Statistical and Predictive Analysis to Identify Risk Factors and Effects of Post COVID-19 Syndrome

Milad Leyli-abadi*, Sonja van Ockenburg †, Axel Tahmasebimoradi*, Jean-Patrick Brunet‡
* IRT SystemX, Palaiseau, France
† Department of Endocrinology, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
‡ Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
email: {milad.leyli-abadi}@irt-systemx.fr

***Abstract*—Some corona virus disease 2019 (COVID-19) symptoms can persist for months after infection, leading to what is termed Post COVID-19 condition. Factors such as vaccination timing, patient characteristics, and pre-existing conditions may contribute to the prolonged effects and intensity of Post COVID-19 condition. Each patient, based on their unique combination of factors, develops a specific risk or intensity of Post COVID-19 condition. In this work, we aim to achieve two objectives: (1) conduct a statistical analysis to identify relationships between various factors and Post COVID-19 condition, and (2) perform predictive analysis of Post COVID-19 condition intensity using these factors. We benchmark and interpret various data-driven approaches using data from the Lifelines COVID-19 cohort. Our results show that Neural Networks (NN) achieve the best performance in terms of Mean Absolute Percentage Error (MAPE), with predictions averaging 19% error. Additionally, interpretability analysis reveals key factors such as loss of smell, headache, muscle pain, and vaccination timing as significant predictors, while chronic disease and sex are critical risk factors. These insights provide valuable guidance for understanding Post COVID-19 condition (PCC) and developing targeted interventions.**

Keywords-Post COVID-19 syndrome; PCC; predictive analysis; Machine learning; Explainability.

I. Introduction

In May 2023, after 3 years of global pandemic, the WHO declared the end of the global Public Health Emergency for COVID-19. Although this indicates an improvement, especially with general access to vaccines, it does not mean the end of the presence and effects of COVID-19 which can now be considered endemic [1]. One lasting effects being post-COVID-19 condition (PCC), which presents by the continuation of physical and cognitive symptoms after recovery from acute COVID-19 [2][3]. PCC prevalence is not exactly known with recent worlwide estimates varying from 6% to 10% lowered from initial WHO estimates of 10 to 20% [4][5]. Many countries are now developing dedicated health care paths for PCC and as such means to identify at risk population would be beneficial for improved early referrals.

Although the condition has been extensively studied, there are still many uncertainties regarding the exact characterization and risk factors associated. One major challenge in studying this subject is the lack of comprehensive data. As an evolving crisis, initial datasets had to be created and collected in real time with limited understanding of the virus and lasting effect. Thus, most data were collected retrospectively from incomplete patient medical files, clinical cohorts of hospitalized patients or patients in dedicated PCC recovery care. However, data suggest that most people affected by PCC were never hospitalized and would not necessarily seek direct care for the condition. Alternatively, there is often limited knowledge of participants' pre-existing conditions, making it hard to verify that persistent symptoms are new and attributable to COVID-19 [2][5].

This study uses a unique dataset collected and maintained by Lifelines that addresses some of these concerns. Lifelines is a multi-disciplinary, prospective cohort study examining the health and health-related behaviors of 167,729 individuals in Northern Netherlands over three generations. It assesses biomedical, socio-demographic, behavioral, physical, and psychological factors.

From April 2020 to November 2022, a COVID-19 specific branch involving 31 questionnaires was sent to Lifelines adult participants without inclusion criteria. Frequency varied from weekly to bi-monthly. 76,503 participants answered at least one questionnaire, with a mean of 13.5 questionnaires (standard deviation 10.5). The cohort's duration and size provide valuable data on pre-existing conditions, control groups, and factors influencing PCC's emergence, evolution, and severity.

A number of studies have explored the use of data-driven approaches to predict and analyze the attributes developing PCC [6][7]. The use of unsupervised clustering on time series of early development of COVID-19 is investigated in [7] that could be predictive of the need for high-level care in individuals more likely to seek medical help. A recent study employed a gradient boosting classifier for diagnosis of PCC [6]. They obtain similar results using a dataset retrieved from a panel of primary care practices in Germany.

The aim of this study is to explore the following critical research question: *"Can specific pre-infection parameters be identified to predict the severity of post-COVID-19 condition?"*. To answer this question, an analysis was performed using machine learning techniques. The ability to predict PCC and identify relevant pre-infection symptoms and risk factors holds significant societal implications, impacting physical and mental health, daily functioning, and productivity. To facilitate this, we introduced the concept of Post-COVID-19 Symptom Intensity (PCSI) as a measure of the persistence and impact of symptoms after COVID-19 infection. As such, a continuous measure of PCC is proposed allowing for a more accurate measure of the impact of the condition compared to the com-

monly used binary definition. Using various machine learning models, we focused on predicting PCSI using demographic and clinical characteristics. This study constitutes the first predictive analysis conducted on Post-COVID-19 Lifeline data through the application of machine learning algorithms. The principal contributions of this work are as follows:

- Conducting a comprehensive statistical analysis to identify influential factors associated with the study of PCC;
- Performing predictive analysis of *Post COVID-19 Symptom Intensity* using data-driven approaches;
- Interpreting and analyzing the impact of diverse variables on *Post COVID-19 Symptom Intensity*, offering valuable information for medical decision-making;
- Developing a Python package [8] for evaluating ML algorithms on health-related (Lifelines) datasets, facilitating reproducibility and further research in the domain.

The remainder of this article is structured as follows. Section 2 describes the data preprocessing steps and provides statistical insights into the dataset. Section 3 presents the methodology for predicting PCSI, along with results and an analysis of key influential factors identified by each model. Finally, Section 4 provides a discussion and concludes the paper.

II. Preprocessing and data analysis

This section presents the data used for the analysis and describes pre-processing steps undertaken to format the data suitably. Additionally, it includes a preliminary statistical analysis to reveal global tendencies.

A. Data description

The dataset comprises two main types of variables:

- Static Variables: These denote fixed attributes of individuals, recorded as single entries in the database. Examples include age, sex, SARS-CoV-2 variant, income, smoking status, overall health status, presence of chronic diseases, vaccination status, and time between vaccination and infection.
- Dynamic Variables: These variables capture the presence and intensity of symptoms at different time intervals (before, during, and after SARS-CoV-2 infection). Symptoms include headache, dizziness, heart or chest pain, lower back pain, nausea, muscle pain, difficulty breathing, feeling warm or cold, numbness or tingling, sore throat, dry or wet cough, fever, diarrhea, loss of smell or taste, and sneezing, among others.

Several challenges emerged while working with the data. Similar to many questionnaire-based datasets, there were considerable amounts of missing or aberrant data. Additionally, since the data was collected during an active epidemic, the scope and phrasing of the questionnaires evolved over time, resulting in inconsistencies. Extensive preprocessing was undertaken to address these issues, standardizing the dataset and ensuring a uniform structure suitable for analysis.

B. Definition of Post COVID-19 symptoms intensity (PCSI)

Post COVID-19 condition is a systemic condition in which individuals experience persistent symptoms following a SARS-CoV-2 infection. While the WHO provides a general definition, it does not specify which symptoms or measurement methods to use [9][10], leading to inconsistencies across studies in terms of time frames, symptom types, and severity criteria. In this study, we adopted the WHO time frame definition: symptoms that cannot be explained by an alternative diagnosis, appearing three months after infection and lasting for at least two months. Symptom selection was based on 10 core PCC symptoms identified in prior research using the same dataset [2].

Symptom intensity was rated on a 5-point Likert scale (1 = not at all, 5 = extremely) based on the participant's experience during the previous seven days (see Figure 1). Symptoms were considered present if rated at least 3 (moderate). Each participant's baseline was defined as the mean intensity of symptoms from all questionnaires completed at least seven days before infection; individuals without such data were excluded.

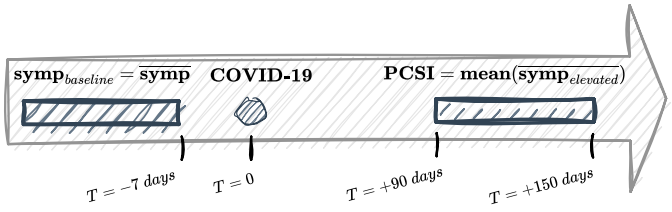

Figure 1. The overall process for defining Post COVID-19 symptom intensity (PCSI) using symptoms (symp) scores. All analyses were centered around the time of the first reported SARS-CoV-2 infection.

PCC was defined as the presence of at least one persistent symptom (mean score $\geq$ 3) between 90 and 150 days post-infection, with an increase of at least one point from baseline.

We further defined a continuous measure, *Post COVID-19 Symptoms Intensity (PCSI)*, as he highest mean score among symptoms meeting the PCC criteria defined above. PCSI preserves symptom severity granularity, facilitating more nuanced modeling and analysis. It supports both statistical and machine learning approaches and can serve as a proxy for the binary PCC definition when needed. For non PCC participant, a proxy was used by taking the value of the symptom with the highest mean score in the 90-150 days post-infection.

C. Data cleaning and preprocessing

The raw data from different questionnaires were organized into multiple tables, each containing information collected at the participant level for specific dates. After cleaning and preprocessing, participants with a sufficient number of shared variables were filtered. This filtering process resulted in the creation of a merged database that consolidated all the necessary information required for the study and analysis. For the predictive analysis, we adopted the steady-state hypothesis, utilizing only the pre-infection period for feature extraction.

TABLE I. POPULATION CHARACTERISTICS. BLUE REPRESENT PROPORTION OVER KNOWN VALUES.

		SARS-CoV-2 positive n=13191		Included n=4657		Excluded n=8534		PCC Positive n=715 | 15.4%		PCC Negative n=3942 | 84.6%	
Characteristics	**Modalities**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Age	18–39	1520	12	411	9	1109	13	61	9	350	9
	40–59	7006	53	2315	50	4691	55	426	60	1889	48
	≥60	4665	35	1931	41	2734	32	228	32	1703	43
Gender	Male	4631	35	1679	36	2952	35	190	27	1489	38
	Female	8560	65	2978	64	5582	65	525	73	2453	62
BMI	<25	5830	44	2111	45	3719	44	276	39	1835	47
	25≤ BMI <30	5173	39	1827	39	3346	39	297	42	1530	39
	≥30	2188	17	719	15	1469	17	142	20	577	15
Chronic disease	None	7948	67	3118	67	4830	68	381	53	2737	69
	One	2212	19	914	20	1298	18	178	25	736	19
	Multiple	1643	14	625	13	1018	14	156	22	469	12
	Unknown	1388	11			1388	30				
Smoking	Yes	1292	10	438	9	854	10	79	11	359	9
	No	11783	90	4219	91	7564	90	636	89	3583	91
	Unknown	116	1			116	1				
Self-assessed health prior to infection	Excellent	1189	11	492	11	697	10	41	6	451	11
	Very good	3886	34	1631	35	2255	34	185	26	1446	37
	Good	5645	50	2302	49	3343	50	406	57	1896	48
	Mediocre/poor	580	5	232	5	348	5	83	12	149	4
	Unknown	1891	14			1891	22				
Educational level	High	4907	38	1035	22	3181	38	272	38	895	23
	Medium	5054	39	1751	38	3303	39	297	42	1454	37
	Low	2777	21	1726	37	1742	21	140	20	1454	37
	Other	305	2	112	2	193	2	12	2	100	3
	Unknown	148	1	33	1	115	1	6	1	27	1
Vaccination prior to infection	Full	6701	57	3149	68	3552	50	417	58	2732	69
	Partial	562	5	0	0	562	8				
	No	4492	38	1508	32	2984	42	298	42	1210	31
	Unknown	1436	10			1436	17				
Variant	Original	2747	21	987	21	1760	21	193	27	794	20
	Alpha	1417	11	190	4	1227	15	40	5	150	4
	Delta	1096	8	444	6	652	8	80	11	364	9
	Omicron	7931	60	3066	66	4865	57	402	57	2662	68
Hospitalization	Yes	190	1	44	1	146	2	15	2	29	1
	No	12663	99	4512	99	8151	98	683	98	3829	99
	Unknown	338	3	101	2	237	3	17	2	84	2

As the result of preprocessing, a total of 4,657 participants were included in this study. Table I illustrates the characteristics of the total population observed (subset of the cohort with a covid-19 diagnosis), included and excluded group (based on missingness of information) and finally the subgroups with positive or negative post-covid assessment. Base characteristics of the included and excluded population are similar. It is to be noted that women account for 73% of the cases while representing 64% of the base dataset. This indicates that women are more likely to be at risk for Post COVID-19 condition than men. Conversely, for low PCC symptom intensities, the proportion of women is smaller.

D. Preliminary statistics

To assess the impact of input variables and investigate potential dependencies between the input variables and the outcome (presence of PCC), we applied two statistical tests. These tests are outlined below:

- *Chi-square test*: This test assesses whether two categorical variables are independent [11] and used to study the relation between two categorical variables, i.e., vaccination and PCSI. By evaluating the p-value obtained from the test statistic at the chosen confidence level, we determine whether to reject the null hypothesis (indepedence) in favor of the alternative hypothesis (dependence). A confidence level of 95% is typically used and the null hypothesis is rejected if $p - value < 0.05$.
- *Cramer's V test*: This test quantifies the strength of association between two categorical variables [12]. A value close to zero indicates a weak dependency, while a value approaching 1 suggests a strong dependency.

Using these tests, we analyzed the influence of vaccination on PC symptom intensity, with the results depicted in Figure 2. This analysis was also conducted for other variables; however, we present only the results for vaccination, as it serves as a crucial preventive measure against COVID-19. To simplify the interpretation, we rounded the PCSI score. From the figure, it is evident that most participants who are fully vaccinated are less likely to experience high levels of PCSI (2,790 out of 3,149 or 88% vaccinated participants report intensity levels 1 or 2). However, due to a lack of representative observations for higher intensity levels, we cannot confidently establish a relationship between vaccination and PCSI for these cases. The Chi-square test statistic ($p < 0.05$) confirms the significance of this relationship, even though the strength of the association is weak (Cramer's $V = 0.072$).

VACCINE	*PC_INTENSITY* 1	2	3	4	5	***Total***
complete vaccin	2514 79.8 % 71 %	276 8.8 % 54.9 %	225 7.1 % 59.7 %	108 3.4 % 54.5 %	26 0.8 % 70.3 %	3149 100 % 67.6 %
no	1028 68.2 % 29 %	227 15.1 % 45.1 %	152 10.1 % 40.3 %	90 6 % 45.5 %	11 0.7 % 29.7 %	1508 100 % 32.4 %
Total	3542 76.1 % 100 %	503 10.8 % 100 %	377 8.1 % 100 %	198 4.3 % 100 %	37 0.8 % 100 %	4657 100 % 100 %

χ^2=81.995 · df=4 · Cramer's V=0.133 · p=0.000

Figure 2. Chi-square test between vaccination and PCSI scores. The test results indicate a significant relationship ($p < 0.05$) between vaccination and PCSI scores.

To further examine the relationships between multiple variables simultaneously, the Multiple Correspondence Analysis (MCA) [13] is used. It allows identification and visualization of underlying structures in a set of nominal categorical data as is the case in this study. It can be seen as the categorical equivalent of principal component analysis (PCA), projecting data points into a low-dimensional Euclidean space where each axis represents a component, with the corresponding variance explained in percentage. Figure 3 depicts the obtained results.

The MCA plot reveals that high PCSI (5) is linked to the presence of chronic diseases and poorer overall health. Additionally, it appears that women are more likely to experience higher PCSI compared to men. The original SARS-CoV-2 variant does not show a strong correlation with PCC, suggesting a lower risk. Lastly, individuals in better general health seem to have a reduced risk of developing PCC.

III. Methodology and Results

In this section, we outline an evaluation pipeline designed to select and benchmark various predictive models using the data obtained from the pre-processing stage. The goal of this study is to predict the target variable, y, which represents the intensity of Post COVID-19 condition. The intensity is modeled as a continuous variable ranging between 1 (low intensity) and 5 (high intensity). Given its continuous nature, the problem is formulated as a regression task, where the

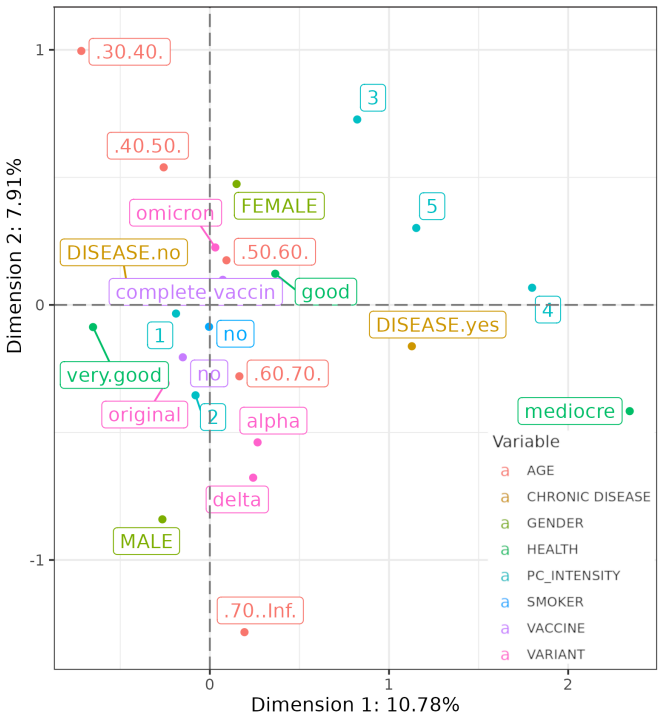

Figure 3. Multiple Correspondence Analysis considering static and vaccination variables. The PCSI variable is discretized (1-5 in clear blue).

models aim to approximate the mapping $f : \mathbf{X} \to y$, with $\mathbf{X} \in \mathbb{R}^p$ being the set of p explanatory variables (features). The overall structure of the proposed pipeline is illustrated in Figure 4.

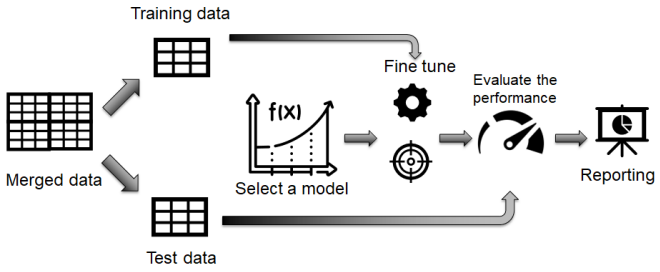

Figure 4. Benchmark and evaluation pipeline

In the context of statistical learning, the data are partitioned into three subsets:

- **Training set** ($\mathcal{D}_{\text{train}}$): It involves 60% of all the participants (4657) and is used to estimate the parameters θ of the predictive model f_θ;
- **Validation set** ($\mathcal{D}_{\text{val}}$): It involves 10% of the participants and is used to estimate the hyperparameters θ_{hyp} of the predictive model f_θ;
- **Test set** ($\mathcal{D}_{\text{test}}$): It involves 30% of all the participants, and it is used to evaluate the performance of the trained model on unseen data and assess the generalization ability of the model.

After selecting the models, their hyperparameters (θ_{hyp}) are

fine-tuned to optimize performance. This crucial step enhances the model's predictive capabilities and is elaborated on in Section III-C. The optimization process may involve techniques such as grid search or gradient-free optimization methods (e.g., Nevergrad), depending on the model's complexity.

Subsequently, each model's performance is evaluated based on a set of criteria measuring accuracy and reliability. The results are presented using both tabular and graphical tools to facilitate comparison and interpretation. These results offer insights into the models' predictive capabilities and help identify the most suitable approach for modeling PCSI.

Lastly, to identify patient profiles and implement preventive measures against Post COVID-19 condition, it is crucial to assess the significance of the explanatory variables used for model training and parameter adjustment. Depending on the model utilized, we employ explanation and interpretation tools to extract meaningful insights. These insights can offer valuable guidance for the medical field.

A. Evaluated Methods

To tackle the regression problem, we evaluated and compared several data-driven models, including Linear Ridge Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Multi-Layer Perceptron (MLP). LR is a linear model enhanced with regularization to address multicollinearity and reduce overfitting. RF is an ensemble technique that builds multiple decision trees and aggregates their predictions for robust regression. GB sequentially combines weak learners, typically decision trees, to minimize errors and improve predictive accuracy. MLP is a feed-forward neural network excelling at modeling non-linear relationships with fully connected layers of neurons and non-linear activation functions.

B. Evaluation criteria

Considering that PCSI is a continuous target variable, we have selected four evaluation criteria to assess the model's performance, which are: *MAPE* (Mean Absolute Percentage Error), *MAE* (Mean Absolute Error), *MSE* (Mean Squared Error) and *Pearson correlation* between predicted and actual values.

C. Experimental setup

We fine-tuned all the presented models to determine the optimal set of hyperparameters. For hyperparameter optimization, we employed the Nevergrad library [14]. The best hyperparameters for MLP were: 3 hidden layers with 126 neurons each, ReLU activation function, Adam optimizer with a learning rate of 9×10^{-4}, and 200 training epochs. For RF, the optimal settings included 500 estimators, a maximum depth of 12, a maximum sample fraction of 0.4, and 25 maximum features. Similar hyperparameters were achieved for GB. Lastly, for LR, the L2 regularization strength multiplier was set to 1.0. To ensure the stability and robustness of the results, we conducted K-fold ($K = 5$ cross-validation and the results are reported using mean and standard deviation across the five folds.

D. Results

This section presents and discusses the results obtained by the methods introduced and summarizes their performance in Table II. Using each method, different combinations of features are compared through the introduced evaluation criteria. The "All" feature combination represents the integration of all characteristics, including static variables, symptoms, and vaccination data. For clarity, the best results for each method are marked in bold, while the best performance for each evaluation criterion is highlighted in green. Additionally, all performance metrics are averaged across $K = 5$-fold cross-validation and results are reported as MEAN $\pm$ STD (refer to Section III-C for details on the experimental setups). Pearson's correlation is reported using the pair (test statistic, p-value).

TABLE II. COMPARISON BETWEEN VARIOUS INTRODUCED MODELS AND FEATURES COMBINATION FOR PREDICTION OF *PCSI*.

		Evaluation criteria			
Methods	Features	MAE	MSE	MAPE	Pearson
LR	All	**.61 ± .01**	**.68 ± .02**	**.29 ± .01**	**(.56, 6e-70)**
	Static	.71 ± .02	.91 ±.05	.35 ± .01	(.28, 2e-16)
	Symptoms	.62 ± .02	.70 ± .04	.30 ± .01	(.57, 2e-69)
	Vaccination	.81 ± .02	.99 ± .05	.41 ± .01	NaN
RF	All	**.60 ± .01**	**.67 ± .02**	**.28 ± .01**	**(.58, 7e-73)**
	Static	.72 ± .02	.93 ± .05	.35 ± .01	(.26, 1e-15)
	Symptoms	**.60 ± .01**	.66 ± .03	**.28 ± .01**	(.57, 5e-72)
	Vaccination	.79 ± .02	.99 ± .06	.39 ± .01	(.04, 1e-1))
GB	All	**.61 ± .01**	**.66 ± .01**	**.28 ± .01**	**(.57, 4e-74)**
	Static	.72 ± .02	.90 ± .05	.35 ± .01	(.29, 7e-17)
	Symptoms	**.61 ± .01**	.68 ± .02	.28 ± .01	(.55, 8e-82)
	Vaccination	.81 ± .02	.99 ± .06	.41 ± .01	(.05, 6e-1)
MLP	All	**.45 ± .05**	**.90 ± .12**	**.19 ± .03**	**(.25, 3e-18)**
	Static	.87 ± .18	1.4 ± .78	.43 ± .07	(.21, 4e-9)
	Symptoms	.76 ± .11	.98 ± .38	.34 ± .05	(.43, 5e-33)
	Vaccination	.80 ± .03	1.03 ± .05	.41 ± .03	(.04, 2e-1)

As shown in Table II, the best performance for each method is achieved when all features are combined. However, with the exception of MLP, the performance remains comparable even when only symptom-based features are used. It is worth noting that neural network-based methods, such as MLP, have the capability for automatic feature extraction, whereas traditional statistical approaches like LR, RF, and GB require a dedicated feature engineering step.

We observe that the performance, in terms of the MAE metric, remains very similar across the four approaches when all features are combined. An MAE value of 0.60 indicates that, on average, the predicted values deviate by 0.60 points from the actual observations. Given that the PCSI ranges from 1 to 5, a deviation of 0.60 in intensity is unlikely to significantly affect the overall conclusions.

Finally, we note that the best result in terms of MAPE is achieved using MLP, with a value of 0.19. This indicates that, on average, the predictions deviate by 19% from the actual intensity values. Interestingly, the highest Pearson correlations between predictions and actual values are obtained with RF and GB, rather than MLP. This discrepancy can be attributed to the differences in how these models capture relationships within the data. RF and GB are ensemble-based methods that excel in capturing complex interactions between features,

which may result in higher linear correlations (as measured by Pearson correlation) between predicted and actual values. On the other hand, MLP, being a neural network, is better suited for non-linear patterns and optimization for specific loss functions, which may explain its superior performance in minimizing relative errors (as captured by MAPE).

E. Interpretation

Using explainability tools, this section allows to better understand the models' decision through some statistics such as estimated feature coefficients and feature importance.

The top 9 most influential features, along with their corresponding Linear Ridge Regression (LR) coefficients, averaged over 5-fold cross-validation are presented in Table III. These coefficients indicate the direction and magnitude of each feature's contribution to the prediction of PCSI. Many common acute symptoms, such as loss of sense of smell, headache, and muscle pain, exhibit strong positive contributions, suggesting they are associated with a higher risk of Post COVID-19 condition. Conversely, certain acute symptoms like fever or pain when breathing show significant negative contributions, indicating that their presence is less likely to increase the risk of Post COVID-19 condition. This distinction highlights the nuanced relationship between acute and long-term COVID symptoms.

TABLE III. ESTIMATED COEFFICIENTS OF LINEAR REGRESSION FOR PREDICTION OF POST COVID-19 CONDITION

Variable	Coef	Variable	Coef
Loss of sense of smell/taste	0.32	Pain when breathing	-0.58
Headache	0.28	Fever (38° or higher)	-0.27
Muscle pain/aches	0.27	Omicron variant	-0.26
Lower back pain	0.23	Heaviness in arms/legs	-0.08
Original variant	0.17	Very good health	-0.07
Feeling warm & cold	0.16	No chronic disease	-0.07
Red, painful eyes	0.16	Age group	-0.06
Sneezing	0.16	Smoker	-0.05
Difficulty breathing	0.14	Male	-0.03

The importance of features obtained by the Random Forest (RF) model is illustrated in Figure 5 using a bar plot. For clarity and brevity, only the top 10 most important features were extracted from the full set. The identified features show some overlap with those presented in Table III, although their relative importance differs. Notably, muscle pain emerges as the most important predictor of PCSI. Additionally, the feature representing the time interval between vaccination and infection (VACCIN_TTI in the bar plot) is highlighted as a significant contributor. This finding supports the hypothesis that vaccination timing influences the risk and severity of Post COVID-19 condition, emphasizing its potential impact on disease outcomes.

Based on the SHAP explanation tool, the most influential features for the MLP model predicting PCSI are identified in Figure 6. Key symptoms such as difficulty breathing, diarrhea, fluctuating body temperature, muscle pain, and sneezing had high positive SHAP values, indicating strong contributions to increased symptom intensity.

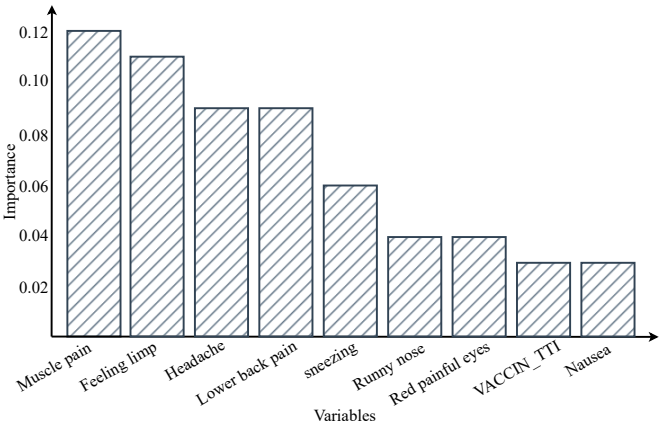

Figure 5. Feature importances resulted using Random Forest model for prediction of PCSI

Smoking was associated with higher PCSI, likely due to its impact on respiratory health. In contrast, the absence of chronic diseases and prior vaccination were linked to reduced intensity, emphasizing the protective role of good baseline health and immunization. Additionally, female sex was associated with higher PCSI, in line with existing research on sex-based vulnerability to post-viral syndromes [15]. These findings highlight the complex interplay of symptoms and individual factors in shaping Post COVID-19 outcomes.

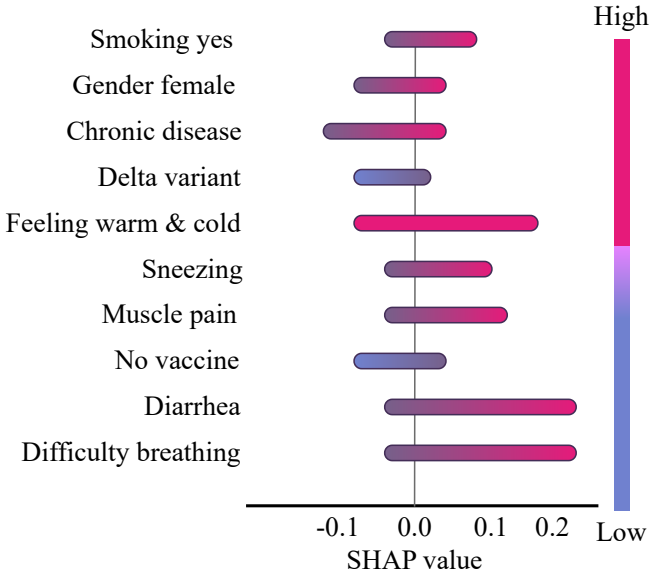

Figure 6. Interpreting MLP influential factors using SHAP

IV. CONCLUSION AND PERSPECTIVES

This study aimed to identify patient profiles at higher risk of developing PCC and predict its intensity using machine learning approaches. We utilized features that were grouped into static, vaccination, and symptom-related variables. Statistical analyses revealed that women and patients with chronic diseases are more susceptible to PCC. Predictive analysis using

four different models demonstrated strong performance across all methods when combining all features, with MLP showing slightly better results in terms of MAPE. The interpretability analyses identified key predictors, including loss of smell, headache, muscle pain, and vaccination timing, as well as protective factors like the absence of chronic diseases. These insights provide valuable information for tailoring interventions and understanding the underlying risk factors of PCC.

Limitations and future works. The steady-state assumption in our analysis limits the ability to capture temporal relationships between symptoms or events. Model performance is also constrained by the quality and completeness of the dataset, highlighting the need for validation on independent datasets to ensure robustness in real-world scenarios. Additionally, while the models offer predictive value, they are intended as tools to complement clinical judgment rather than replace it. These gaps will be addressed in future studies.

Societal Impact. Post COVID-19 condition has profound societal implications, affecting physical and mental health, daily functioning, and productivity [16][17]. It disrupts educational and professional activities, with children and adults experiencing isolation, stress, and cognitive impairments. Predicting PCC symptoms intensity can inform early interventions, alleviate healthcare burdens, and improve patients' quality of life.

ACKNOWLEDGMENT

This work was supported by the ZonMw COVID-19 programme (10430302110002). The Lifelines initiative has been made possible by subsidy from the Dutch Ministry of Health, Welfare and Sport, the Dutch Ministry of Economic Affairs, the University Medical Center Groningen (UMCG), Groningen University and the Provinces in the North of the Netherlands (Drenthe, Friesland, Groningen).

REFERENCES

[1] T. Lancet, *The covid-19 pandemic in 2023: Far from over*, 2023.

[2] A. V. Ballering, S. K. van Zon, T. C. olde Hartman, and J. G. Rosmalen, "Persistence of somatic symptoms after covid-19 in the netherlands: An observational cohort study," *The Lancet*, vol. 400, no. 10350, pp. 452–461, 2022.

[3] F. Callard and E. Perego, "How and why patients made long covid," *Social science & medicine*, vol. 268, p. 113 426, 2021.

[4] https://www.who.int/europe/news-room/fact-sheets/item/post-COVID-19-condition. retrieved: September, 2025.

[5] C. E. Hastie *et al.*, "Natural history of long-covid in a nationwide, population cohort study," *Nature Communications*, vol. 14, no. 1, p. 3504, 2023.

[6] R. Kessler, J. Philipp, J. Wilfer, and K. Kostev, "Predictive attributes for developing long covid—a study using machine learning and real-world data from primary care physicians in germany," *Journal of Clinical Medicine*, 2023.

[7] C. Sudre and al., "Symptom clusters in covid-19: A potential clinical prediction tool from the covid symptom study app," *Science advances*, 2021.

[8] M. Leyli-abadi, *ML4HEALTH Python Package*. [Online]. Available: https://github.com/Mleyliabadi/ML4HEALTH.

[9] J. B. Soriano, S. Murthy, J. C. Marshall, P. Relan, and J. V. Diaz, "A clinical case definition of post-COVID-19 condition by a delphi consensus," *The Lancet. Infectious Diseases*, 2022.

[10] S. Srikanth, J. R. Boulos, T. Dover, L. Boccuto, and D. Dean, "Identification and diagnosis of long covid-19: A scoping review," *Progress in biophysics and molecular biology*, vol. 182, pp. 1–7, 2023.

[11] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, 2013.

[12] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 43.

[13] M. Greenacre and J. Blasius, *Multiple correspondence analysis and related methods*. CRC press, 2006.

[14] J. Rapin and O. Teytaud, *Nevergrad - A gradient-free optimization platform*, https://GitHub.com/FacebookResearch/Nevergrad, 2018.

[15] F. Bai *et al.*, "Female gender is associated with long covid syndrome: A prospective cohort study," *Clinical microbiology and infection*, vol. 28, no. 4, 611–e9, 2022.

[16] M. Mayhew, G. Kerai, and D. Ainslie, "Coronavirus and the social impacts of 'long covid'on people's lives in great britain: 7 april to 13 june 2021," *Newport, Wales: Office for National Statistics*, pp. 1–21, 2021.

[17] A. MacLean *et al.*, "Impact of long covid on the school experiences of children and young people: A qualitative study," *BMJ open*, vol. 13, no. 9, e075756, 2023.