CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG - Brasil

Signs and symptoms analysis of SARS-CoV-2 virus infection waves

Análise de sinais e sintomas da infecção pelo vírus SARS-CoV-2

Análisis de síntomas de la infección por el virus SARS-CoV-2

Felipe Cassemiro Ulrichsen[1], Alexandre Costa Sena[2], Luís Cristóvão Porto[3], Karla Figueiredo[2]

1 PhD student, Matematical and Statistics Institute, State University of Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil.
2 PhD/Associate Professor, Matematical and Statistics Institute, State University of Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil.
3 PhD/Associate Professor, Piquet Carneiro University Polyclinic (PPC), State University of Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil

Autor correspondente: MSc. Felipe Cassemiro Ulrichsen
E-mail: felipeulrichsen@eng.uerj.br

Abstract

Objective: Develop a Data Mining and Machine Learning methodology for COVID-19 diagnosis. Method: Create diagnostic models, evaluate differences in symptoms between pandemic waves. Results: Diagnose symptomatic SARS-CoV-2 infection. Conclusion: Highlight the effectiveness of the methodology in pandemic management.

Keywords: SARS-CoV-2; COVID-19; Machine Learning.

Resumo

Objetivo: Desenvolver metodologia de Mineração de Dados e Aprendizado de Máquina para diagnóstico de COVID-19. Método: Criar modelos de diagnóstico, avaliar diferenças nos sintomas entre ondas pandêmicas. Resultados: Diagnosticar infecção sintomática pelo SARS-CoV-2. Conclusão: Destacar a eficácia da metodologia na gestão pandêmica.

Palavras-chave: SARS-CoV-2; COVID-19; Aprendizado de Máquina;

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

Resumen

Objetivo: Desarrollar una metodología de Minería de Datos y Aprendizaje Automático para el diagnóstico de COVID-19. Método: Crear modelos de diagnóstico, evaluar diferencias en síntomas entre olas pandémica. Resultados: Diagnosticar infección sintomática por SARS-CoV-2. Conclusión: Destacar la eficacia de la metodología en la gestión pandémica.

Descriptores: SARS-CoV-2; COVID-19; Aprendizaje Automático;

Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, was first reported in Wuhan, Hubei province, China, in 2019 [1]. By March 2020, the WHO (World Health Organization) declared it a global pandemic due to its highly contagious nature and the emergence of new variants, including mutations in the spike protein [2]. Testing became crucial to curb the spread, particularly among symptomatic individuals or those in contact with infected individuals. In the absence of tests, many doctors rely on clinical signs and symptoms for diagnosis, highlighting their importance in disease assessment.

Machine Learning (ML) has seen widespread application across various sectors, including healthcare, since the mid-20th century [3] with its use intensifying due to the digital storage of patient data. ML offers numerous advantages in healthcare, allowing the manipulation of large volumes of variables quickly and safely [4]. However, many ML algorithms are perceived as black boxes, posing challenges for acceptance within the healthcare domain.

In this context, the objective of this work is to investigate and evaluate methodologies and models based on ML for the diagnosis of COVID-19, based only on the signs and symptoms of patients, to assist health professionals that can also be extended to other outbreaks or pandemic. The main contributions of this paper are: (i) evaluate the use of ML techniques to infer the diagnosis of COVID-19 considering different waves of contagion; (ii) increase the quality and explainability of COVID-19 diagnoses, to help

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

healthcare professionals decide the best treatment for patients; (iii) indicate the most prevalent signs and symptoms in the different waves of contagion; (iv) evaluate and identify the variation of signs and symptoms in the different waves of contagion.

The following paragraphs describe relevant related work found in the literature. In a study conducted in Jordan [5], an online form was used to collect data for developing a diagnostic tool for COVID-19 using a Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM). The attributes used were signs and symptoms, gender and age. The study also used X-ray images in the inference which provided an accuracy above 90% for both models. Another study, carried out in England [6], used data from more than one million participants who took part in the REACT-1 survey on SARS-CoV-2 infection. The data used were: symptoms, results of the RT-PCR tests, and results of the genetic analysis of the virus SARS-CoV-2, which were divided into two groups. The LASSO algorithm was used to perform the analyses. The study obtained 72% of sensitivity and 64% of specificity for the first group and 74% of sensitivity and 64% specificity for the group with the Alpha variant.

In turn, the study carried out in the United Kingdom [7], used a cell phone application for users to inform the signs and symptoms after the third day of the first symptom and the presence of pre-existing diseases. The Hierarchical Gaussian model, Bayesian framework and Logistic Regression were used. There was a division into groups: health professionals or not, gender, age, body mass index and date of onset of symptoms. There was no data equalization, but an attempt was made to reduce the imbalance between negatives and positives in the database using bootstrapping. The best result was obtained in the groups of health workers using the hierarchical Gaussian model, achieving a sensitivity of 76%. A similar study in England[8] used data from the United Kingdom and the United States of America that were obtained through a cell phone application in which patients reported symptoms, BMI (body mass index), sex, pre-existing diseases, demographic data and the result of the RT-PCR test. The Logistic Regression algorithm was used to make the inferences and the data were not equalized. Instead, the inputs were divided into groups

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

by sex, age, and BMI. The study had a mean sensitivity of 65% and mean specificity of 78% for UK data and mean sensitivity of 66% and a mean specificity of 83% for USA data.

Differently from the works presented previously, in the work proposed in this article, only signs and symptoms described by health professionals and the COVID-19 test results are used as attributes. Moreover, data were divided and analyzed in waves to assess the impact on the model. This work is also the only one that equalizes the data between positive and negative for COVID-19 tests. Although the real world is not equalized, evaluating models with balanced data is important to eliminate possibly biased results. In addition, this work also analyzes the influence of signs and symptoms on the result generated by the machine learning model through explainable and Shapley techniques.

Methods

Machine Learning algorithms in Data Mining extract insights by evaluating, preprocessing, and optimizing databases, including organizing, identifying outliers, normalizing, and selecting variables. The methodology in this study comprises three phases: data acquisition, preprocessing and application, and analysis of classification results using Machine Learning [3].

Five algorithms were chosen: Random Forrest, Multi-Layer Perceptron, XGBoost, Logistic Regression, and the Shapley Additive Explanation. They were selected to predict COVID-19, considering the specifics of each wave, and for explainable analysis. The selection allows the exploration of both linear and nonlinear characteristics of the database, examining each algorithm individually and in an ensemble.

Random Forest (RF) is an ensemble decision tree algorithm that combines multiple classifiers by using bootstrap aggregating (bagging) to reduce variance while maintaining bias [9]. To mitigate the correlation between decision trees, Breiman [10] introduced random attribute selection for tree construction, decreasing the correlation among trees.

Multi-Layer Perceptron (MLP) constructs Neural Networks (NN) inspired by biological neurons, learning input-output relationships by adjusting synaptic weights based

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

on errors during supervised learning [11]. In turn, Logistic regression generates a linear model from input data to predict values of a categorical variable, typically binary [12]. Finally, Shapley Additive Explanations, introduced by Lundberg and Lee,[13] interprets machine learning model outputs by measuring each variable's contribution to the final result. This approach is crucial for understanding models that lack intrinsic explainability.

All data used in this research refer to the COVID-19 tests carried out at the Piquet Carneiro Polyclinic, which is part of the health complex of the State University of Rio de Janeiro. This study was conducted following ethical principles outlined in the Declaration of Helsinki and was approved by the Pedro Ernesto University Hospital Ethical Committee (CAAE: 30135320.0.0000.5259). Moreover, the software developed is registered with the National Institute of Industrial Property through process number BR512023000197-0 and is available for academic use.

The patients' self-described symptoms on the form were consolidated, and laboratory tests considered for diagnosis included RT-PCR, Rapid Antibody Test (RT-antibody), and Rapid Antigen Test (RT-antigen).

Results and Discussion

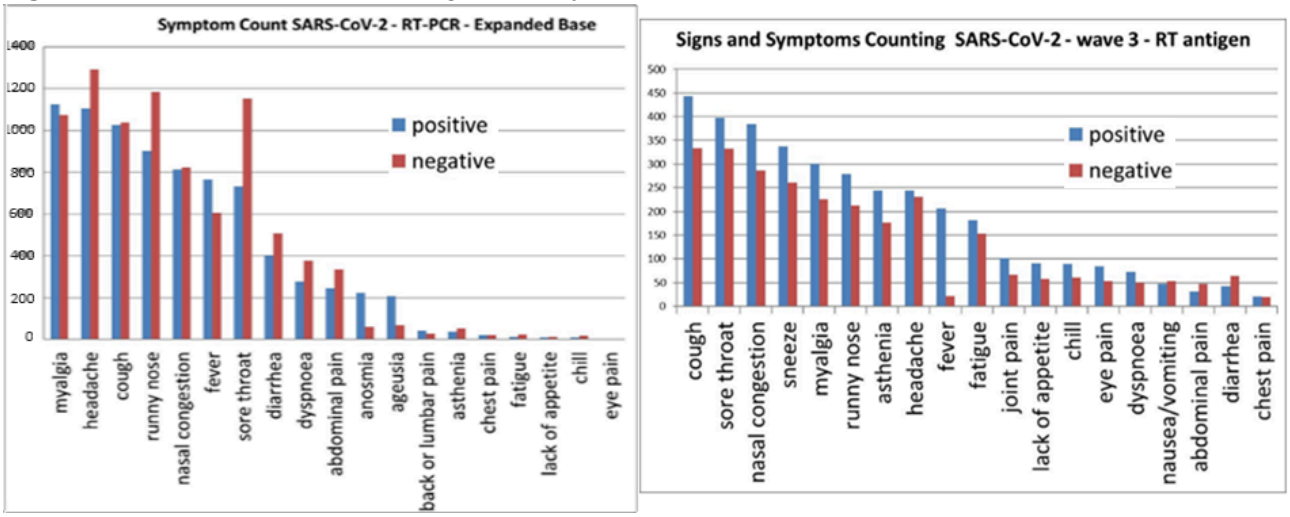
Initially, an analysis of the waves of contagion in the city of Rio de Janeiro was conducted. The start and end dates of each wave were determined using open data obtained from the Rio de Janeiro Municipal Health Department. The start and end dates of the first wave were 03/18/2020 and 06/18/2020 and the start and end dates of the second wave were 10/18/2020 and 2/18/2021. In turn, the start and end dates of the third wave were 12/25/2021 and 2/25/2022.

After evaluation by health specialists, 19 prevalent signs and symptoms were selected for the 1st and 2nd waves, and 19 for the 3rd wave (though not exactly the same signs and symptoms). These 19 signs and symptoms were considered broadly, encompassing even those with low representation in the database.

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

Figure 1 illustrates the prevalence of 19 signs and symptoms in patients with positive and negative results from the RT-PCR test during the 1st and 2nd waves (left graph) and even displays the primary signs and symptoms reported by patients during the 3rd wave (right graph), all tested with the Rapid Antigen, using nasal swab samples.

Figure 1 – Prevalence of the 19 signs and symptoms.



For a comprehensive analysis, data were categorized based on different waves and types of tests (RT-PCR, RT-Antigen, and RT-antibody) used to diagnose patients as positive or negative. This categorization yielded 10 datasets, listed in the first column of Table 1, as in subsequent tables labeled "Data group".

Table 1 – Number of records for each database: using approximately 90% for training and validation, and 10% for testing.

Data group	Training and validation	Test
1 - RT-PCR (1st wave + 2nd wave)	2436	272
2 - RT-PCR 1st wave	1228	138
3 - RT-PCR 2nd wave	839	93
4 - RT-antigens 3rd wave	1080	120
5 - RT-antibody (1st wave + 2nd wav)	1990	222
6 - RT-antibody 1st wave	1748	196
7 - RT-antibody 2nd wave	67	9
8 - RT-PCR + RT-antibody (1st wave + 2nd wave)	4426	494
9 - RT-PCR + RT-antibody 1st wave	2926	332
10 - RT-PCR + RT-antibody 2nd wave	769	87

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

The patient records with signs and symptoms related to different test types (RT-PCR, RT-antibody, and RT-antigen) and their results (negative and positive) across the first, second, and third waves were imbalanced. To ensure unbiased model performance, data equalization was conducted by randomly undersampling the class with the higher number of records to achieve balance. Symptoms reported during the waves were encoded as binary attributes ("1" for presence, "0" for absence), as was the diagnosis column ("1" for positive test results, "0" for negative). Subsequently, approximately 10% of positive and negative cases were set aside as a test set, while the remainder was allocated for model training and validation. This procedure was applied to all datasets, with Table 1 displaying the data distribution for each phase (i.e., Training and Validation or Test).

A thorough attribute evaluation was conducted individually for each dataset. A sensitivity analysis type evaluation was performed, involving the removal of each sign and symptom to assess the impact on model performance. This analysis, restricted to data groups 2, 3, and 4, involved sequentially removing attributes and evaluating the improvement in validation results. Through exploratory analysis, variable selection, and optimization of hyperparameters using a search grid, the most effective signs and symptoms for enhancing models were identified. This attribute selection process was integrated with a search for optimal hyperparameters using cross-validation, which divided the dataset into 5 equal parts for training and validation. Models were trained on four parts and validated on the fifth, repeated five times with each part used once for validation. The model with the best average accuracy was selected for each dataset, with accuracy being the primary metric, alongside precision, recall, F1, AUC, sensitivity, and specificity.

After assessing the impact of removing each remaining attribute individually, it was determined that further removal of attributes from the databases was not appropriate following the exclusion of "nasal congestion" and "chest pain" for the first and second waves, and solely "chest pain" for the third wave. Interestingly, the signs and symptoms that produced the best validation results were consistent for the first and second waves,

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

whereas, for the third wave, certain signs and symptoms aiding the model's classification of SARS-CoV-2 infection differed.

Table 2 displays the best results achieved for the 10 datasets (Data group). Each database had its hyperparameters determined through a search algorithm based on validation metrics.

The MLP algorithm demonstrated slightly superior results for the RT-PCR 1st wave (Data group 2) compared to RF and RL. In turn, RF and MLP algorithms exhibited similar performance for the RT-PCR 2nd wave database (Data group 3), with RL showing superior results for recall, F1, and AUC metrics. In turn, for RT-antigens 3rd wave (Data group 4), RF and MLP algorithms produced nearly identical metrics, while RL displayed slightly lower results.

Overall, RL outperformed RF and MLP algorithms across all RT-antibody bases (Data groups 5, 6, and 7), particularly achieving 64% accuracy for the base corresponding to patients in the second wave (Data group 7). However, it's essential to note the limited data available for this base, which may have influenced the outcome.

Table 2 – Metrics - average validation errors.

Method	Data group	#records	Accuracy	Precision	Recall	F1	AUC
RF	1	2852	0.65	0.67	0.60	0.63	0.70
	2	1228	0.62	0.61	0.53	0.57	0.66
	3	839	0.70	0.70	0.64	0.67	0.71
	4	1080	0.67	0.67	0.65	0.66	0.68
	5	1990	0.64	0.64	0.53	0.58	0.65
	6	1748	0.64	0.64	0.48	0.55	0.65
	7	67	0.64	0.64	0.67	0.66	0.66
	8	4426	0.65	0.67	0.59	0.62	0.68
	9	2956	0.62	0.64	0.54	0.58	0.66
	10	769	0.69	0.70	0.67	0.68	0.72
MLP	1	2852	0.65	0.67	0.60	0.63	0.70
	2	1228	0.62	0.62	0.61	0.62	0.65
	3	839	0.69	0.70	0.64	0.67	0.71
	4	1080	0.66	0.66	0.68	0.66	0.68
	5	1990	0.64	0.65	0.59	0.62	0.66
	6	1748	0.61	0.66	0.47	0.55	0.65
	7	67	0.62	0.66	0.47	0.55	0.64
	8	4426	0.67	0.65	0.67	0.67	0.68
	9	2956	0.68	0.65	0.68	0.68	0.65
	10	769	0.68	0.66	0.68	0.68	0.61
RL	1	2852	0.65	0.65	0.62	0.62	0.70

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

	2	1228	0.61	0.62	0.54	0.57	0.64
	3	839	0.68	0.66	0.76	0.71	0.74
	4	1080	0.65	0.66	0.61	0.64	0.68
	5	1990	0.62	0.63	0.59	0.61	0.65
	6	1748	0.61	0.64	0.51	0.56	0.64
	7	67	0.57	0.65	0.50	0.53	0.51
	8	4426	0.63	0.64	0.60	0.62	0.67
	9	2956	0.61	0.66	0.43	0.52	0.64
	10	769	0.66	0.67	0.66	0.66	0.72

The MLP algorithm demonstrated superior performance across all metrics when considering the combined databases of patients tested by RT-PCR and RT-antibody, except for the second wave where RF achieved the best outcome. Upon conducting the Mann-Whitney test on the average accuracy values, p-values > 0.05 were obtained for comparisons between MLP and RF, MLP and RL, and RF and RL, indicating that the results are not significantly different.

Table 3 shows that the Random Forest model achieved 79% accuracy in the RT-PCR 2nd wave dataset (data group 3) with 85% precision. Sensitivity and specificity were 76% and 82% respectively, indicating its ability to correctly identify positive and negative cases.

Table 3 – Metrics - average tests errors (Random Forrest).

Data group	Accuracy	Precision	Recall	F1	AUC	Sensitivity	Specificity
1	0.64	0.65	0.64	0.64	0.70	0.62	0.66
2	0.65	0.66	0.65	0.65	0.66	0.62	0.69
3	0.79	0.85	0.79	0.79	0.71	0.76	0.82
4	0.72	0.72	0.72	0.71	0.68	0.70	0.73
5	0.59	0.55	0.59	0.59	0.65	0.58	0.60
6	0.62	0.65	0.62	0.59	0.65	0.51	0.67
7	0.44	0.40	0.44	0.50	0.66	0.50	0.40
8	0.60	0.61	0.60	0.60	0.68	0.57	0.64
9	0.62	0.68	0.62	0.62	0.66	0.63	0.64
10	0.60	0.54	0.60	0.59	0.72	0.63	0.67

The exchange of test data between models trained with data from different waves helped in identifying differences among datasets and evaluating their consistency. Therefore, test data from the first symptomatic wave of SARS-CoV-2 were used in models trained with second-wave and third-wave data. Likewise, test data from the second wave

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

were used in models trained with first-wave and third-wave data. Also, test data from the third wave were used in models trained with first-wave and second-wave data. The results can be seen in Table 4, where the first line for each wave shows the results using data from the same wave and the other two lines with distinct waves.

For all wave models, there is a decrease in sensitivity and recall when using test data from other waves. In the 1st wave model, results from the second wave generally outperform those from the first wave tests, sensitivity and recall are exceptions. The 2nd wave model achieves excellent results with its own test set but performs less optimally when tested with data from other waves. Notably, sensitivity and recall drop to 38% when using third-wave data. The 3rd wave model's tests with first and second-wave datasets reveal a significant decrease in most metrics compared to using its own data.

Table 4 – Wave models tested with data from1st, 2nd and 3rd wave test sets - RF.

Wave	Data group	Accuracy	Precision	Recall	F1	AUC	Sensitivity	Specificity
1st	2	0.65	0.66	0.65	0.65	0.65	0.62	0.69
	3	0.70	0.71	0.58	0.70	0.69	0.58	0.80
	4	0.61	0.61	0.40	0.61	0.59	0.40	0.78
2nd	3	0.79	0.85	0.79	0.79	0.79	0.76	0.82
	2	0.62	0.61	0.60	0.62	0.62	0.60	0.65
	4	0.62	0.64	0.38	0.62	0.60	0.38	0.81
3rd	4	0.72	0.72	0.70	0.72	0.71	0.70	0.73
	2	0.58	0.58	0.46	0.58	0.58	0.46	0.70
	3	0.63	0.62	0.51	0.63	0.62	0.51	0.74

Figures 2 display the SHAP distribution for each record in their respective datasets. In these graphs, blue represents low values and red represents high values, with positive red values indicating that the presence of the symptom aids in identifying symptomatic infection by SARS-CoV-2. The distribution of points along the axis reveals their impact on positive or negative classification.

The charts in Figure 2 show that for the best model, there is a difference in signs and symptoms between the different waves. A significant difference in signs and symptoms can be observed in the third wave in particular. Although fever remains the primary indicator of virus presence, nasal congestion took second place in the third wave.

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

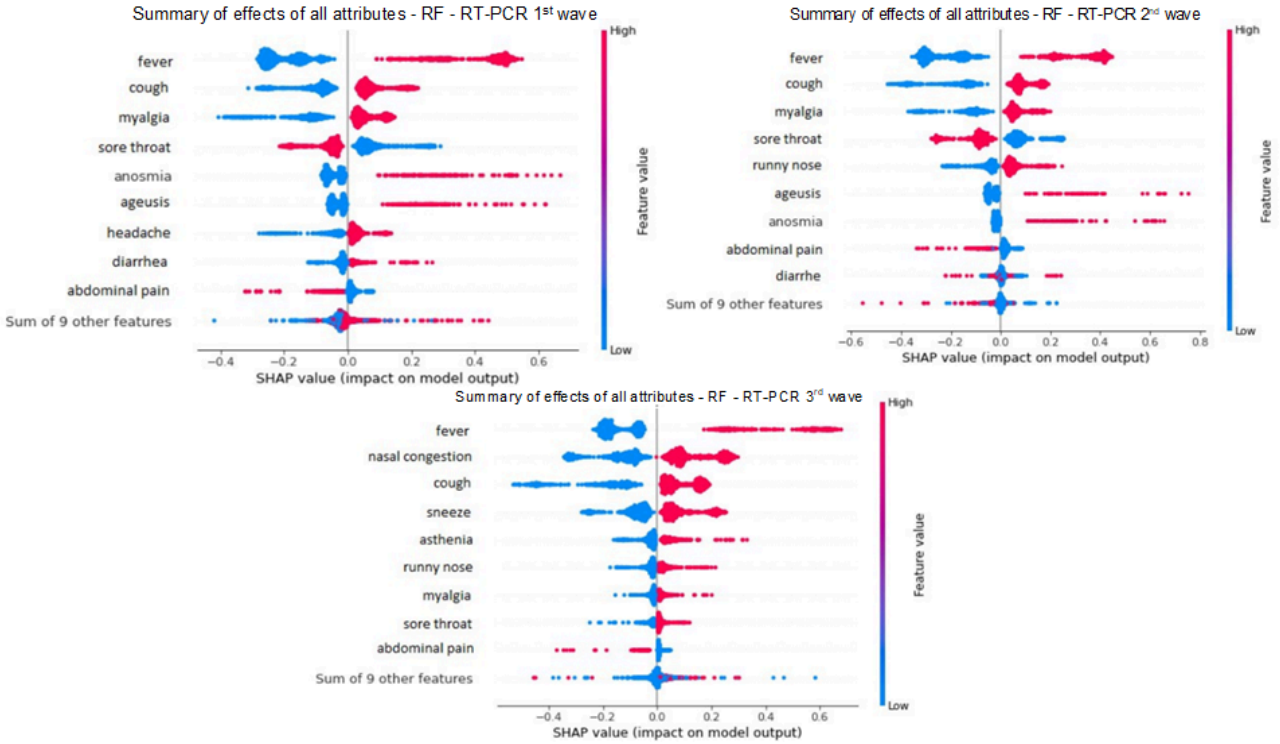
Additionally, there was a change in the importance of other attributes; the higher they are on the chart, the more they assist the model in classification.

Discussion

Data were analyzed and divided into waves. In addition to signs and symptoms, other attributes were explored to improve the sensitivity and specificity of the models. The gender attribute was irrelevant for predicting the outcome of symptomatic infection with the SARS-CoV-2 virus. The information if the patient had contact with confirmed cases, markedly worsened the results. Analyzing the data from the third wave forms, it can be seen that many patients confirmed contact after a date after the onset of signs and symptoms, indicating that contact was not the cause of the infection. Data such as place of residence and age could not be used, as most patients did not fill in this information. In the database for the third wave, there was already information about the vaccine. However, as the vast majority of patients were vaccinated (97.5% vaccinated), it was not possible to use this information as an attribute to assess this base specifically.

Figure 2 – Summary of effects of all attributes - RF - RT-PCR 1st wave

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339



It can be seen, through the results generated by the ML models, that the RT-PCR exam is much superior in detecting symptomatic infection by the SARS-CoV-2 virus; this is demonstrated by the metrics of the models when using the exams as labels separately. The best result of the model that considered the RT-PCR exam used data from the second wave, reaching 79%, 76%, and 82%, respectively of accuracy, and sensitivity specificity, while the best model using the RT of Antibodies was 62% accuracy, with only 51% sensitivity and 67% specificity, considering the data from the first wave. Another aspect analyzed was the division into waves. The models were trained with data from the first and second waves and data outside a specific wave, that is, the period between waves. In this way, much more data is obtained; however, when mixing wave data, the models obtained inferior results; this fact motivated the testing of the models by training and validating them with data from a specific wave and testing with data from another wave, which was presented in the results section.

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339

Conclusion

This study aimed to develop a methodology for diagnosing symptomatic SARS-CoV-2 infection, assisting in patient triage and social distancing efforts. The models enabled the identification of main signs and symptoms across different waves of infection. Several models achieved over 70% accuracy, sensitivity, and specificity, indicating successful achievement of the objectives. The methodology proved effective in studying symptom prevalence across different infection waves. It was observed that using data from the same wave improved model performance, suggesting that the Omicron variant may have different symptom profiles. These models can aid in isolating potentially infected patients more efficiently and affordably than traditional tests, with RT-PCR and RT Antigen tests proving more reliable than RT antibody tests. Future work will involve gathering data on variants and vaccines received to identify symptom patterns. This methodology may also be applied to other diseases like Dengue or Zika, with the potential inclusion of additional data types such as imaging or lab tests using multimodal approaches for diagnosis and outcome investigation.

Acknowledgements

This work was carried out with the support of the The National Council for Scientific and Technological Development - CNPq 308717/2020-1, the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001 and CAPES 88881.506840/2020-01.

References

1. Andersen KG, Rambaut A, Lipkin WI. The proximal origin of SARS-CoV-2. Nature Medicine. 2020; 26:450-452.
2. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. ZGenomics. 2020; 112(5):3588–3596.
3. Ming-Syan C, Jiawei H, Yu PS. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering. 1996; 8(6):866-883.

4. Addis TR. Towards an ”expert” diagnostic system. ICL Technical Journal. 1956; 1:79-105.

5. Fayyoumi E, Idwan S, AboShindi H. Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan. (IJACSA) International Journal of Advanced Computer Science and Applications. 2020; 11:122–126.

6. Elliott J, Whitaker M, Bodinier B, et al. Predictive symptoms for COVID-19 in the community: REACT-1 study of over 1 million people. PLoS Med. 2021; 18(9).

7. Canas LS, Sudre CH, Pujol JC, et al. Early detection of COVID-19 in the UK using self-reported symptoms: a large-scale, prospective, epidemiological surveillance study. Lancet Digit Health. 2021;

8. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nature Medicine. 2020; 26:1037–1040.

9. Breiman L. Random Forests. Machine Learning. 2001; 45(1):5-32.

10. Breiman L. Bagging predictors. Machine Learning. 1996; 24:123-140.

11. Haykin, S. Neural Networks and Learning Machines. Third Edition, Pearson Education, Inc., McMaster University, Hamilton, 2009

12. Brant R. Digesting logistic regression results. The American Statistician. 1996; 50(2):117-119.

13. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017.

J. Health Inform. 2024, Vol. 16 Especial - ISSN: 2175-4411 - jhi.sbis.org.br
DOI: 10.59681/2175-4411.v16.iEspecial.2024.1339