A Novel Recommendation System for Vaccines Using Hybrid Machine Learning Model

Nishant Singh Hada, Sreenu Maloth, Chandrashekar Jatoth, Ugo Fiore, Sangeeta Sharma, Subrahmanyam Chatharasupalli, and Rajkumar Buyya

1 Introduction

Advancement in medical science over the centuries has gifted humanity with a crucial boon. Vaccines are one such offering that has brought a much-needed change in the medical field. Both non-infectious and infectious diseases are now within the realm of vaccinology [1]. The journey from the discovery of the first vaccine by Edward Jenner in 1796 [2] to the continuous work being done by medical researchers over the continents on COVID-19 vaccines is remarkable. The virus disturbed the world economies and impacted millions of people. Luckily, with constant development, trials and testing by the R&D departments, several vaccines are in use currently to

N. S. Hada · S. Maloth · S. Sharma
National Institute of Technology Hamirpur, Hamirpur, HP 177005, India
e-mail: hadanis.singh@gmail.com

S. Maloth
e-mail: sreenu@nith.ac.in

S. Sharma
e-mail: sangeetas@nith.ac.in

C. Jatoth (✉)
National Institute of Technology Raipur, Raipur, CG 492010, India
e-mail: chandrashekar.jatoth@gmail.com

U. Fiore
Federico II University of Naples, Naples, Italy
e-mail: ugo.fiore@unina.it

S. Chatharasupalli
Union Public Service Commission, New Delhi, India
e-mail: subrahmanyamch.1981@gov.in

R. Buyya
The University of Melbourne, Parkville VIC 3010, Australia
e-mail: rbuyya@unimelb.edu.au

D. S. Sisodia et al. (eds.), *Machine Intelligence Techniques for Data Analysis and Signal Processing*, Lecture Notes in Electrical Engineering 997,
https://doi.org/10.1007/978-981-99-0085-5_35

protect us from COVID-19. The first mass vaccination programme started in early December 2020 [3]. WHO issued an Emergency Use Listing for the Pfizer COVID-19 vaccine (BNT162b2), two versions of the AstraZeneca/Oxford COVID-19 vaccine and the vaccine Ad26.COV2.S, developed by Janssen (Johnson & Johnson) [3].

Not just COVID-19, but for other diseases and outbreaks there have been multiple vaccines from different manufacturers [16]. The efficacy rate of vaccines varies from region to region and from person to person because of substantial variation in how an individuals' immune response to the vaccine. The study conducted by Zimmerman et al. [4] investigates and talks about various factors that influence humoral and cellular vaccine responses in humans. Some of these factors are age, sex, gestational age, extrinsic factors like medical history, allergies, pre-existing immunity, infections and antibiotics. There are the vaccine factors as well such as vaccine type, product, adjuvant and dose that governs how effective a vaccine will be on the host. Studying all these factors can help medical officers in suggesting the right manufacturer of vaccine that shows the highest efficacy based on the previous records of patients having similar factors to the current host.

Over the last decade, we saw various recommendation systems for drugs and medicines based on different algorithms and techniques [5–11]. The common goal behind the development of these systems is to build a platform that helps the medical officers in decision-making. Considering all the factors quantitatively can be a tedious process which these systems can perform swiftly and help the doctors in making more concise decisions. This field of intelligent recommendation systems is unexplored for vaccines. In our paper, we propose an unprecedented recommendation system for vaccines that produce decisions after considering host-based and vaccine-based factors. This is a new step towards helping doctors in suggesting the right vaccine for the patient.

In our system, we consider data points like age, sex, medical history, allergies of the host and recovery rate, death rate, after vaccination symptoms of the vaccine to make a recommendation. The system stands upon a score-based algorithm that utilizes machine learning to recommend the right vaccine.

The rest of the paper is organized as follows. Section 2 contains the background literature of previously developed recommendation systems for drugs. We brief about the preliminaries in Sect. 3. Section 4 gives an overview of the recommendation system is provided. Section 5 displays the varying experimental results based on changes made in the input. In Sect. 6, the paper is concluded by mentioning the future scope of improvement in the work.

2 Related Work

In the literature, we have various drug recommendation systems that with the help of leading algorithms and techniques like machine learning consider all the factors quantitatively and suggest the right drug. Stark et al. [5] presented an approach for a drug recommendation system using Neo4J, a graph database with high scalability.

The system considered the individual features of the patient and assigned scores to the drugs. The drugs with the highest relevance scores were recommended. This would help physicians to know which drug fits the patient best based on related factors. In 2019, Stark et al. described and compared various recommendation systems based on various features [7]. Hossain et al. in 2020 implemented a drug recommender system that applied sentiment analysis on drug reviews to generate ratings on drugs. They carried out the rating generation using decision tree, K-nearest neighbours and linear support vector classifier algorithm. Finally, linear support vector classifier was selected for rating generation to obtain a good trade-off among model accuracy, efficiency and scalability and the hybrid model for recommendation [6]. Garg also proposed recommendation system based on sentiment analysis [10].

Bao and Jiang in 2016 implemented a recommendation system for medicines using data mining models like ID3 decision tree algorithm, BP neural network and SVM. Finally, the SVM model was selected to obtain a good trade-off. They also proposed a mechanism to ensure the diagnosis accuracy and service quality and showed that their results had excellent accuracy [8]. Abbas et al. in 2020 proposed a drug supply chain management and recommendation system based on blockchain and machine learning. The N-gram and LightGBM models were used to recommend the medicines. These models were trained on the publicly available drug reviews dataset provided by the UCI [9]. Yong et al. in 2020 developed a system based on blockchain and machine learning to support vaccine traceability and smart contract functions [15]. Chen et al. in 2018 proposed a diagnosis system for diseases and also a system to recommend treatment. They introduced density-peaked clustering analysis algorithm to cluster the disease symptoms and perform association analyses based on D-D and D-T rules. They achieved high performance and low latency using the parallel solution provided by Apache Spark [11].

3 Proposed System

In Sect. 2, we discussed various recommendation system for drugs. In this paper, we introduce a novel recommendation system for vaccines that considers age, sex, allergies and medical history to recommend a vaccine to a subject.

3.1 Dataset Collection

In our system, we use the VAERS dataset that is a national early warning system to detect possible safety problems in U.S. licensed vaccines. Healthcare professionals and vaccine manufacturers are required to report adverse events that come to their attention. We use those adverse effects to recommend the best possible vaccine for a patient depending on his age, sex, allergies and medical history. We get a tremendous amount of information from the dataset and have to preprocess it to extract the

Table 1 Selected attributes for the system

Attribute name	Attribute meaning
vaers_id	Unique ID
Recvdate	Received date
age_yrs	Age of patient
Sex	Sex of patient
History	Medical history of the patient
Allergies	Allergies of the patient
symptom_text	Post-vaccination symptoms
Died	Post-vaccination death
Recovd	Post-vaccination recovery
vax_type	Disease type
vax_manu	Vaccine variant

important information for our system. The next section explains the preprocessing of data.

3.2 Data Preprocessing and Analytics

To create an efficient system, we extract the important attributes from the dataset shown in Table 1. In our system, we have three preprocessing stages. In the first stage, we remove data rows where the vaccination information is not available for any user. In the second stage, we remove entries where vaccine manufacturer or type is unknown. In the third stage, we restructure the dataset to improve the performance. The dataset contains various vaccines for different diseases. There are various vaccine types for FLU3, flu4, hepa, flun(h1n1), flux(h1n1), chol, COVID-19, etc.

We further in the process to recommend the vaccine, normalize the data to remove any disadvantage or advantage of having more data as compared to another vaccine. After the preprocessing and analysis, we apply our scoring-based algorithm to recommend the vaccine based on the patient's information. This process is explained in the next section.

3.3 Recommendation Methodology

The user is asked for age, sex and type (disease/outbreak type). Medical history is an optional input. Using the input and processed dataset, the algorithm recommends the vaccine by following the steps given in sections below. Various elements of the system are shown in Fig. 1.

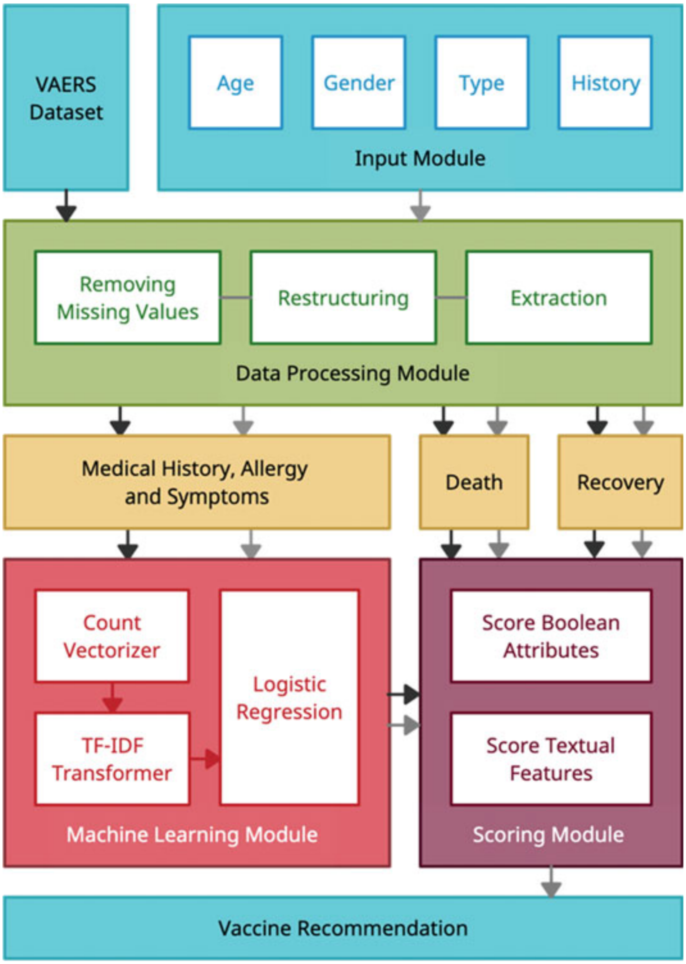


Fig. 1 Vaccine recommendation system design

3.3.1 Data Extraction

We remove all the vaccination data other than the type (disease/outbreak type) that user inputs. The user input age is converted to a range from age −5 to age +5. Next, we remove all the data rows that are of the opposite sex and lie outside the age range.

3.3.2 Scoring Based on Type, Age and Sex

After we get the required data for the recommendation system, we score different vaccines for the given type based on defined criteria. Finally, the vaccine with the highest score is recommended to the user. Each post-vaccination death and recovery contributes -2 and $+1$ to the score of a vaccine, respectively.

3.3.3 Scoring Based on Medical History and Allergies

Based on general idea, if we know that the patient is going through a problem or has allergies, we will not recommend the vaccine that leads to those problems or allergies. Next in the process, all the post-vaccination symptoms and allergies for different vaccines available as textual data in the dataset are fed into the Count Vectorizer [12] and then to TF-IDF transformer [13] to extract the features. These features are unigrams or bigrams based on the vocabulary extracted from user input patient history. Extracting vocabulary from user input limits the number of features extracted and improves the performance. User input medical history is also fed to the same vectorizer and transformer to break down into tokens. As an additional step before vectorization, we strip the accents and remove all the stop words.

After the feature extraction from the textual data and user input medical history, the data is fitted to a logistic regression classifier [14]. The features extracted from the user history are passed to the classifier to generate prediction probabilities. Most probable vaccine will contribute the least score towards vaccine recommendation. The reason behind this statement is that the most probable vaccine is most likely to give negative side effects to the patient. So, the scores generated from the classifier for each vaccine is the inverse of its probability.

3.3.4 Normalization and Recommendation

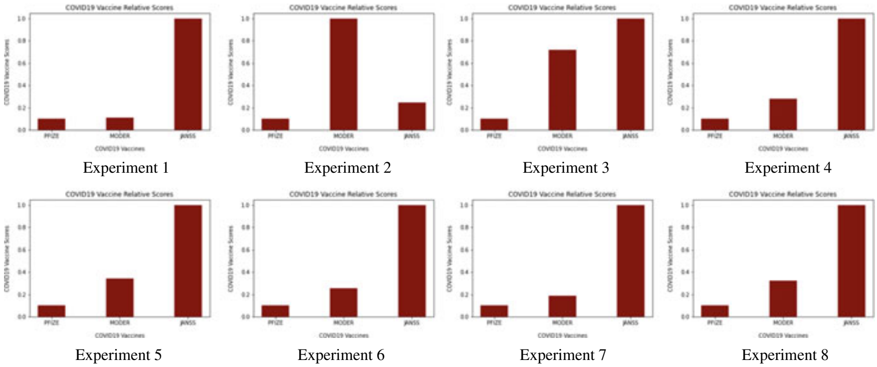
The score for each vaccine calculated by post-vaccination death and recovery is normalized by dividing it with the count of data entries for that particular vaccine. This helps in removing the merits that a vaccine could have had by having more data than others. Find and store the maximum normalized score over all vaccines. The scores for each vaccine calculated after logistic regression classification are also normalized by bringing them in range [0, 1] and then multiply by the maximum stored earlier. Finally, the two scores are added for each vaccine and are brought again to the range [0.1, 1] by rescaling. Finally, the vaccine with 1.0 score is recommended by the system for the particular patient.

4 Performance Evaluation

The results were calculated on a computer with a 1.8 GHz Dual-Core Intel Core i5 processor and 8 GB 1600 MHz DDR3 RAM. The results were produced by creating simulated patient information. For each experiment, a single input was varied to

Table 2 Normalized scores of COVID-19 vaccines in experiments 1–8

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8
Pfizer\BioNTech	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Moderna	0.1128	1.0	0.7196	0.2815	0.3427	0.2537	0.1889	0.3230
Janssen	1.0	0.2445	1.0	1.0	1.0	1.0	1.0	1.0



Experiment 1

Experiment 2

Experiment 3

Experiment 4

Experiment 5

Experiment 6

Experiment 7

Experiment 8

Fig. 2 Visualization of normalized scores of COVID-19 vaccines in experiments 1–8

observe the change in vaccine scores. A total of 16 experiments were conducted on COVID-19 and FLU3 vaccines for eight different kinds of simulated patient's data.

Experiments 1–8 were conducted for the COVID-19 vaccine, and the normalized scores for the different vaccines are shown in Table 2. The graphical representation of the results is shown in Fig. 2.

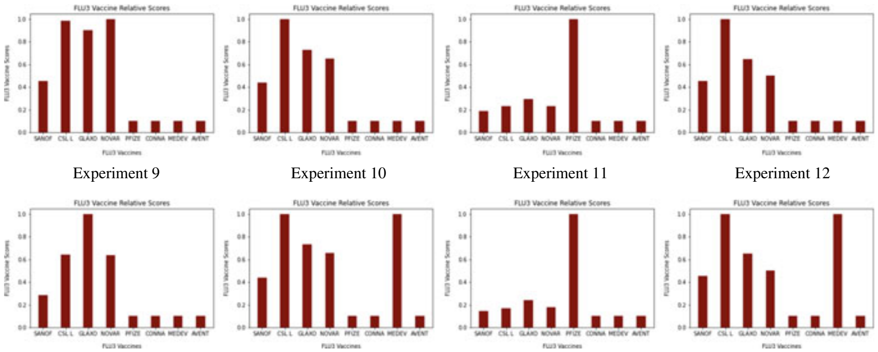
- Experiment 1: age = 32, sex = M. Result = Janssen.
- Experiment 2: age = 32, sex = F. Result = Moderna.
- Experiment 3: age = 52, sex = M. Result = Janssen.
- Experiment 4: age = 52, sex = F. Result = Janssen.
- Experiment 5: age = 32, sex = M, history = 'I have problem with penicillin also I usually have high headache and mild fever'. Result = Janssen.
- Experiment 6: age = 32, sex = F, history = 'I have problem with penicillin also I usually have high headache and mild fever'. Result = Janssen.
- Experiment 7: age = 52, sex = F, history = 'I have problem with penicillin also I usually have high headache and mild fever'. Result = Janssen.
- Experiment 8: age = 52, sex = M, history = 'I have problem with penicillin also I usually have high headache and mild fever'. Result = Janssen.

Experiments 9–16 were conducted for the FLU3 vaccine, and the normalized scores for the different vaccines are shown in Table 3. The graphical representation of the results is shown in Fig. 3.

- Experiment 9: age = 32, sex = M. Result = Novartis Vaccines and Diagnostics.

Table 3 Normalized scores of FLU3 vaccines in experiments 9–16

	Exp. 9	Exp. 10	Exp. 11	Exp. 12	Exp. 13	Exp. 14	Exp. 15	Exp. 16
Sanofi Pasteur	0.4541	0.4382	0.1895	0.4519	0.2873	0.4389	0.1448	0.4519
CSL Limited	0.9844	1.0	0.2328	1.0	0.6421	1.0	0.1671	1.0
GlaxoSmithKline Biologicals	0.9015	0.7287	0.2928	0.6477	1.0	0.7342	0.2393	0.6518
Novartis Vaccines and Diagnostics	1.0	0.6537	0.232	0.5007	0.6392	0.6567	0.1764	0.5042
Pfizer\Wyeth	0.1	0.1	1.0	0.1	0.1	0.1	1.0	0.1
Connaught Laboratories	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Medeva Pharma, Ltd	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.9997
Aventis Pasteur	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1



Experiment 9

Experiment 10

Experiment 11

Experiment 12

Experiment 13

Experiment 14

Experiment 15

Experiment 16

Fig. 3 Visualization of normalized scores of FLU3 vaccines in experiments 9–16

- Experiment 10: age = 32, sex = F. Result = CSL Limited.
- Experiment 11: age = 52, sex = M. Result = Pfizer\Wyeth.
- Experiment 12: age = 52, sex = F. Result = CSL Limited.
- Experiment 13: age = 32, sex = M, history = ‘I have problem with penicillin also I usually have high headache and mild fever’. Result = GlaxoSmithKline Biologicals.
- Experiment 14: age = 32, sex = F, history = ‘I have problem with penicillin also I usually have high headache and mild fever’. Result = CSL Limited.
- Experiment 15: age = 52, sex = M, history = ‘I have problem with penicillin also I usually have high headache and mild fever’. Result = Pfizer\Wyeth.
- Experiment 16: age = 52, sex = F, history = ‘I have problem with penicillin also I usually have high headache and mild fever’. Result = CSL Limited.

5 Conclusions and Future Work

Recommendation systems are becoming a boon for the medical officers and are helping in considering factors that were difficult to consider by humans while suggesting a vaccine. The system is based on scoring that is supported by machine learning modules containing Count Vectorizer, TF-IDF transformer and N-grams. Our system is the first of its kind that recommends vaccines based on the patient's age, sex, medical history and allergies. These are one of the most important factors that decide how a vaccine will perform after entering the patient's body. The score-based algorithm displays the score of not just the recommended vaccine but also provides the results for other vaccines to allow medical officers and other researchers to compare the vaccines for different patients. In the results, we demonstrated the working of our system for COVID-19 and FLU3 vaccines. We can clearly observe the difference in scores generated by the system for different patient data. As a future work, we will increase the dataset involving vaccination data from various other countries. We will collaborate with pharmaceutical companies and laboratories to test the performance and validity of our system and further improve our algorithm in terms of accuracy and recommendation results.

References

1. Plotkin S (2005) Vaccines past, present and future. Nat Med 11(4):S5–S11
2. Stewart A, Devlin P (2016) The history of the smallpox vaccine. J Infect 52(5):329–334
3. WHO Coronavirus disease (COVID-19): vaccines. https://www.who.int/news-room/q-a-detail/coronavirus-disease-(covid-19)-vaccines. Last accessed 2021/04/17
4. Zimmermann P, Curtis N (2019) Factors that influence the immune response to vaccination. Clin Microbiol Rev 32(2):e00084-e118
5. Stark B, Knahl C, Aydin M, Samarah M, Elish K (2017) BetterChoice: a migraine drug recommendation system based on Neo4J. In: Proceedings of the 2nd IEEE international conference on computational intelligence and applications, Beijing, China, pp 382–386
6. Hossain MD, Azam MS, Ali MJ, Sabit H (2020) Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning. In: Proceedings of the 2020 emerging technology in computing, communication and electronics, Bangladesh, pp 1–6
7. Stark B, Knahl C, Aydin M, Elish K (2019) A literature review on medicine recommender systems. Int J Adv Comput Sci Appl 10(8)
8. Ni J, Huang Z, Cheng J, Gao S (2021) An effective recommendation model based on deep representation learning. Inform Sci 542:324–342. ISSN 0020-0255, https://doi.org/10.1016/j.ins.2020.07.038
9. Abbas K, Afaq M, Ahmed Khan T, Song WC (2020) A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry. Electronics 9(5):852
10. Garg S (2021) Drug recommendation system based on sentiment analysis of drug reviews using machine learning. In: 2021 11th international conference on cloud computing, data science and engineering (confluence), pp 175–181. https://doi.org/10.1109/Confluence51648.2021.9377188
11. Chen J, Li K, Rong H, Bilal K, Yang N, Li K (2018) A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inf Sci 435:124–149

12. Count Vectorizer. https://www.studytonight.com/post/scikitlearn-countvectorizer-in-nlp. Last accessed 2021/04/17
13. TF-IDF Transformer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html. Last accessed 2021/04/17
14. Wikipedia n-Grams. https://machinelearningmastery.com/logistic-regression-for-machine-learning/. Last accessed 2021/05/04
15. Yong BB, Shen J, Liu X, Li F, Chen H, Zhou Q An intelligent blockchain-based system for safe vaccine supply and supervision. Int J Inform Manage 52:102024
16. Top 10 Vaccine Manufacturers in the World 2020. https://blog.technavio.org/blog/top-10-vaccine-manufacturers. Last accessed 2022/02/15