2023 International Conference on Computer Communication and Informatics (*ICCCI*), Jan. 23 – 25, 2023, Coimbatore, INDIA

2023 International Conference on Computer Communication and Informatics (ICCCI) | 979-8-3503-4821-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCI56745.2023.10128441

COVID 19 post-vaccination adverse effects prediction with supervised machine learning models

Vaishali Ravindranath [1*]
Research Scholar, Dept. of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
rvaishali4@gmail.com
ORCID: 0000-0002-5992-0958

Sarojini Balakrishnan [2]
Assistant Professor (SS) in Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India
balakrishnan.sarojini@gmail.com
ORCID: 0000-0002-7740-7058

Abstract—**This electronic Pharmacovigilance with AI helps to trace possible adverse events among the vaccinated population. The symptom pattern discovery from the vaccinated population obtained through post-vaccination surveys provides insights for medical practitioners to study the possibility of adverse events among vulnerable populations. This work contains postvaccination survey data from an Indonesian national referral hospital with 840 instances, 6 categorical input features, and 15 binary target attributes. As there were multiple symptoms as a target, multi-target classification algorithms experimented on the dataset. The inadequate sample size resulted in poor performance of the algorithm. To improve model prediction performance, the target was converted into binary format. The population that exhibited at least one symptom is considered symptomatic by the binary classification models. The supervised machine learning model of test train split (80% - 20%) produced 89% accuracy with a decision tree classification algorithm in the classification of symptomatic or non-symptomatic patients.**

Keywords- Machine learning; Pharmacovigilance; Supervised algorithm; multi-target classification

I. INTRODUCTION

Covid 19 pandemic brought the necessity of emergency vaccination among the world population [1,2]. Vaccinated populations are prone to a variety of symptoms post-shot [3]. Medical practitioners obtain surveys on a large scale to study the behavior and symptom pattern post-vaccination. Manual interpretation of patterns among a large group is highly timeconsuming and could lead to bias. Sometimes hidden patterns are missed in the process of manual pattern discovery. Hence, autonomous pattern discovery methods that can discover hidden patterns are considered potential solutions to predict adverse events. In the process of predicting adverse events, they adopt pharmacovigilance techniques in pattern discovery [4]. The process involves a clinical survey, statistical inference, and pattern recognition. Collection and evaluation of such samples on a large scale become complex and the time incurred in the process may outdate the prediction. Hence, an automation pipeline is required to perform data collection, transformation, and learning simultaneously in a cloud environment. As the number of samples increases, the patterns become more complex for manual pattern discovery. In modern forecasting and prediction systems, the advantage of employing machine learning models has simplified the process involved. Machine learning methods can capture hidden relationships among variables and form distinct patterns to predict the outcome with a lesser error rate. The quality of a machine learning algorithm relies completely on the quality of the dataset. The cleaner the data, the more distinct the patterns would be. Factors that could affect the quality of a dataset are missing values, outliers, class imbalance, insignificant attributes, data collection anomalies, etc. Covid 19 vaccination dataset is a real-time dataset that could possess all the above data anomalies. Particularly data collection anomalies are highly impactful on the results. These anomalies could happen due to differences in equipment, differences in lab setup, atmosphere, and other potential variations in the medium of a collection. In such cases, discovering patterns from such datasets becomes non-reliable. Hence, it is important to consider the data, that was collected from the same data collection center. To ensure environmental anomalies are avoided.

979-8-3503-4821-7/23/$31.00 ©2023 IEEE

II. PROBLEM STATEMENT

In Pharmacovigilance, AI is deployed in drug discovery and drug performance analysis due to its ability to bring everything related to the ETL process into one stage. Covid 19 tracking dashboards trained with historical records are the validation source for vaccine performance post-shots. Hence, it is necessary to study the practical difficulties involved in building a machine-learning solution to the real-time vaccine data that are prone to high levels of noise.

In this research work, the performance of machine learning methods was tested in three criteria 1. Multi-target classification 2. Binary classification 3. Binary classification with feature selection. Further results are evaluated to generate insights.

III. LITERATURE REVIEW

Before you begin to format your paper,

The emergence of covid 19 as a deadly disease, urged researchers to find potential ways to find reliable solutions to prevent, cure and track the disease. To track vulnerable populations with machine learning methods [1] had done an elaborate review on machine learning and applications in covid tracking. The work discussed various machine learning methods including Deep learning methods that could serve the purpose of early diagnosis. Similar work was done in [2] where three machine learning models: Neural Networks, Random Forest, and Classification regression Trees performance were reviewed in various environments including SPSS 25 and Javabased statistical processor (JASP). The work claims machine learning models exhibit poor target prediction due to laboratory anomalies in the data collection phase. A review on Pharmacovigilance and opportunities in the field are reviewed in [4]. The importance of data collection in machine learning is emphasized in the above reviews. There are multiple works available in the literature in the Covid-19 disease tracking.

Vaccination drives happened in many countries and recorded post-vaccination symptoms like Fever, body pain, headache, etc. The pattern of symptoms is recorded and analyzed in multiple works. In 2021, an Indonesian National referral hospital released a covid 19 post-vaccination dataset [3]. Data analysis was performed on the vaccinated population to find statistical patterns. In the data on the post-vaccination symptoms, there are multiple target attributes which makes it a multi-output classification problem. In literature, multi-output classification models were used in problems of Predictive maintenance [5]. The same strategy could be tried on the postvaccination dataset [3]. The machine learning models could follow strategies like tree-based learning, Ensemble learning, Vector based learning, Lazy learning, probabilistic learning, and Function based learning. However, if the target groups undergo data imbalance, the results could get affected.

IV. DATASET

The dataset [3] used in this work is the survey data obtained from the hospital staff of Dr. M. Djamil Hospital Padang, a national referral hospital in Indonesia. The collection of data happened through an online questionnaire mode, assessing COVID-19 vaccination side effects between 9th February 2021 and 13th February 2021.

Table 1. *The attribute information of the post-vaccination dataset*

	Attributes	***Count***
Inputs	*'Sex'* *'Age'* *'Professions'* *'Education'* *'Living_Area'* *'Symptoms_time'*	*6*
Symptoms	*'Swelling'* *'Redness'* *'Itching'* *'Fever'* *'Headache'* *'Muscle_Pain'* *'Tiredness'* *'Coughing'* *'Diarrhoea'* *'Nausea & vomiting* *'Breathlessness'* *'Joint_Pain'* *'Fainted'* *'Anaphylactic_Reaction'* *'Tingling'*	*15*

Among the vaccinated population, the most common symptoms reported were muscle pain (40%), tiredness (36%), and headache (22%). Other symptoms like Swelling, Coughing, Tingling, Diarrhea, Redness, Fever, Nausea and vomiting, Joint Pain, Breathlessness, Anaphylactic Reaction, and Fainting were noticed in moderate to lower numbers. In the obtained samples, no one reported 'itching' as a symptom in post-vaccination. As the 'itching' attribute does not have any impact on the data, it is ignored by the target. The statistics discussed are visualized in Fig.1.

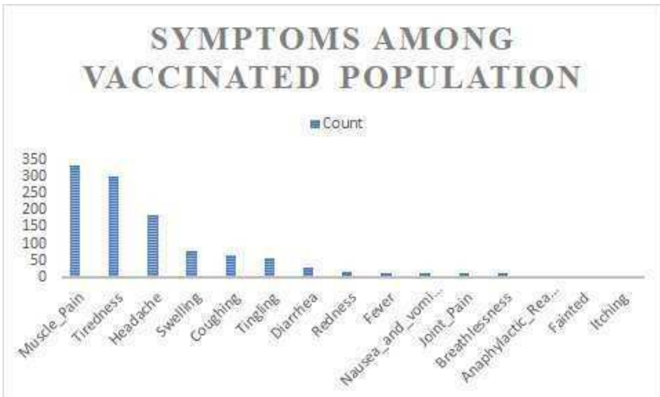

Fig.1. *Symptom distribution among the vaccinated population*

Among the vaccinated population in the survey, 62.14% of them had at least one symptom post-vaccination. 37.86% of the population realized no symptoms at all.

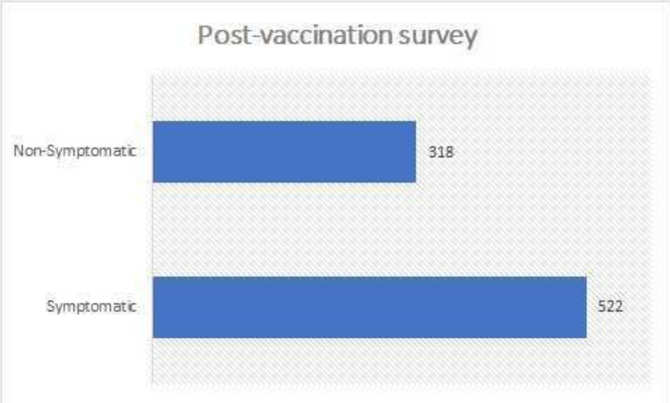

Fig.2 *Distribution of symptomatic and non-symptomatic population*

Table 2. *Cohorts' analysis on combinations of symptoms*

Cohorts	Count	%
Only Muscle_Pain	110	21.07
Only Tiredness	65	12.45
Muscle_Pain+Tiredness	45	8.62
Headache+Muscle_Pain+ Tiredness	42	8.05
Only Headache	19	3.64
Headache+Tiredness	17	3.26
Headache+Muscle_Pain	13	2.49
Only Swelling	12	2.30
Muscle_Pain+Swelling	11	2.11

As shown in Table 2, the symptom combination has a huge data imbalance. Due to this approach, this problem as a multitarget classification becomes the worst solution. To overcome the class imbalance due to the scattering of symptoms, a generic binary target attribute was created with conditions 'symptomatic' or non-symptomatic.

Symptomatic: At least one symptom is true

Non-Symptomatic: All symptoms are false

The distribution of the symptomatic and non-symptomatic population is displayed in Fig.2.

V. PROPOSED SOLUTION

As the dataset has a huge class imbalance, a three-phase modeling experiment was proposed.

In the first phase, the modeling experiment was conducted on the whole data sample with 840 instances with 101 unique symptom combinations with the multi-target classification model.

As the results were not in acceptable ranges, a second phase experiment was conducted on a filtered population with the condition (a minimum of 10 counts should be there per combination). This resulted in a sample size of 470 instances with 48 unique combinations. It means out of 101 total unique symptom combinations, only 48 symptom combinations had a count greater than or equal to 10.

The result of Phase 2 was better than phase 1, still, due to the huge number of classes in the target, the patterns were not captured well by the model. Hence, multi-target classification models were not performing as expected. Another experiment was conducted by creating a sample with symptoms that had >= 25% samples each. Though the performance improved in certain training o

To overcome the issue in Phase 3, the target attribute of the dataset was feature engineered into a binary column with the symptomatic and non-symptomatic assumptions as discussed in section IV. The performance evaluation of the results of the models was detailed in section VI.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed three-phased model experiments were carried out in a Python environment in Google cloud. The metrics of evaluation used in this work include Accuracy, Precision, recall, and Delta. The formulas of the metrics are listed as follows

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \qquad (1)$$

In short, accuracy is the measure of finding rightly predicted instances among the total number of instances. Higher the accuracy, the better the model. However, considering only accuracy as a measure for validation is not recommended.

Abbreviations of the above notations are

TP - True Positive

FP- False Positive

TN- True Negative

FN- False Negative

Delta = Train accuracy - Test accuracy (2)

Delta is the difference between Train and test accuracy scores. Lesser the difference, the better the model provided accuracy is above the threshold.

Precision = TP/ (TP+FP) (3)

Precision is the score of correctly classified positive instances among total positive instances as classified by the model.

Recall = TP/(TP+FN) (4)

The recall is the score of correctly classified positive instances among actual positive instances in the dataset.

Table 3. *Phase 1 Prediction results with all symptoms*

Metric= Accuracy	**Target: All Symptoms (15)**	
Algorithm	**Training (80%)**	**Testing (20%)**
Decision Tree	**0.55**	0.14
Gaussian Naïve Bayes	0	0
Random Forest	0.54	0.2
K-Nearest Neighbour	0.26	0.2
MLP Perceptron	0.32	0.13

When all 15 symptoms are considered as targets, the accuracy of prediction in both test and train delivered underfit models. The reason behind the worst performance was the huge class imbalance that led to poor pattern discovery in the training phase. Poor training models lead to poor training results. In other words, there are scanty samples available in the dataset for each unique combination which leads to poor training and testing accuracy. The models are in the underfit category due to this reason. To improve the accuracy of the model, the class imbalance among unique combinations should be addressed. On the other hand, the symptoms that are rare among the population can be trimmed to prevent diversion from broad patterns.

To improve the accuracy of symptom prediction, the feature importance approach is deployed to choose features that have an influential position in prediction. A statistical filterbased approach that picks features with at least 10 values is used to pick features.

Swelling	32
Headache	110
Muscle Pain	247
Tiredness	195
Coughing	18
Tingling	10

A second experiment was carried out on the selected symptoms listed above. The results are analyzed to check for any significant improvement.

Table 4. *Phase 2a - Prediction results with selected symptoms with at least 10 counts*

Metric=Accuracy	**Target: Selected Symptoms (>=10 counts each)**	
Algorithm	**Training (80%)**	**Testing (20%)**
Decision Tree	0.58	0.21
Gaussian Naïve Bayes	0.01	0.02
Random Forest	0.57	0.22
K-Nearest Neighbour	0.31	0.15
MLP Perceptron	0.40	0.21

Though the result shows improvement in numbers, the model remains underfit. Hence, the feature importance was refined with a criteria minimum of 25% of representation, and the modeling experiment was carried out.

Table 5. *Phase 2b- Prediction results with selected symptoms with at least 25% samples.*

Metric=Accuracy	**Target: Selected Symptoms (>= 25% samples each)**	
Algorithm	**Training (80%)**	**Testing (20%)**
Decision Tree	0.70	0.23
Gaussian Naïve Bayes	0.07	0.05
Random Forest	0.70	0.30
K-Nearest Neighbour	0.40	0.29
MLP Perceptron	0.58	0.27

The performance of the models with a minimum 25% class representation was better than n=15 and the Sample size >=10. Still, the delta between the training and testing performance shows a huge negative number.

Due to these practical difficulties in real-time datasets, multi-target classification remains a challenge for modeling. Hence, patterns discovered from such models are non-reliable.

To overcome the data imbalance problem, the multi-targets are converted into binary targets and a modeling experiment was carried out for various split ratios of random state 42.

Table 6. *Performance of the binary classification models are tabulated - Accuracy & Delta*

Split Ratio	**90%-30%**			**80%-20%**			**70%-30%**			**60%-40%**		
Algorithm	**Train**	**Test**	**Delta**	**Train**	**Test**	**Delta**	**Train**	**Test**	**Delta**	**Train**	**Test**	**Delta**
DT	**0.87**	**0.80**	0.07	**0.89**	0.77	0.12	**0.88**	0.70	0.18	0.88	0.75	0.13
GNB	0.40	0.40	0.00	0.39	0.43	-0.04	0.40	0.36	0.04	0.41	0.38	0.03
RF	0.87	0.80	0.07	0.89	0.76	0.13	0.87	0.73	0.14	0.88	0.80	0.08
KNN	0.82	0.70	0.12	0.82	0.77	0.05	0.81	0.70	0.11	0.82	0.77	0.05
MLP	0.82	0.80	**0.02**	0.88	0.78	0.10	0.80	**0.86**	**-0.06**	0.82	0.87	-0.05

DT-Decision tree, GNB- Gaussian Naive Bayes, RF-Random Forest, KNN- K-Nearest Neighbours, and MLP - Multi-Layered Perceptron.

In the above table, better accuracy was produced by MultiLayered perceptron, Decision Tree, and Random Forest. The performance has far improved due to the handling of the class imbalance in this experiment. Compared to previous experiments, the delta between the train and test scores remains lower and the accuracy measure is comparatively higher. Hence, the model created in the final experiment is taken as a better fit. The recall and precision scores of the final experiment were studied for further validation. The results were tabulated as follows

Table 7. *Performance of the binary classification models are tabulated - Precision & Recall*

Split	90-30%		80-30%		70-30%		60-40%	
Algorithm	**precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
DT	0.83	0.83	0.9	0.64	0.8	0.63	0.86	0.72
GNB	0	0	0	0	0	0	0	0
RF	0.83	0.83	0.85	0.79	0.79	0.79	0.77	0.84
KNN	0.7	0.71	0.92	0.86	0.77	0.73	0.78	0.88
MLP	0.9	1	0.92	0.86	0.94	0.84	0.91	0.88

The precision and recall rates analysis help to check the false-positive rates among the classified results. From Table 7, it was observed that the MLP Neural network model produced higher precision and recall with the least false-positive rates. The Gaussian Naive Bayes model underperforms all other models with a 0 score for precision and recall. It means none of the instances is correctly classified as positive due to bias toward the Negative class.[11] The models taken in these experiments work with default parameters provided by sklearn. Tuning the hyperparameters could improve the precision and recall scores.

VII. CONCLUSION

In real-time scenarios like vaccination adverse effects where multiple symptoms were recorded as targets, it is not possible to construct a single classification model to predict all symptoms at one go due to a huge imbalance in the unique patterns' population. The experiments conducted to support the claim, conclude that huge pattern imbalance delivered worst-fit models which were not reliable for real-time prediction. Alternatively, building multiple models for multiple symptoms would increase the number of prediction models, and hence, model management becomes a complex task. To ease the model management and simultaneously improve the accuracy of prediction, in the proposed method, a generic binary target attribute was created, which in turn improved the accuracy of prediction with better fit and better delta values. However, scrutinizing the data further with attribute selection, the test of significance, and better data wrangling methods could further improve the model.

In future work, an additional layer could be added to the model architecture, where the symptomatic population could be further studied for generating personas and tracking patterns of symptoms.

Ethical Statement

The dataset used in this work was obtained from the article titled 'Survey data of COVID-19 vaccine side effects among hospital staff in a national referral hospital in Indonesia'. We thank the authors of the survey article for listing the dataset in the annexure.

Acknowledgment

Author 1 thanks Dr. Sarojini Balakrishnan for her valuable mentorship.

References

[1] Clement, J. C., Ponnusamy, V., Sriharipriya, K. C., & Nandakumar, R. (2021). A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis. IEEE reviews in biomedical engineering, 15, 325-340.

[2] Assaf, D., Gutman, Y. A., Neuman, Y., Segal, G., Amit, S., GefenHalevi, S., ... & Tirosh, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Internal and emergency medicine, 15(8), 1435-1443.

[3] Djanas, D., Martini, R. D., Putra, H., Zanir, A., & Nindrea, R. D. (2021). Survey data of COVID-19 vaccine side effects among hospital staff in a national referral hospital in Indonesia. Data in brief, 36, 107098.

[4] Kompa, B., Hakim, J. B., Palepu, A., Kompa, K. G., Smith, M., Bain, P. A., ... & Beam, A. L. (2022). Artificial intelligence based on machine learning in pharmacovigilance: a scoping review. Drug Safety, 45(5), 477-491.

[5] Last, M., Sinaiski, A., & Subramania, H. S. (2010, March). Predictive maintenance with multi-target classification models. In Asian Conference on Intelligent Information and Database Systems (pp. 368377). Springer, Berlin, Heidelberg.

[6] Suthaharan, S. (2016). Support vector machine. In Machine learning models and algorithms for big data classification (pp. 207-235). Springer, Boston, MA.

[7] Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.

[8] Wijayanto, U. W., & Sarno, R. (2018, September). An experimental study of supervised sentiment analysis using gaussian naive Bayes. In 2018 International Seminar on Application for Technology of Information and Communication (pp. 476-481). IEEE.

[9] Biau, G., & Scornet, E. (2016). A random forest-guided tour. Test, 25(2), 197-227.

[10] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbour rule. IEEE Transactions on Systems, Man, and Cybernetics, (4), 325-327.

[11] Kruse, R., Mostaghim, S., Borgelt, C., Braune, C., & Steinbrecher, M. (2022). Multi-layer perceptrons. In Computational Intelligence (pp. 53124). Springer, Cham.

[12] Shraboni Rudra, Minhaz Uddin, & Mohammed Minhajul Alam. (2019). Forecasting of Breast Cancer and Diabetes Using Ensemble Learning. International Journal of Computer Communication and Informatics, 1(1), 1-5