuram Campus
Chennai, India
Email: rajak1@srmist.edu.in

Amira Alturki
Department of Computer Science
University of Business & Technology
Saudi Arabia
Email: a.turki@ubt.edu.sa

Mithileysh Sathiyanarayanan
Research & Innovation
MIT Square Services Private Limited
London, UK
Email: mithileysh@mitsquare.com

2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) | 978-1-6654-3773-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISAECT53699.2021.9668587

***Abstract*— Many large pandemics have occurred throughout human history, and pandemic-related crises have had massive detrimental effects on global health, economics, and even national security. Aside from the potentially lethal spread of infections, one key problem is persuading people to get vaccines. For decades, problematic negative vaccine beliefs have been the largest predictor of opposition to vaccination efforts. In this research, we attempted to examine the probability of an individual to be vaccinated or not and offer a strategy for efficiently spreading awareness among people via a well-executed network. Finding a set of seed nodes that maximizes the spread of influence in a social network is known as influence maximization. The seed nodes are employed in viral marketing to maximize profit by leveraging effective word-of-mouth. By finding the most important seeds in the network, we want to use this strategy to raise vaccination awareness in communities. For prediction, we compared a few deep learning models and a random forest classifier. Then we examined the model with the best performance and conducted a basic study on employing influence maximization to raise awareness in distinct communities. The best model was roughly eighty four percent accurate and forecasted us those who are most likely to be vaccinated, which could be utilized as seeds for our Influence Maximization model, to spread vaccination awareness among the unvaccinated, which can also be discovered by our prediction model.**

***Keywords—vaccination, awareness, pandemic, prediction, influence maximization, social network*.**

I. INTRODUCTION

Vaccination is considered to be one of the most notable achievements of humankind. It has made an immense contribution towards the health factor of human beings. It has aided us in eradicating life-threatening diseases such as smallpox and rinderpest. Vaccines provide human beings and other living things, the necessary immunity to fight off diseases. In 1974, the World Health Organization (WHO) proposed the extended program on Immunization to typically develop and expand immunization programs throughout the world. Later in 2000, WHO created the Global Alliance for Vaccination and Immunization (GAVI) to improve access to vaccines for children living in under developed countries.

Despite such big measures taken, approximately 2.3 million children continue to die each year from vaccine-preventable diseases. Availability, Acceptability, Affordability, Accessibility, Quality and lack of awareness of vaccination were considered to be the biggest reasons contributing to such cases. According to various studies, vaccine hesitancy was also a factor contributing to the above-specified mortality rate. Evidence to this claim can be seen in the paper authored by Dubé et al. [1] who did an overview regarding this topic. According to the study, several factors such as cognitive, social, emotional and environmental factors were involved in an individual's decision to take a vaccine or not.

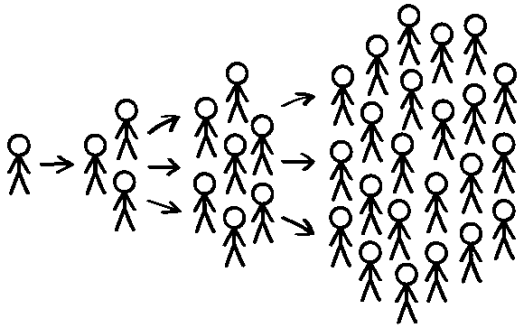

Fig. 1.Influence Maximization

As stated above, lack of access to these vaccines were also considered to be an issue for the death rate due to vaccine preventable diseases. Lydon et al. [2] conducted a study on vaccine stockouts around the world. The studies revealed that stockouts occurred at a regular basis in some areas leading to non-availability of vaccines when required. To make sure that these situations don't occur often, Mueller et al. [3] proposed the idea of a demand forecasting system which fulfilled the vaccine demand requirements.

To make sure that every individual gets vaccinated, one can spread awareness about the importance of vaccination in several ways such as social media awareness, organizing campaigns, etc. When it comes to spreading awareness in social media, the method of influence maximization can be used. Influence maximization [4] is the method to find a subset of nodes

called as seed set, which are responsible for the maximum influence spread in a given network. Using this set of seed nodes, one can make sure that they can spread the importance of awareness to reach a higher number of people.

In this paper, we will be incorporating the concepts of deep learning and machine learning algorithms to predict the unvaccinated individuals in the given H1NI vaccination dataset. We will also propose the idea of an influence maximization model to make sure that one gets educated about the importance of taking vaccines, thereby motivating them to get vaccinated.

II. RELATED WORK

Vaccination is an important health management requirement as it helps us from preventing diseases and other several complications. Several studies and research were conducted with respect to this topic. For example, Brian Greenwood [5] conducted an in-depth analysis of the contribution of vaccines in the past, present and also future. Vaccination trends and other statistical patterns were also observed by several researchers. Inampudi et al. [6] in their research paper, predicted whether a person will take vaccination or not based on their behavioral patterns. This was performed on an H1N1 and flu survey database conducted in 2009. A total of 9 algorithms were utilized for this purpose. The best accuracy was provided by an open library algorithm called CatBoost [Categorical Boosting] [7]. It gave an accuracy rate of 86% when performed in the given database. This could be used for predictions in other models of similar categories for future purposes.

Hariharan et al. [8] proposed a prediction model to provide accurate utilization forecast for vaccinations. This was done to make sure that the specific location workers/doctors have access to information regarding the stock availability of vaccines. For this research, data was retrieved from the regional health facilities of the region of Tanzania. They used a random forest regressor [9] to predict the bi-weekly vaccine utilization because the model outperformed majority of the commonly used machine learning algorithms. Two different models named RFR1 and RFR2 were proposed by the authors. RFR1 model 1 had a high prediction rate compared to RFR2 but RFR2 has a much broader applicate range and scope. This model proved to be useful to determine the vaccine availability.

Studies were carried out on identifying the number of people who got vaccinated or not. Dick. K et al. [10] proposed a predictive model for identifying the number of unvaccinated people using the 2009-14 Canadian Community Health surveys (CCHS). The survey relied on several features such as behavioral patterns, pre-medical conditions, living environment etc. Based on these features, two random forest classifiers were made based on a policy system (age<60 and age>=60). This policy system was further used to spread the awareness about vaccines to the unvaccinated set of people accordingly.

The awareness part where the importance of taking a vaccination can be spread is mainly done using the method of influence maximization. Chen et al. [11] proposed a model to find selective seed notes which helps in maximum influence spread otherwise known as influence maximization. In their model, topic-aware influence maximization queries (TIM) were used to find the seed nodes which can be used to increase the influence spread. An efficient algorithm with an approximation ratio of 1-1/*e* was proposed to find the upper tight bounds requirement. Another topic aware algorithm was also proposed for achieving efficient and fast performances while having an high influence spread. The results outperformed other baseline approaches which were available.

Further studies on influence maximization were done by Rabadiya K et al. [12] where they compared two different influence diffusion models named linear threshold and independent cascade model which are basically used to spread influence in networks. This provided scope for further improvisation in making the models even more scalable to achieve influence maximization. Litou et al. [13] proposed a model which used influence maximization that explained how a campaign can influence a user decision to support them or not.

There are several machine learning algorithms available to be used for a prediction model. To find such an efficient and compatible method, Spieser et al. [14] in their paper tried to find the best algorithm that can be used for prediction purpose by comparing each algorithm's common features and accuracy rate. Random forest proved to be one of the best algorithms to be used for prediction models as a result. Hence in our model, we will be using a random forest classifier along with deep learning models and feature engineering [15] concept for the implementation purpose.

III. METHODS

A. Data Collection

Data from The National 2009 H1N1 Flu Survey (2009) conducted by U.S. Department of Health and Human Services (DHHS) was used for our prediction. The study questioned respondents if they had got the H1N1 and seasonal flu vaccines, as well as personal questions. These additional questions inquired about respondents' backgrounds (social, economic, and demographic), attitudes on illness risks and vaccine effectiveness, and so forth.

The variables were organized into various groups, and these groups were manually filtered to identify those that were most suited to our application. Due to a lack of information, entire groupings of variables were eliminated. To make interpretation more intuitive, categorical data were turned into binary variables for each category after manually deleting groups of variables.

To remove observations with missing data, several sample limits have to be imposed. These constraints included the deletion of persons who did not submit information, as well as observations missing behavioral information, and missing income information. Out of total 33723 observations, many of the features have at least 20% of their values missing in their entirety. Some characteristics have missing values

that account for more than half of their data. All of these exclusion counts are for exclusions related with the dataset.

B. Data Preprocessing

a) Preprocess 1(Feature Engineering):

We considered demographic information about individuals who are likely associated with vaccination decisions (e.g. a child under six months) as well as health information associated with vaccination decisions (e.g. having chronic medical condition) to identify variables that were important in predicting the likelihood that an individual is vaccinated. In addition to these factors, this analysis incorporated behavioral characteristics (e.g., mask use, hand washing) to better explain the decisions. As a result, there may be underlying personality traits that inspire all of these sorts of behaviors that must be accounted for in the model. Table I shows the mean and standard deviation of all variables that were considered.

We presented three columns in feature engineering that combined behavioral features, medical features (i.e., frequent or seasonal doctor consultation), and awareness (i.e., h1n1 knowledge, concern etc.). Each column represented the sum of the values assigned to the features when they were combined.

b) Preprocess 2:

We tried three different steps or processes. To begin, we employed label encoding for all ratings-based columns; we used a negative value for negative responses, a positive value for positive responses, and a 0 for don't know, not sure, and nan values. Then, for category-based columns ("employment industry", "employment occupation"), we used one-hot encoding. Finally, we deleted location-based columns ("state", "census msa"), because we had 136 features with location-based columns one-hot encoded, and eliminating them reduced us to 86 features with no significant impact in accuracy (if needed we could also try label encoding location-based columns)

C. Model Building

Using the preprocesses outlined in the preceding part, we used deep learning architecture to generate two models. The variables that were found became binary characteristics that were considered in the building of our predictive model. These models were pitted against the Random Forest Classifier.

a) Random Forest: The most important metrics for our evaluation are the Positive Predictive Value (PPV) , which is defined as the ratio of correctly identified positives by our classifier as shown by (1), the Negative Predictive Value (NPV), which is defined as the ratio of correctly identified negatives by our classifier as shown by (2) , and the accuracy (ACC), which is defined as the ratio of correctly classified individuals as shown by (3).

$$PPV = TP / (TP + FP) \quad (1)$$

$$NPV = TN / (TN + FN) \quad (2)$$

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

TABLE I. SAMPLE MEANS WITH STANDARD DEVIATIONS

Variable	Mean	S.D
h1n1_concern	1.613448	0.910372
h1n1_knowlede	1.259236	0.618602
behavioral_antiviral_meds	0.051912	0.221854
behavioral_avoidance	0.727320	0.445436
behavioral_face_mask	0.068554	0.252699
behavioral_wash_hands	0.824682	0.380246
behavioral_large_gatherings	0.353942	0.478200
behavioral_outside_home	0.336933	0.472671
behavioral_touch_face	0.677104	0.467592
doctor_recc_h1n1	0.240360	0.427309
doctor_recc_seasonal	0.333426	0.471445
chronic_med_condition	0.250796	0.433478
child_under_6_months	0.084119	0.277571
health_insurance	0.892853	0.309308
health_worker	0.111409	0.314644
education_comp	2.981597	1.000320
raceeth4_i	2.828989	0.653395
sex_i	1.570886	0.494957
inc_pov	2.130979	0.990891
marital	1.462780	0.498623
rent_own_r	2.931631	12.654156
census_region	2.711532	1.032819
n_adult_r	1.941252	0.747974
household_children	0.797264	1.036703
n_people_r	2.741894	1.393036
hhs_region	5.364410	2.704540
vacc_h1n1_f	0.239065	0.426519

b) Model1: Model 1 is a deep learning-based binary classification model. Using the features collected as input, we developed a fully connected (FC) neural network to categorise our data into two classes: vaccinated (0) and not vaccinated (1). We add FC layers to make the model end-to-end trainable. The fully connected layers develop a (potentially non-linear) function shown by (4), between the high-level characteristics provided by the convolutional layers as an output. We used the rectified linear activation function, or ReLU, to directly output the input if it is positive; else, it will output zero.

$$f(x)=\max(0,x) \quad (4)$$

Both the ReLU function and its derivative are monotonic. If it receives any negative input, it returns 0, but

if it receives any positive input, it returns that value. As a result, it produces an output with a range of 0 to infinity.

Then we apply sigmoid activation function on this FC layer. This function converts the input into a value between 0.0 and 1.0. Inputs that are significantly larger than 1.0 are changed to 1.0, and values that are much lower than 0.0 are snapped to 0.0.

$$Y = 1 / 1+e^{-z} \quad (5)$$

So in (5), if the value of z approaches positive infinity, the expected value of y becomes 1, and if it approaches negative infinity, the predicted value of y becomes 0. And if the result of the sigmoid function is greater than 0.5, we classify that label as class 1 or positive class, and if it is less than 0.5, we classify it as class 0 or negative class.

The loss function of our model 1 was BCELoss, and the optimizer was Adam. It takes approximately 50-300 epochs with a learning rate of 0.01.

c) Model2: Model 2 is a multiclass classification model based on deep learning. In this model, we use the Leaky ReLU function and Softmax to create a fully connected (FC) neural network. The Leaky ReLU function improves on the ReLU activation function. In terms of the ReLU activation function, the gradient is 0 for all input values less than zero, which deactivates the neurons in that region and may cause the dying ReLU problem. To solve this issue, Leaky ReLU is defined. We specify the ReLU activation function as an extremely small linear component of x rather than as 0 for negative values of inputs(x). This activation function's formula is as represented by (6).

$$f(x)=\max(0.01*x , x) \quad (6)$$

If it receives a positive input, it returns a positive x, but if it receives a negative input, it returns a very little number equal to 0.01 times x. As a result, it also produces an output for negative values. By making this minor change, the gradient on the left side of the graph becomes non-zero. As a result, we would no longer meet dead neurons in that region, improving the accuracy of our model.

In this model, instead of sigmoid, we utilize the Softmax activation function (7) in the output layer.

$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)} \quad (7)$$

The Z in this case represents the values from the output layer's neurons. The non-linear function is the exponential. These values are then normalized and converted into probabilities by dividing them by the sum of exponential values.

The loss function of our model 1 was CrossEntropyLoss, and the optimizer was Adam. It also takes approximately 50-300 epochs with a learning rate of 0.01.

IV. RESULTS

The outcome of Model 1 is a number in the range of (0, 1). If the number is more than 0.5, the individual is vaccinated; otherwise, the individual is not vaccinated (no). The function has an S-shape for all potential inputs ranging from zero to 1 whereas Model 2's output has two numbers. The first is a likelihood of no (not vaccinated), while the second is a probability of yes (vaccinated). When two numbers are added together, we get one.

Table II displays the performance metrics of our three models, as well as their behavior when run individually with both preprocessors. In total, we receive five alternative models: four are preprocess 1 with model 1(P1M1), preprocess 1 with model 2(P1M2), preprocess 2 with model 1(P2M1) and preprocess 2 with model 2(P2M2), and one is our Random Forest Classifier. The Random Forest classifier uses preprocess 2 to attain the best performance and accuracy.

TABLE II. COMPARISON OF PERFORMANCE METRICS OVER FIVE MODELS

Model	Values	Precision	Recall	F1 score	Support	Accuracy
P1M1	0	0.85	0.94	0.89	6365	83%
	1	0.72	0.49	0.59	2066	
P1M2	0	0.85	0.94	0.89	6365	83%
	1	0.72	0.95	0.59	2066	
P2M1	0	0.84	0.94	0.89	6365	82%
	1	0.71	0.45	0.56	2066	
P2M2	0	0.84	0.94	0.89	6365	82%
	1	0.71	0.45	0.55	2066	
Random Forest	0	0.84	0.95	0.89	6365	83%
	1	0.74	0.45	0.56	2066	

It was observed that models 1 and 2 with preprocess 1 and Random Forest had the same accuracy, i.e. 83 percent (which is the highest accuracy among our all models). When the performance indicators were examined further, the total f1 score of model 1 with preprocess 1 and that of model 2 with preprocess 2 were somewhat higher than Random Forest. However, when the AUC values of each model were compared, Random Forest provided the highest value (0.843). As a result of the AUC values, the Random Forest classifier has the best performance, however the difference is

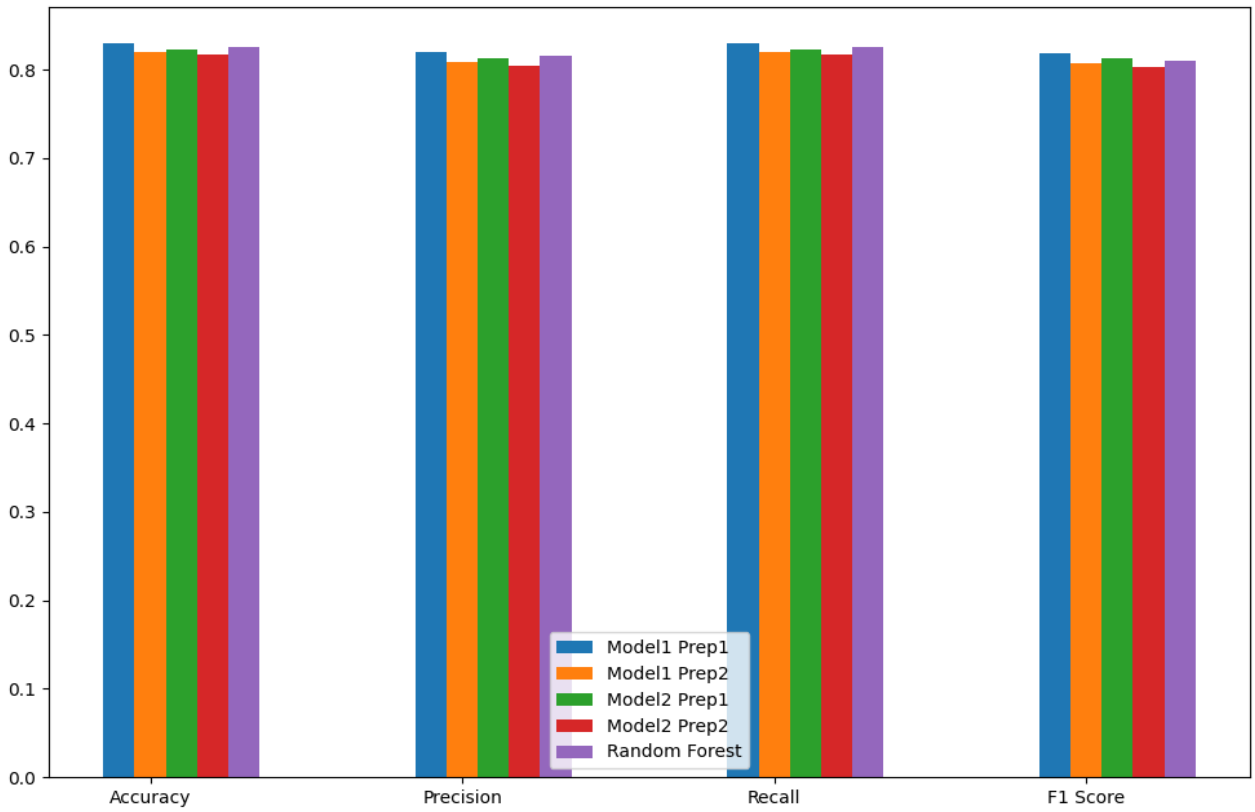

Fig. 2.1. Comparison of Parameters of all models

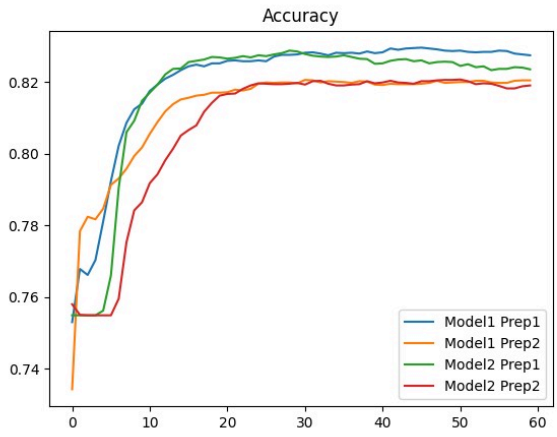

Fig. 2.2. Accuracy of all models

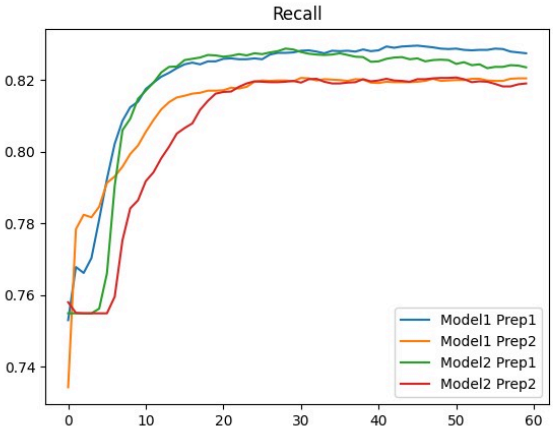

Fig. 2.4. Recall of all models

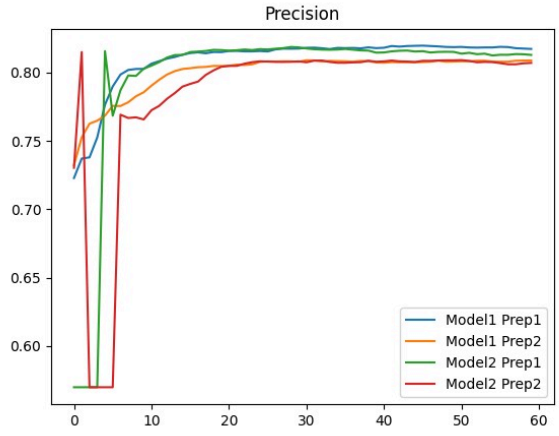

Fig. 2.3. Precision of all models

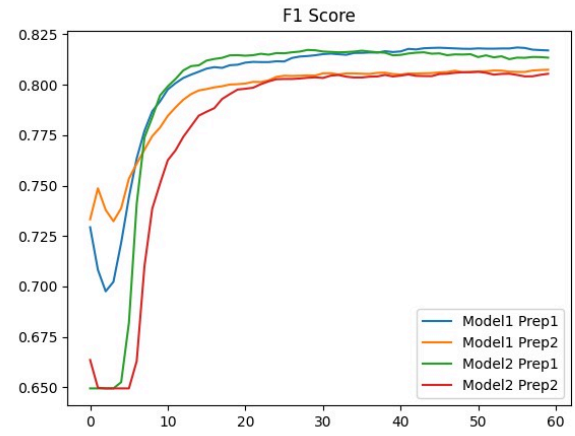

Fig. 2.5. F1 Score of all models

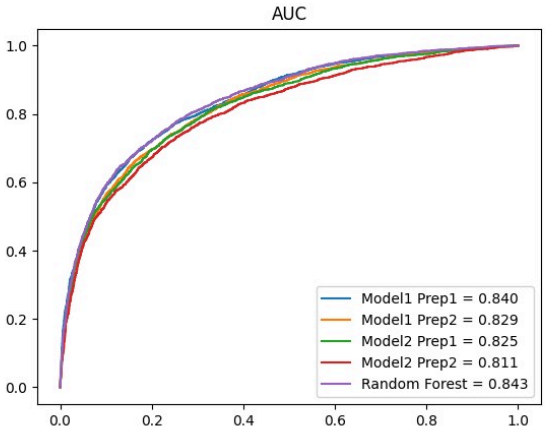

Fig. 3. Area Under Curve(AUC) values of all models

extremely little when compared to our deep learning models. Furthermore, the models that followed preprocess 1 produced the greatest results on all models examined.

V. Conclusion and Future Work

Our best-performing algorithm returns nodes or targets (individuals) who are most likely not vaccinated. Resources allocated to promote vaccination should then be targeted to this individual on the grounds that such promotion would benefit both the individual and their surrounding community. To maximize the value of these resources, our model can identify people who have most likely already been vaccinated, directing vaccine promotion efforts to all individuals in the broader community who have not been classified as positives in our model. To maximize the transmission of influence through vaccination campaigns in all communities, we must create a strategy to maximize the effectiveness of targeted promotions. One way we offer is to use Influence Maximization on the targets predicted by our model. One of the methods offered would be to use Influence Maximization on social networks with the projected goals.

Social networks have an important role in the diffusion of knowledge and influence, raising awareness among individuals. Influence Maximization is used to identify a small group of important persons (as seeds) who can eventually influence as many people as feasible in a social network using a specific influence cascade model. In this instance, individuals who are likely to be vaccinated can be chosen as seeds to affect the other unvaccinated individuals. We will be conducting deeper research into influence maximization frameworks, performance, challenges and directions on healthcare social networks. We will also be testing our approaches using many innovative visualizations to provide better insights [16-19].

There are numerous variations of the impact maximization problem that can be used to fulfil various real-world objectives. A number of cascade models are been presented to model information propagation. To handle the influence maximization problem effectively and efficiently, some greedy algorithms and heuristic methods are proposed. More research is needed to determine which model of Influence Maximization is the greatest fit for our suggested strategy.

References

[1] E. Dubé, C. Laberge, M. Guay, P. Bramadat, R. Roy, and J. A. Bettinger, "Vaccine hesitancy," *Human Vaccines & Immunotherapeutics*, vol. 9, no. 8, pp. 1763–1773, Aug. 2013, doi: 10.4161/hv.24657.

[2] P. Lydon, B. Schreiber, A. Gasca, L. Dumolard, D. Urfer, and K. Senouci, "Vaccine stockouts around the world: Are essential vaccines always available when needed?," *Vaccine*, vol. 35, no. 17, pp. 2121–2126, Apr. 2017, doi: 10.1016/j.vaccine.2016.12.071.

[3] B. Y. Lee, L. E. Mueller, and C. G. Tilchin, "A systems approach to vaccine decision making," *Vaccine*, vol. 35, pp. A36–A42, Jan. 2017, doi: 10.1016/j.vaccine.2016.11.033.

[4] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient algorithms for influence maximization in social networks," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 577–601, Sep. 2012, doi: 10.1007/s10115-012-0540-7.

[5] B. Greenwood, "The contribution of vaccination to global health: past, present and future," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1645, pp. 20130433–20130433, May 2014, doi: 10.1098/rstb.2013.0433.

[6] S. Inampudi, G. Johnson, J. Jhaveri, S. Niranjan, K. Chaurasia, and M. Dixit, "Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination," *Communications in Computer and Information Science*, pp. 139–150, 2021, doi: 10.1007/978-981-16-0401-0_11.

[7] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

[8] R. Hariharan *et al.*, "An Interpretable Predictive Model of Vaccine Utilization for Tanzania," *Frontiers in Artificial Intelligence*, vol. 3, Oct. 2020, doi: 10.3389/frai.2020.559617.

[9] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272.

[10] Dick, K., & Nordstrom, A. (2016). Identifying unvaccinated individuals in Canada: A predictive model. *arXiv preprint arXiv:1607.08656*.

[11] S. Chen, J. Fan, G. Li, J. Feng, K. Tan, and J. Tang, "Online topic-aware influence maximization," *Proceedings of the VLDB Endowment*, vol. 8, no. 6, pp. 666–677, Feb. 2015, doi: 10.14778/2735703.2735706.

[12] Rabadiya, K., Makwana, A., Jardosh, S., & Changa, I. C. (2017). Performance analysis and a survey on influence maximization. In *International conference on telecommunication, power analysis and computing techniques-2017. At Bharath University, Chennai*.

[13] I. Litou and V. Kalogeraki, "Influence Maximization in Evolving Multi-Campaign Environments," *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, doi: 10.1109/bigdata.2018.8622591.

[14] 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, Nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[15] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," *IEEE Xplore*, Mar. 01, 2016. https://ieeexplore.ieee.org/abstract/document/7506650.

[16] M. Sathiyanarayanan, C. Turkay, and O. Fadahunsi. "Design of Small Multiples Matrix-based Visualisation to Understand E-mail Socio-organisational Relationships." In 2018 10th International Conference on Communication Systems & Networks (COMSNETS). 2017.

[17] M. Sathiyanarayanan,, and N. Burlutskiy. "Design and evaluation of euler diagram and treemap for social network visualisation." In 2015 7th international conference on communication systems and networks (COMSNETS), pp. 1-6. IEEE, 2015.

[18] M. Sathiyanarayanan,, and D. Pirozzi. "Spherule diagrams with graph for social network visualization." In 2016 8th International Conference on Communication Systems and Networks (COMSNETS), pp. 1-6. IEEE, 2016.

[19] M. Sathiyanarayanan,, and D. Pirozzi. "Social network visualization: Does partial edges affect user comprehension?." In 2017 9th international conference on communication systems and networks (COMSNETS), pp. 570-575. IEEE, 2017.