Med Innovations Journal

Received: 17.06.2026
Revised: 05.09.2026
Accepted: 07.09.2026
Published: 08.09.2026

Original Article

Evaluation of principal component–based and synthetic feature representations in machine learning prediction of hospital length of stay in COVID-19 patients using XGBoost algorithm

Jacek Jankowski [1*], Joanna Bargieł-Kostuj [2], Marcin Rojek [1,3], Jakub Kufel [4]

1. Department of Biophysics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, ul. Jordana 19, 40-055 Katowice, Poland
2. Department of Internal Medicine, Diabetology and Nephrology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, ul. Jordana 19, 40-055 Katowice, Poland
3. Department of Histology and Cell Pathology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, ul. Jordana 19, 40-055 Katowice, Poland
4. Department of Radiology and Nuclear Medicine, Faculty of Medical Sciences in Katowice, Medical University of Silesia, 40-752 Katowice, Poland

*Corresponding author: jankowskijacek93@interia.pl

Abstract

Background: The COVID-19 pandemic highlighted limitations of healthcare systems and the need for early identification of patients at risk of prolonged hospitalization. This study aimed to develop a machine learning model for predicting the length of hospital stay in patients with non-severe COVID-19 and to identify key predictive factors using SHAP and PCA analyses.

Material and methods: Anonymized records from 1,027 patients hospitalized due to SARS-CoV-2 infection between November 2020 and May 2021 were retrospectively screened. Five records with corrupted documentation were excluded, leaving 1,022 evaluable patients; after exclusion of severe COVID-19 cases, 786 patients were included in the predictive analysis. A total of 58 numerical predictor variables were used: age at admission and 57 laboratory-derived features comprising the first available, mean, and median values of 19 laboratory parameters. XGBoost regression models with quantile regression were developed using different data representations: RAW variables, PCA_FULL, PCA reduced to 10 components, and synthetic variables. Model performance was assessed using RMSE, MAE, and interval classification agreement.

Results: The RAW model achieved the best predictive performance (RMSE 6.66 days, MAE 4.49 days, agreement 0.276). PCA-based dimensionality reduction substantially reduced prediction accuracy. SHAP analysis showed the largest contributions to model predictions for inflammatory and tissue injury markers, particularly CRP, D-dimer, IL-6, PCT, and AST. PCA revealed biologically interpretable axes associated with inflammatory response, cardiorenal dysfunction, and coagulation abnormalities.

Conclusion: Machine learning models based on raw predictor variables enabled moderately effective modeling of hospitalization length in non-severe COVID-19 patients. Although PCA preserved clinically interpretable biological patterns, it reduced predictive performance. Hospitalization duration appears to be a multifactorial phenomenon requiring broader clinical context in future predictive models. Because the mean and median laboratory values summarized measurements obtained throughout hospitalization, the findings should not be interpreted as admission-time prognostic performance.

Keywords: COVID-19, Principal Component Analysis (PCA), Machine Learning, Length of Stay, Biomarkers

1. Introduction

1.1. COVID-19 and Length of Hospitalization

The COVID-19 pandemic, caused by infection with SARS-CoV-2, represented the greatest global healthcare challenge of the 21st century. The rapid increase in cases led to the overload and failure of healthcare systems in many countries, exceeding their technical, infrastructural, staffing, and organizational capacities. In response to the crisis, temporary healthcare structures were established, while existing medical facilities were transformed into dedicated infectious disease hospitals focused on COVID-19 treatment. At the same time, substantial differences in hospital preparedness were observed. Some institutions were equipped with advanced life-support systems and specialized therapies, whereas others lacked even sufficiently efficient oxygen supply infrastructure. One of the major challenges was the appropriate allocation of patients to healthcare facilities with an adequate level of reference and treatment capability. Cases were reported both of clinical deterioration among patients managed in outpatient settings or inadequately equipped hospitals, and of inefficient utilization of highly specialized resources by patients with mild COVID-19. Therefore, the early identification of patients at risk of severe disease became a matter of critical importance.

1.2. Clinical Implications of Length-of-Stay Prediction Models

In pandemic situations exceeding the capacity of healthcare systems, the development of rapid and threat-specific triage algorithms becomes particularly important. Early assessment of the probable disease course may improve the allocation of medical personnel and equipment by enabling prediction of the likely scope of required therapies. For this purpose, the use of machine learning algorithms appears justified, as they may be capable of extracting laboratory and clinical variables associated with severe COVID-19, prolonged hospitalization, and the need for advanced therapeutic interventions from large volumes of input data. The development of effective predictive modeling strategies may therefore improve healthcare management during potential future pandemic threats.

1.3. Component Analysis and XGBoost as Exploratory Techniques

Principal Component Analysis (PCA) and XGBoost are classical machine learning techniques. PCA is widely used both for exploratory data analysis and data preprocessing in machine learning pipelines [1, 2]. The aim of PCA is dimensionality reduction through transformation of the original correlated variable set into a new set of uncorrelated variables known as principal components. This transformation is particularly useful for reducing the curse of dimensionality, a phenomenon in which the amount of training data becomes insufficient for reliable model training without the risk of overfitting due to excessive multidimensionality. Additionally, PCA enables empirical analysis of individual principal components to determine the clinical nature of the biological phenomena they represent [3]. XGBoost is an advanced implementation of the gradient boosting algorithm based on decision trees, primarily used for predictive tasks involving tabular data. Unlike single decision trees, the model sequentially constructs multiple weak learners, with each subsequent model minimizing the errors of the previous ones by fitting their residuals. This approach enables high predictive accuracy while maintaining control over overfitting through regularization mechanisms [4]. An additional advantage of using classical machine learning techniques, particularly compared with neural networks, is the relatively straightforward application of SHAP (SHapley Additive exPlanations) analysis, which allows evaluation of how individual input features contribute to model predictions [5].

1.4. Aim of the Study

The primary aim of this study was to develop a predictive model of hospitalization length in patients with non-severe SARS-CoV-2 infection. Additionally, the data structure was analyzed to identify the biological variables and components most strongly associated with model predictions using SHAP analysis. As part of the secondary objectives, Principal Component Analysis (PCA) was used to identify the main sources of variability within the dataset and to attempt their clinical interpretation. Furthermore, the effectiveness of different data representation strategies in the context of machine learning was compared, including the full variable set, reduced representations, and synthetic variables.

1.5. Materials and Methods

The dataset used in this study initially consisted of anonymized medical records from 1,027 patients hospitalized due to SARS-CoV-2 infection between November 2020 and May 31, 2021. Five records were

excluded because the corresponding medical documentation was corrupted and could not be reliably evaluated, leaving 1,022 patients in the evaluable cohort. The medical documentation was used to create a database containing information on diagnoses (according to ICD-10 classification), intubation status, mortality, intensive care unit (ICU) hospitalization, patient age, and the complete profile of ordered biochemical laboratory tests, including measurement results and examination dates. Among the numerous available biochemical parameters, 19 parameters with the highest measurement availability were selected for analysis. These included alkaline phosphatase (ALP), alanine aminotransferase (ALT), activated partial thromboplastin time (APTT), aspartate aminotransferase (AST), creatine kinase myocardial band (CK-MB), C-reactive protein (CRP), D-dimers, fibrinogen, gamma-glutamyl transferase (GGTP), hemoglobin concentration (HGB), interleukin-6 (IL-6), international normalized ratio (INR), creatinine concentration, lactate dehydrogenase (LDH), N-terminal pro-B-type natriuretic peptide (NT-proBNP), procalcitonin (PCT), platelet count (PLT), troponin T (TnT), and estimated glomerular filtration rate (eGFR). Laboratory tests were not performed at predefined timepoints but were ordered according to clinical indications and routine hospital practice; consequently, their number and timing varied between patients. For each parameter and patient, the first available value recorded during hospitalization and the median and mean values across the entire hospitalization period were calculated. The first-available values are identified as day 1 values in the figures and internal variable labels. Together with age at admission, the 57 laboratory-derived features resulted in a total of 58 numerical predictor variables. Distributions of selected patient characteristics and laboratory parameters were visualized according to membership in the population with a particularly severe disease course (intubated patients, deceased patients, and patients transferred to the ICU) and the remaining study population (Figure 1).

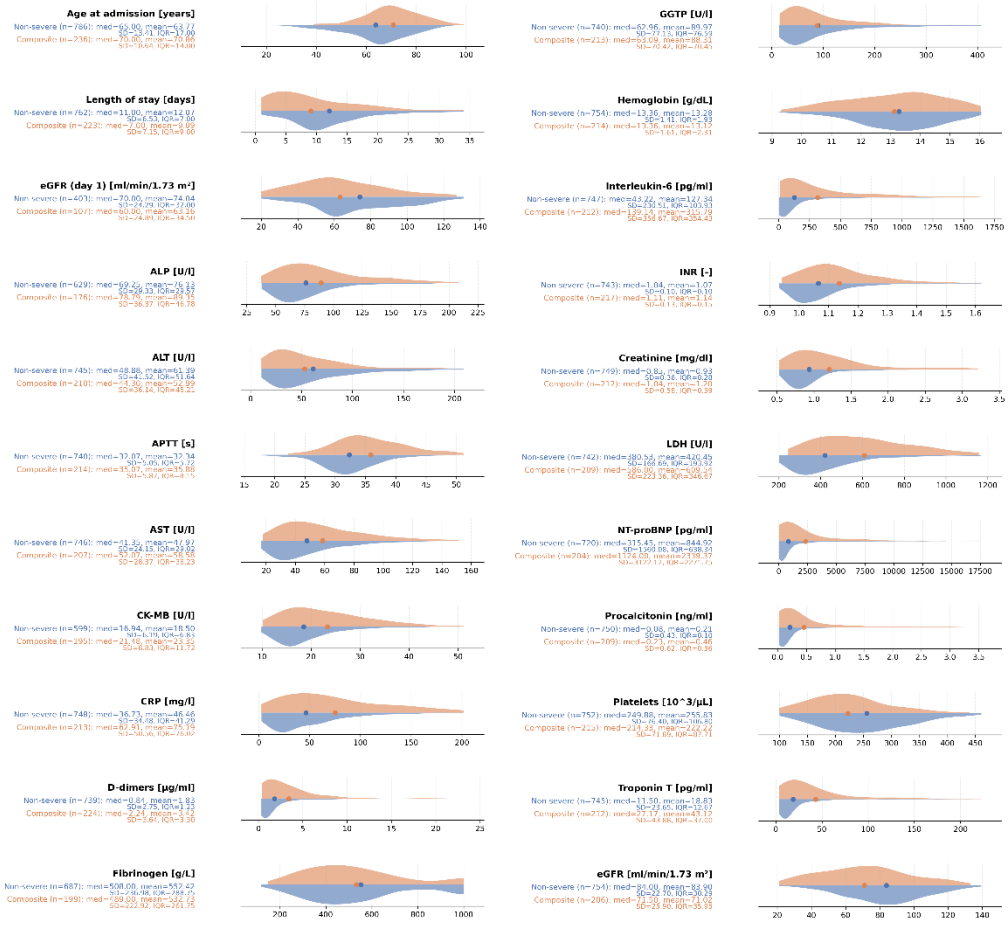

Figure 1. Distribution of age, hospitalization length, and laboratory parameters according to disease severity.

The study population was subsequently restricted to patients whose hospitalization outcome excluded particularly severe disease progression. This decision was justified by the limited availability of data from ICU-transferred patients and by the need to avoid situations in which rapid death resulting from severe disease

would be incorrectly interpreted as rapid clinical recovery associated with short hospitalization duration. After applying these exclusion criteria, the final study population consisted of 786 patients.

2. Methodology

2.1. Statistical Analysis Environment

All statistical analyses were performed using Python 3.12.8 with the use of the NumPy, SciPy, Pandas, and Plotly libraries.

2.2. Principal Component Analysis

A total of 58 numerical variables described above were included in the PCA analysis. The purpose of PCA application was to develop simplified synthetic variables that could subsequently be used for model training. The use of PCA aimed to provide a compromise between dimensionality reduction—particularly in the context of potentially redundant measurements derived from the same laboratory parameter—and preservation of the predictive information contained within the originally multidimensional dataset. Dimensionality reduction was considered particularly important due to the occurrence of the so-called curse of dimensionality.

2.3. Construction of Synthetic Variables

The PCA results were subjected to empirical interpretation in order to identify the biological relationships represented by the dominant variables contributing to subsequent principal components. Ten principal components accounting for the highest variance (cumulatively 60.49%) were selected for interpretation. For each component, the ten predictor variables with the highest contribution were identified. Subsequently, overlapping variables between components were removed based on the highest available loading of a given variable. The final result was a set of 10 synthetic variables derived from the original predictor set without duplication between individual synthetic features. The final stage involved human interpretation of the principal biological axes represented by these synthetic variables based on the dominant contributing laboratory parameters.

2.4. Training Data Preparation and XGBoost Model Development

To prepare the training and testing datasets, a database containing predictor variables and hospitalization length was created. The evaluable cohort (n = 1,022) was divided into patients with a severe disease course (n = 236; including death, intubation, or ICU hospitalization) and non-severe patients (n = 786). Predictive analyses of hospitalization length were subsequently performed exclusively in the non-severe subgroup.

The dataset was divided into a training set (80%, n = 628) and a testing set (20%, n = 158). Missing values in continuous variables were imputed using mean values calculated exclusively on the training set and subsequently applied to the testing set, reducing the risk of data leakage.

Three parallel feature representations were prepared:

1. A raw feature set consisting of 58 numerical predictor variables,
2. A set of 10 synthetic variables created through aggregation of biologically related parameters and redundancy reduction,
3. PCA-based representations including both the full 58-component rotation (PCA_FULL) and a reduced 10-component representation.

PCA transformations and synthetic feature construction were performed on the training set and subsequently mapped onto the testing set using identical transformation parameters. Each representation was independently used for model training and validation (Figure 2).

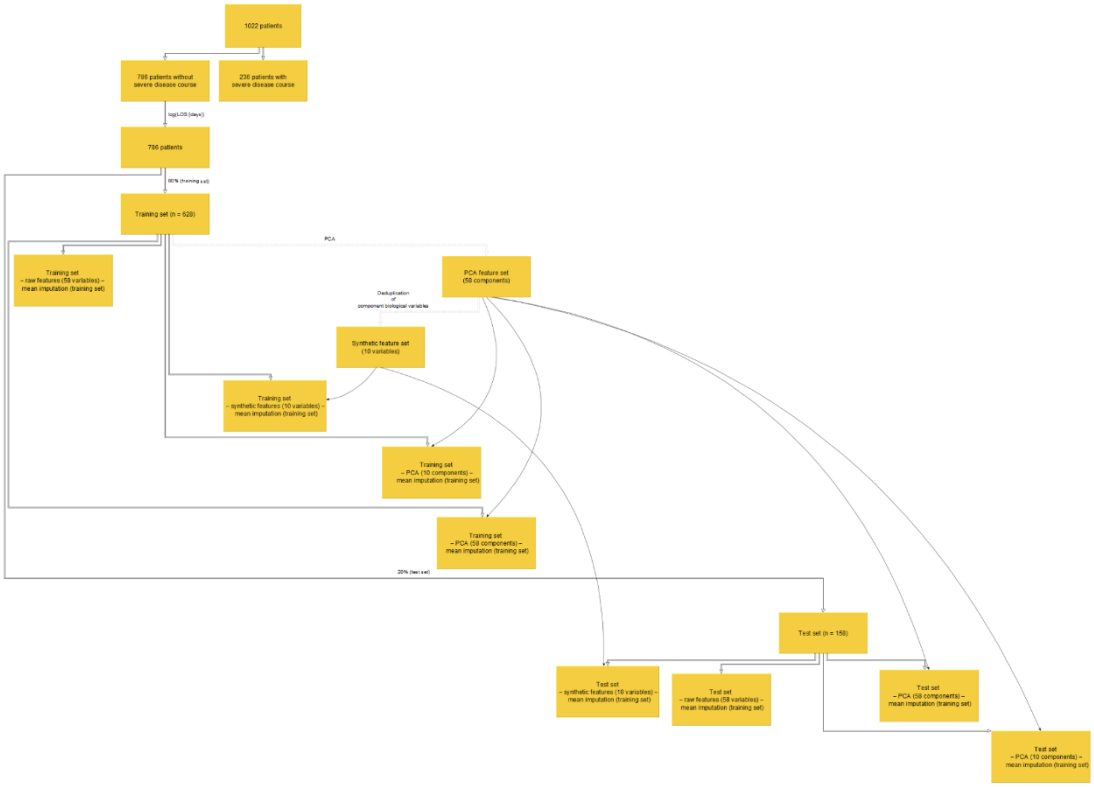

Figure 2. Data preprocessing and feature engineering workflow.

Due to the strongly right-skewed distribution of the target variable, namely hospitalization length, and the presence of outliers, logarithmic transformation using the log1p function was applied. This approach enabled variance stabilization and reduction of the influence of extreme observations while preserving cases with zero-day hospitalization length (Figure 3).

To evaluate the impact of feature representation on predictive performance, four input variants were compared: the full set of 58 numerical predictor variables, a PCA-reduced representation limited to 10 components, PCA_FULL representing full feature-space rotation without dimensionality reduction, and a set of synthetic variables constructed as averages of clinically related parameter groups. Models were trained using the XGBoost gradient boosting decision tree algorithm. Hyperparameter optimization was performed using randomized search. For each data representation, 100 independent hyperparameter configurations were generated. The number of estimators ranged from 300 to 1,200, maximum tree depth from 3 to 8, and learning rate was sampled logarithmically between 10^{-3} and 10^{-1}. Subsample and colsample_bytree parameters were sampled uniformly between 0.6 and 1.0 to introduce stochasticity and reduce overfitting. The min_child_weight parameter ranged from 1 to 8. L1 regularization (reg_alpha) was sampled logarithmically between 10^{-3} and 10^{1}, whereas L2 regularization (reg_lambda) was sampled uniformly between 1 and 5. Hyperparameter exploration was conducted independently for each data representation and model training was performed in parallel across multiple CPU cores.

To reduce the impact of imbalanced hospitalization length distribution, observation weighting was applied during training. Weights were defined as the inverse frequency of rounded hospitalization length values within the training dataset, thereby increasing the relative importance of rare observations and limiting the tendency of the model to predict dominant values from the distribution.

Additionally, a quantile regression approach was implemented by training three independent models for each hyperparameter configuration. These models corresponded to the lower quartile (q = 0.25), median (q = 0.5), and upper quartile (q = 0.75) of the conditional target distribution. The implementation utilized the "quantile loss" function available in XGBoost (objective = "reg:quantileerror") with an appropriately defined quantile_alpha parameter. Median predictions were treated as the primary output, while the difference between upper and lower quartile predictions served as an estimate of prediction uncertainty represented by the interquartile range. Quantile loss additionally improved robustness to outliers and enabled asymmetric penalization of overestimation and underestimation.

Model performance was evaluated on the independent testing dataset. Predictions obtained on the logarithmic scale were transformed back to the original scale using the inverse log1p transformation. Performance metrics included root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), with a small stabilizing constant introduced in the denominator for MAPE calculations to avoid division by zero.

Additionally, prediction quality was evaluated using a discrete classification-based approach. Hospitalization lengths were grouped into three-day intervals starting from zero, and interval agreement between predicted and actual values was calculated as the proportion of observations assigned to the same interval. This metric provided a clinically interpretable assessment of predictive quality.

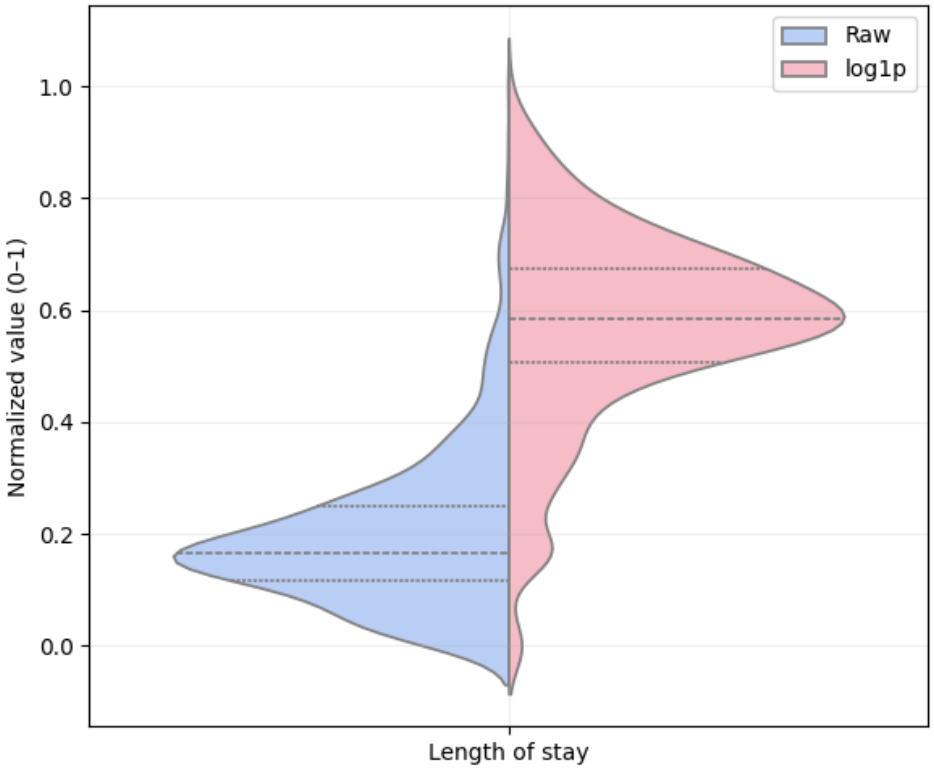

Figure 3. Length-of-stay distribution (raw vs. log1p, normalized separately).

2.5. SHAP Analysis of Trained Models

To interpret the influence of individual variables on model predictions, SHAP (SHapley Additive exPlanations) analysis was performed. SHAP values were calculated using the TreeExplainer algorithm for trained XGBoost median models (q50) on the testing dataset. For each clinical case, SHAP values quantified both the direction and magnitude of influence of individual variables on predicted hospitalization length. Global feature importance was assessed using the mean absolute SHAP value, while the distribution of feature effects was visualized using beeswarm and bar plots. In PCA-based representations, interpretation referred to principal components as linear combinations of original variables, whereas models based on raw and synthetic variables enabled direct clinical interpretation.

3. Results

3.1. Interpretation of Synthetic Variables

Table 1 presents the synthetic variables derived from PCA together with their empirically assigned clinical interpretations and the contributing biological variables with corresponding weights (Table 1).

PC	Synthetic Variable	Biological Variables (loading)
PC1	Hepatic enzyme axis	GGTP (median) (0.287), GGTP (mean) (0.282), ALT (mean) (0.281), ALT (median) (0.274), AST (mean) (0.255), GGTP (day 1) (0.246), ALT (day 1) (0.234), AST (median) (0.234), AST (day 1) (0.220)
PC2	Cardio-metabolic injury axis	LDH (median) (0.222), LDH (mean) (0.220), NT-proBNP (day 1) (0.199), Creatinine (day 1) (0.196)
PC3	Myocardial injury and inflammation axis	CK-MB (mean) (0.286), CK-MB (median) (0.285), IL-6 (median) (0.241), Troponin T (mean) (0.236), Troponin T (median) (0.233), Hemoglobin (mean) (0.205), Hemoglobin (median) (0.203), IL-6 (mean) (0.190)
PC4	Acute phase response axis	Fibrinogen (mean) (0.356), Fibrinogen (median) (0.352), Fibrinogen (day 1) (0.306), CRP (mean) (0.234), CRP (day 1) (0.209)
PC5	Coagulation–liver interaction axis	APTT (median) (0.324), APTT (mean) (0.320), Procalcitonin (day 1) (0.283), APTT (day 1) (0.251), ALP (mean) (0.229), ALP (median) (0.229), INR (median) (0.228), INR (mean) (0.211), ALP (day 1) (0.197)
PC6	Procalcitonin axis	Procalcitonin (median) (0.299), Procalcitonin (mean) (0.286)
PC7	Acute hematologic axis	Platelets (day 1) (0.236), Hemoglobin (day 1) (0.200)
PC8	Thrombo-renal axis	D-dimers (median) (0.372), D-dimers (mean) (0.325), Creatinine (mean) (0.289), Creatinine (median) (0.278), D-dimers (day 1) (0.269)
PC9	Platelet–inflammation axis	Platelets (mean) (0.273), Platelets (median) (0.267), CK-MB (day 1) (0.229), CRP (median) (0.210)
PC10	Cardiorenal perfusion axis	eGFR (mean) (0.335), eGFR (median) (0.335), NT-proBNP (mean) (0.236), NT-proBNP (median) (0.235), LDH (day 1) (0.199), eGFR (day 1) (0.183)

Table 1. Synthetic variables derived from principal component analysis and their biological composition.

Principal Component Analysis (PCA) enabled the identification of dominant biological axes describing variability within the studied population of patients hospitalized due to COVID-19. The first principal component (PC1) was dominated by hepatic enzymes (ALT, AST, GGTP) across all analyzed temporal aggregations, indicating a distinct hepatocellular injury axis. The consistency and high magnitude of the loadings suggest that PC1 reflects the global level of enzymatic activity associated with cellular injury, with relatively limited contribution from other organ systems.

The second and third principal components (PC2 and PC3) represented more complex and clinically relevant patterns. PC2 combined markers of tissue injury (LDH), cardiovascular overload (NT-proBNP), and renal function (creatinine), suggesting a multidimensional axis of multiorgan dysfunction. In contrast, PC3 was primarily associated with markers of myocardial injury (CK-MB, troponin T) and inflammation (IL-6), indicating an axis involving the coexistence of cardiac injury and inflammatory response.

The fourth principal component (PC4) was clearly associated with the acute phase response represented by fibrinogen and CRP. The coherence of this axis highlights its significance as an indicator of active inflammatory processes. In contrast, PC5 reflected more complex interactions between the coagulation system (APTT, INR), hepatic parameters (ALP), and the bacterial infection marker procalcitonin, potentially indicating relationships between hemostatic disturbances and infection severity.

Subsequent components demonstrated more specialized biological profiles. PC6 was dominated exclusively by procalcitonin, suggesting that it reflects a narrow aspect of inflammatory response potentially

related to bacterial superinfection. PC7 included only hematologic admission parameters (platelets and hemoglobin), indicating limited biological complexity and potentially lower clinical relevance.

An important component was PC8, which combined thrombotic activity markers (D-dimers) with renal function parameters (creatinine), forming a distinct thrombo-renal axis. PC9 represented interactions between platelet-related parameters, inflammatory markers (CRP), and myocardial injury markers (CK-MB), potentially reflecting interconnected inflammatory and hemostatic processes. The final analyzed component (PC10) included renal function parameters (eGFR) and markers of cardiac overload (NT-proBNP), suggesting an axis related to cardiorenal perfusion and functional status.

Overall, the obtained results indicate that PCA enables dimensionality reduction while preserving clinically interpretable biological patterns. Components reflecting multiorgan dysfunction (PC2), inflammatory response (PC4), and thrombo-renal disturbances (PC8) appear particularly relevant and may possess predictive significance in subsequent machine learning analyses.

3.2. Results of XGBoost Training

The training results are presented in Table 2. RMSE and MAE reflect absolute prediction accuracy expressed in days, whereas MAPE represents relative prediction error. Accuracy in three-day bins reflects interval-based agreement with potential clinical relevance, while IQR represents the dispersion and uncertainty of model predictions. The interquartile range was calculated as the difference between predictions generated by the internal upper- and lower-quantile models for each observation and subsequently averaged across the entire testing dataset.

Dataset	RMSE (days)	MAE (days)	MAPE	Accuracy (3-day bins)	IQR (days)	Number of Features
PCA	9.00	6.52	0.810	0.179	11.34	10
PCA_FULL	6.92	4.69	0.517	0.256	8.78	58
RAW	6.66	4.49	0.489	0.276	6.56	58
SYNTH	8.94	6.23	0.746	0.244	8.27	10

Table 2. Comparison of XGBoost predictive performance depending on input data representation strategy.

The best results were obtained for the model trained on the full set of raw variables (RAW), which achieved the lowest absolute prediction errors (MAE 4.49 days, RMSE 6.66 days) as well as the lowest relative error (MAPE 0.489). This model also demonstrated the highest interval classification agreement (BIN accuracy 0.276), indicating the greatest potential usefulness for practical stratification of hospitalization duration. Importantly, the RAW model also achieved the lowest IQR value (6.56 days), suggesting relatively high prediction stability and lower uncertainty compared with the remaining approaches. Detailed distributions of prediction errors and agreement matrices are presented in Figure 4 for the RAW model (Figure 4).

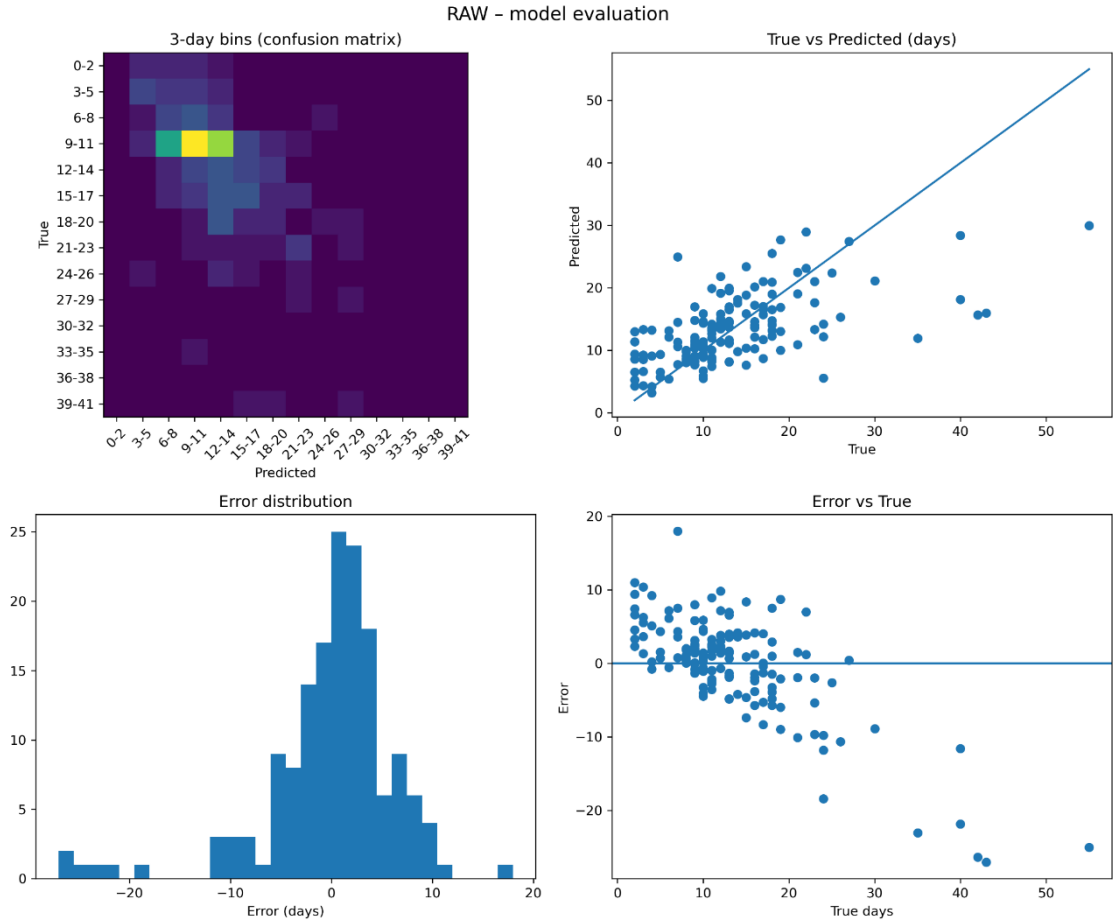

Figure 4. Distribution of prediction errors and agreement matrices for the model trained on the full raw feature set (RAW). The figure presents comparisons between actual and predicted values in logarithmic and real-day scales, together with classification matrices for three-day intervals.

The PCA_FULL model, which utilized full feature-space rotation without dimensionality reduction, achieved performance relatively similar to the RAW model, although consistently inferior across all evaluated metrics. In particular, higher prediction errors were observed (MAE 4.69 days, RMSE 6.92 days, MAPE 0.517), accompanied by reduced interval classification agreement (0.256). Additionally, the IQR value was higher (8.78 days), indicating greater prediction dispersion. These findings suggest that orthogonalization of the feature space does not provide substantial benefit for the applied model architecture, which inherently handles feature correlations effectively. Corresponding prediction distributions and error matrices are presented in Figure 5 for the PCA_FULL model (Figure 5).

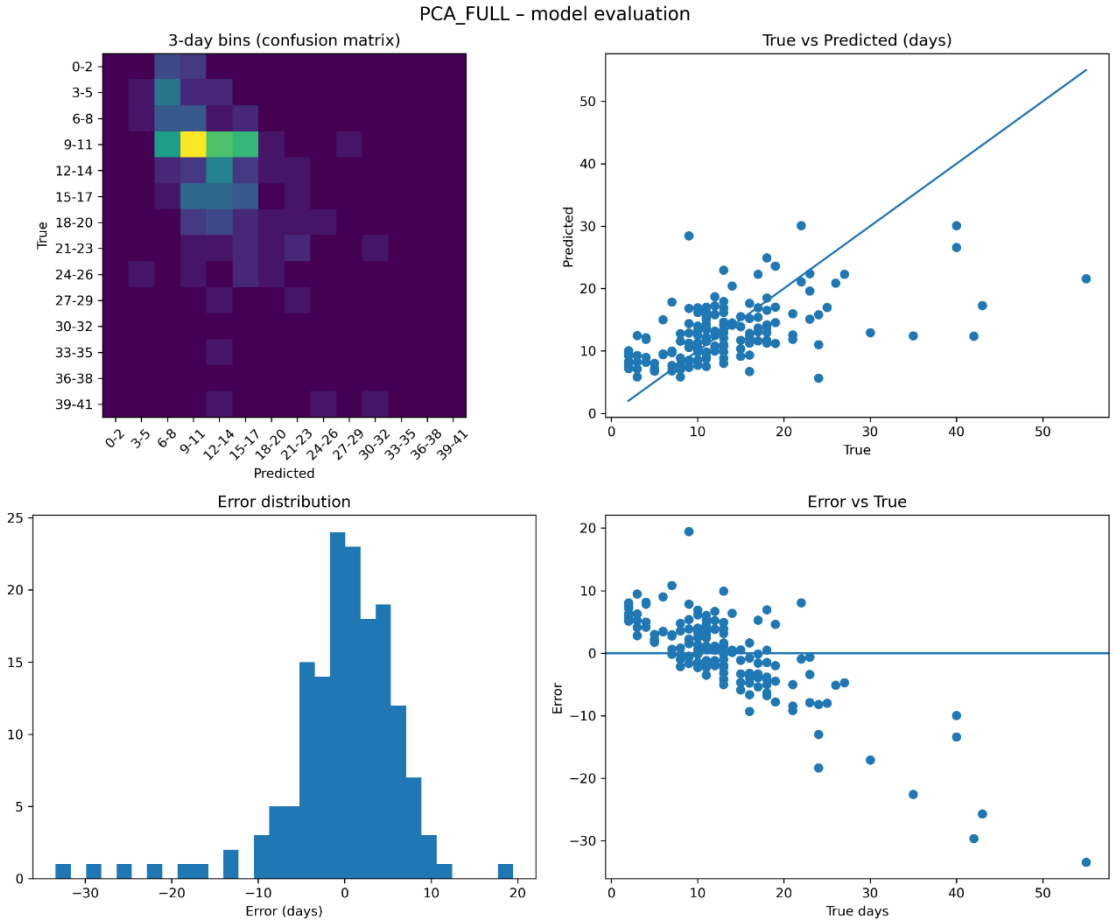

Figure 5. Distribution of prediction errors and agreement matrices for the PCA_FULL model (full feature-space rotation without dimensionality reduction). The visualization includes comparisons of actual and predicted values in logarithmic and real-value scales as well as interval classification agreement analysis.

A substantial deterioration in predictive performance was observed for the PCA model limited to 10 components. Dimensionality reduction resulted in a marked increase in prediction errors (MAE 6.52 days, RMSE 9.00 days, MAPE 0.810) and the lowest interval classification agreement (0.179). This model also demonstrated the highest IQR value (11.34 days), indicating considerable prediction uncertainty. These results suggest a significant loss of predictive information resulting from dimensionality reduction despite preservation of the principal variance directions within the dataset. Visualization of prediction error distributions and agreement matrices for this approach is presented in Figure 6 for the PCA model (Figure 6).

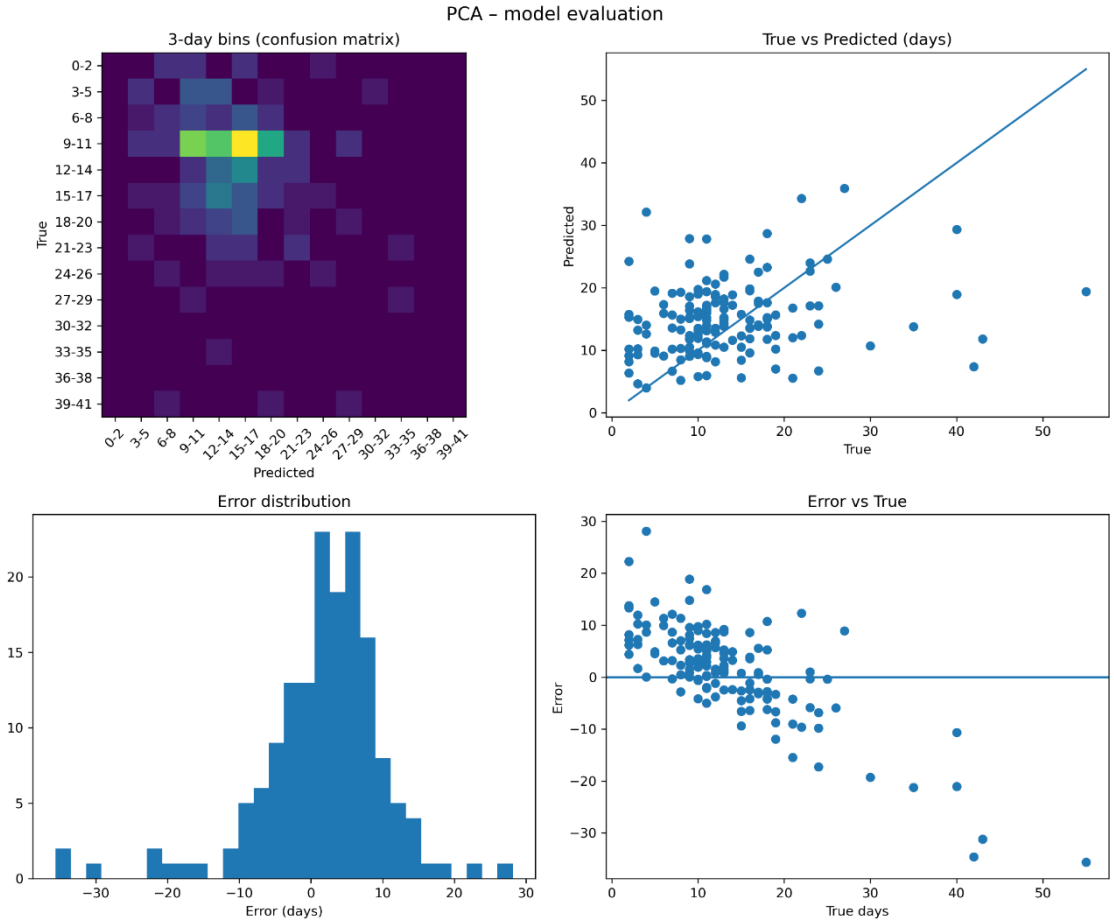

Figure 6. Distribution of prediction errors and agreement matrices for the PCA model limited to 10 principal components. Increased prediction variance and deterioration of interval classification agreement are visible compared with models utilizing the full feature space.

The model based on synthetic variables achieved intermediate performance between PCA and approaches utilizing the full feature space. Prediction errors remained relatively high (MAE 6.23 days, RMSE 8.94 days, MAPE 0.746), although interval classification agreement (0.244) was higher than that observed for PCA. The IQR value (8.27 days) indicated a moderate level of prediction uncertainty. These findings suggest that aggregation of variables based on clinical interpretation partially preserves useful information, although at the cost of losing detailed relationships present within the complete dataset.

3.3. Model Interpretation Using SHAP

Due to the large volume of results, only SHAP analyses for the two best-performing models based on RAW and PCA_FULL representations are presented in this study. The first figure presents the 15 input features with the largest contributions to the output of the RAW model (Figure 7).

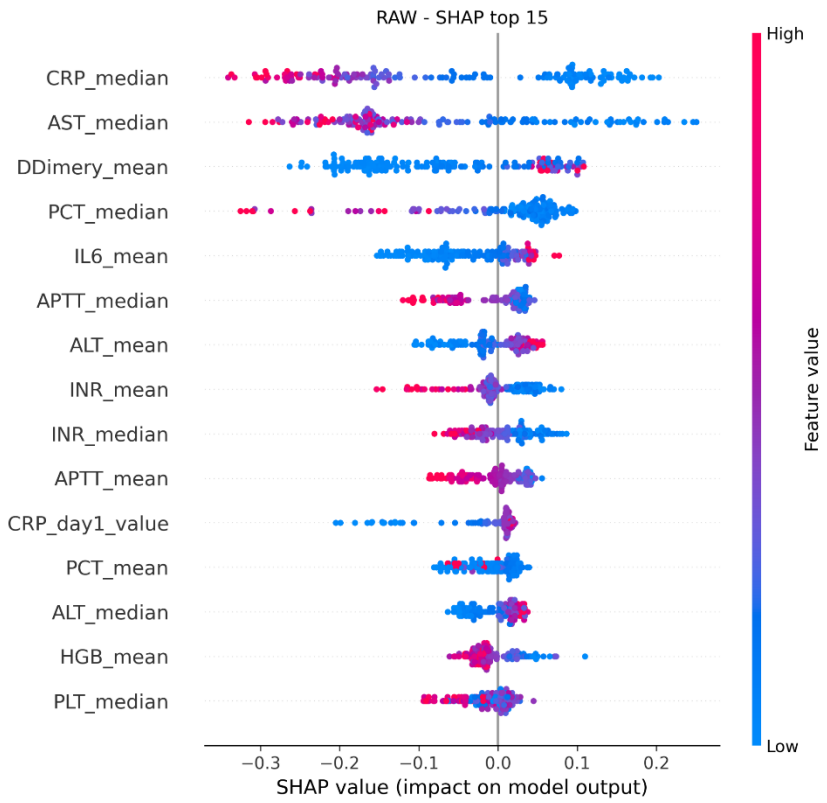

Figure 7. SHAP summary plot for the RAW feature model (top 15 variables).

The largest contributions to hospitalization length predictions in the RAW model were observed for inflammatory and tissue injury markers, particularly CRP, AST, D-dimers, PCT, and IL-6. A clear and consistent direction of association was observed for most of these variables: higher values (red points) shifted predictions toward longer hospitalization duration (positive SHAP values), whereas lower values (blue points) were associated with shorter predicted hospital stay. This pattern was particularly pronounced for D-dimers, IL-6, and PCT, which is consistent with their established role as markers of severe inflammatory response and clinical complications.

A second group consisted of coagulation and organ function parameters (INR, APTT, ALT, HGB, and PLT), whose contributions were more dispersed and directionally less consistent. For these variables, SHAP values remained closer to zero, suggesting lower predictive importance or stronger dependence on the context of other variables. The presence of both aggregated and first-available measurements (e.g., CRP_day1_value) indicates that the model used information from both the beginning and the full course of hospitalization, although the aggregated measures (mean and median values) made the largest contributions.

The following figure presents the 15 principal component signals with the largest contributions to the PCA_FULL model output (Figure 8).

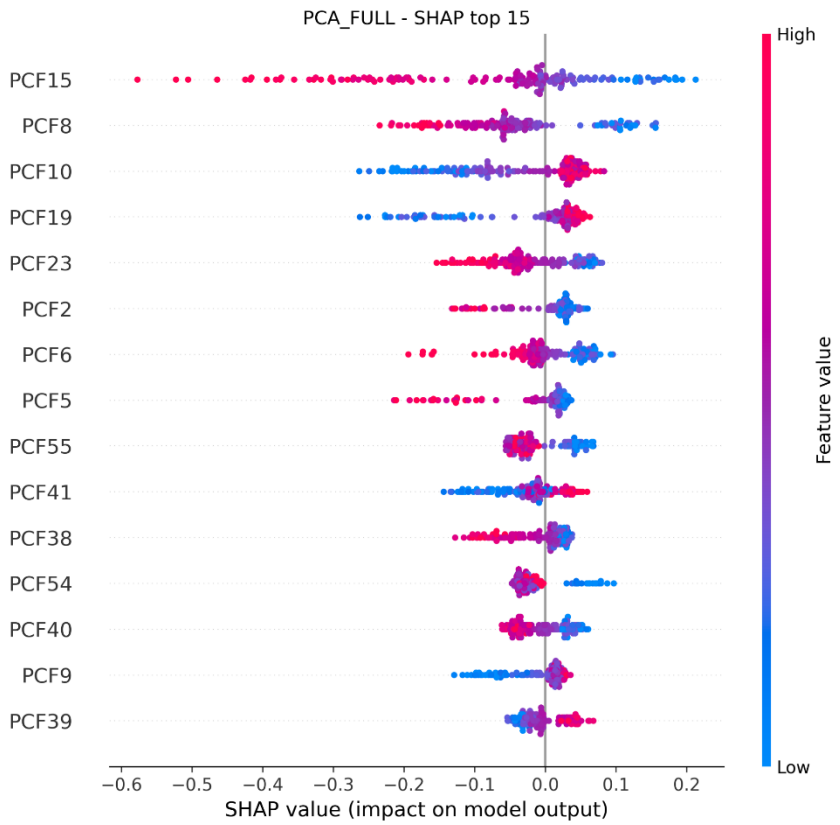

Figure 8. SHAP summary plot for the PCA_FULL model (top 15 principal components).

The largest contributions to predictions in the PCA_FULL model were observed for components PCF15, PCF8, and PCF10. PCF15 demonstrated the widest SHAP value range, indicating that it was the dominant component within the model. High component values (red points) shifted predictions toward shorter hospitalization duration (negative SHAP values), whereas low values (blue points) increased predicted length of stay. This SHAP pattern indicates a strong monotonic association within the model. PCF8 demonstrated a similar, although somewhat weaker, pattern. Higher values of this component were associated with reduced predicted hospitalization duration, whereas lower values were associated with prolonged stay. Its SHAP distribution was more concentrated than that of PCF15, indicating less variability in its contribution despite its continued importance to model performance.

PCF10 displayed an opposite profile: higher values (red points) shifted predictions toward longer hospitalization duration (positive SHAP values), whereas lower values reduced the predicted duration. This suggests that PCF10 represents a clinically distinct pathophysiological dimension compared with the two previous components.

Factor loading analysis of the three most important components (PC8, PC10, and PC15) indicated that each represented a distinct and clinically interpretable dimension of patient condition. PC8 was dominated by strong positive loadings for coagulation and inflammatory markers, particularly D-dimers (first available, mean, and median values), procalcitonin (PCT), and creatinine. Simultaneously, negative loadings were observed for NT-proBNP and hepatic parameters (ALT), suggesting that this component may reflect a thrombo-inflammatory profile associated with renal dysfunction and relatively lower cardiac involvement. In the context of SHAP analysis, the pattern of this component indicates that coagulation abnormalities and elevated inflammatory markers contributed substantially to model estimates associated with longer hospitalization.

PC10 represented a distinct biological profile characterized by strong positive loadings for renal and cardiac function parameters (eGFR, NT-proBNP, creatinine), tissue injury enzymes (LDH), and selected coagulation markers (APTT). Simultaneously, negative loadings were observed for hepatic enzymes (ALT, AST) and INR. This configuration suggests that PC10 reflects a complex systemic condition involving cardiorenal dysfunction and tissue injury, potentially corresponding to more severe multiorgan disease. Consistent with SHAP analysis, this component demonstrated an opposite effect direction compared with PC8, further emphasizing its distinct pathophysiological character.

In contrast, PC15 was strongly associated with inflammatory and coagulation-related markers, with dominant positive loadings for CRP and PLT, accompanied by strong negative loadings for fibrinogen, IL-6, and INR. This

pattern may reflect a more complex and nonlinear representation of the inflammatory–coagulation axis, potentially differentiating phases of inflammatory response or distinct immunological phenotypes among patients. High values of this component may therefore represent a specific inflammatory profile associated with predicted hospitalization duration.

4. Discussion

4.1. Predictive Performance of Machine Learning Models in Relation to Data Representation

The most unexpected finding emerging from the conducted analyses was the limited usefulness of the applied dimensionality reduction techniques in the context of predictive performance. Among all evaluated approaches, the model trained on raw data (MAE = 4.49), consisting of 58 numerical predictor variables including age and the mean, median, and first-available laboratory measurements, achieved the best results. Alternative models trained on the full set of 58 principal components (MAE = 4.69), synthetic variables (MAE = 6.23), and the reduced set of 10 principal components explaining the largest proportion of variance (MAE = 6.52) demonstrated progressively worse performance, approaching the practical limit of clinical usefulness.

These findings suggest both a substantially greater contribution of individual biological signals interacting in complex nonlinear relationships that cannot be effectively extracted by PCA and a fundamental limitation of the proposed dimensionality reduction approach for hospitalization length prediction. The obtained results also suggest that important correlates of hospitalization duration were hidden or unmeasured in the available dataset. This observation may provide valuable information for the future design of analogous predictive models.

4.2. Explainability of Machine Learning Models Using SHAP Analysis

One of the most valuable aspects of the presented analyses, particularly in the context of the moderate predictive performance of the models themselves, was the SHAP analysis of the model trained on raw biological variables. For completeness and transparency, SHAP analyses were presented for all models; however, the most clinically interpretable findings were obtained from the RAW model.

Within this retrospective modeling framework, mean and median laboratory values made larger contributions to hospitalization length estimates than first-available measurements. This pattern indicates that aggregated abnormalities recorded during the hospital course contained information associated with hospitalization duration, but it should not be interpreted as evidence of prospective predictive value at admission. The variables with the largest contributions to model predictions included inflammatory and tissue injury markers such as CRP, AST, D-dimers, PCT, and IL-6. These observations remain consistent with previous reports emphasizing the role of cytokine storm and systemic inflammatory response in COVID-19 severity [6, 7].

4.3. Clinical Interpretability of Principal Component–Derived Phenotypes

Although PCA did not fulfill its intended role as a preprocessing strategy for model training, it provided valuable insight into the dominant biological axes present within the dataset. Once again, mean and median measurements emerged as the most important contributors to principal components. Empirical interpretation of PCA-derived biological axes demonstrated the highest explanatory contribution for the Hepatic Enzyme Axis, Cardio-metabolic Injury Axis, and Myocardial Injury & Inflammation Axis.

This finding is particularly interesting in the context of the poor predictive performance of PCA-based models and the SHAP results obtained for the best-performing RAW model. Together, these observations suggest that while PCA captures clinically meaningful biological trends, the compression of multidimensional laboratory information into orthogonal components may simultaneously eliminate subtle nonlinear interactions crucial for predictive performance.

4.4. Comparison with Previous Studies

To compare the developed model with previously published approaches, the main characteristics of studies investigating the prediction of length of stay in patients with COVID-19 were summarized. The comparison includes study population size, predicted clinical outcomes, and the types of input data used for model development (Table 3).

Study	Population	Outcome	Features
Ebinger et al., 2021 [8]	966 COVID-19 patients	Prolonged LOS	Laboratory, demographic and clinical variables
Mahboub et al., 2021 [9]	2017 patients	LOS, mortality risk	Laboratory, demographic and clinical variables
Olivato et al., 2022 [10]	>6000 patients	Simplified LOS	Laboratory and demographic variables, hospital ward, vaccination status
Amiri et al., 2023 [11]	1291 patients	LOS, mortality risk	Demographic, clinical and comorbidity-related variables
Orooji et al., 2022 [12]	1225 COVID-19 patients	LOS	Demographic, clinical, comorbidity-related and laboratory variables

Table 3. Comparison of study populations, predicted outcomes, and input feature categories in machine learning models developed for length of stay prediction in patients with COVID-19.

The methodological aspects of the included studies were subsequently analyzed, including the machine learning algorithms employed, model explainability techniques, dimensionality reduction methods, and predictive performance metrics. In addition, the most important predictors of length of stay identified by the authors of each study were summarized and compared (Table 4).

Study	ML method	Explainability	Dimensionality reduction	Performance	Key findings
Ebinger et al., 2021 [8]	ENET Blender (days 1–2), AVG Blender (day 3)	Feature importance	No	AUC = 0.819	Age and admission reason were the strongest predictors of prolonged LOS
Mahboub et al., 2021 [9]	Decision Tree	No	No	Accuracy = 96%	Age, BMI, urea level and ventilation status were associated with LOS and mortality
Olivato et al., 2022 [10]	RF, ET, XGBoost, LightGBM, FFNN	SHAP	No	ROC-AUC	D-dimer, age, NLR and fibrinogen were the most important predictors
Amiri et al., 2023 [11]	MLP, ElasticNet, SVR, Lasso, Ridge	LIME	No	MAE = 0.434, RMSE = 0.624, MSE = 0.389	Hyperlipidemia and major comorbidities were key predictors of LOS and mortality

Orooji et al., 2022 [12]	ANN	Feature selection	Feature selection	RMSE = 1.236213	Age, laboratory markers and respiratory disease severity were the strongest predictors of LOS

Table 4. Comparison of machine learning methods, model interpretability approaches, predictive performance, and key predictors identified in studies on length of stay prediction in patients with COVID-19.

To place the obtained results in the context of existing research, previously published machine learning models developed for the prediction of hospital length of stay in patients with COVID-19 were reviewed (Tables 3 and 4). Despite substantial differences in study populations, predicted outcomes, and input variables, most models were based on combinations of laboratory, demographic, and clinical data. In contrast, the present study focused primarily on laboratory parameters, with age at admission as the only demographic predictor, allowing a targeted evaluation of biochemical feature representations.

The reviewed studies consistently highlighted age, inflammatory markers, coagulation abnormalities, and indicators of organ dysfunction as important predictors of hospitalization length. Similar patterns were observed in the present study, where SHAP analysis identified CRP, D-dimers, IL-6, procalcitonin, and AST as the variables with the largest contributions to model predictions. The consistency of these findings with previous reports supports an association between systemic inflammation, related complications, and hospitalization duration among patients with COVID-19.

Comparison of methodological approaches also revealed considerable heterogeneity in the machine learning techniques employed, including decision trees, boosting algorithms, neural networks, and regression-based models. Unlike most previous studies, the present work specifically evaluated the impact of different data representation strategies on predictive performance. The results demonstrated that although PCA enabled the identification of clinically interpretable biological patterns, dimensionality reduction was associated with a loss of predictive information and reduced model performance, indicating that important signals may be contained within complex interactions among individual laboratory variables.

4.5. Limitations of the Study

The presented methodology is associated with several important limitations arising both from the nature of the available dataset and from a partial mismatch between the applied modeling approach and the complexity of hospitalization length prediction.

First, the error analyses presented in Figures 4–6 demonstrate a clear difficulty in predicting extreme values. Quantile regression only partially mitigated this issue. A key contributing factor may be the limited number of extreme clinical observations. A cohort of 786 cases provided limited representation of rare clinical scenarios, restricting model generalizability. Although balancing the target distribution within the training dataset could potentially improve performance, such an approach is difficult to implement reliably in a cohort of this size.

Second, biochemical parameters alone had limited ability to explain variability in hospitalization duration. Although markers such as CRP, IL-6, AST, and D-dimers contributed substantially to model predictions, their informational content was insufficient to achieve high predictive accuracy. Hospitalization length is multifactorial and extends beyond laboratory profiles. Future models should incorporate broader clinical context, including arterial blood gas analysis, comorbidities, pharmacotherapy, vital signs, and imaging-derived data. Reducing variables with marginal predictive relevance may also improve model stability and interpretability.

Third, laboratory tests were obtained according to clinical indications rather than at predefined timepoints, and their timing and frequency therefore varied among patients. Moreover, the mean and median values incorporated measurements obtained throughout hospitalization. Consequently, the models should be interpreted as retrospective analyses of global feature–outcome relationships rather than tools for prospective

prediction at hospital admission. Future studies intended for early clinical prediction should use variables available at admission or within a predefined early time window.

4.6. Future Directions and Potential Applications

The obtained results indicate that biochemical parameters alone possess limited ability to explain hospitalization duration in COVID-19 patients, while still containing clinically meaningful signals enabling partial risk stratification. Consequently, the most natural direction for further development of the presented methodology involves expanding predictive models with additional clinical data sources. Particularly valuable additions could include blood gas parameters, pharmacotherapy information, comorbidities, vital signs, and imaging data. Integration of multimodal datasets could enable the identification of more complex associations with the hospital course.

An important limitation of the current approach also remains the relatively small cohort size and limited representation of extreme clinical cases. Future studies should therefore include larger multicenter datasets encompassing different pandemic waves, viral variants, and therapeutic strategies. Such analyses would improve model generalizability and allow evaluation of the stability of identified biological patterns across diverse patient populations.

Another promising direction involves the development of dynamic models incorporating real-time temporal changes in laboratory parameters. In the present study, aggregated values (means and medians) were primarily used, indirectly reflecting disease progression but not directly modeling biological dynamics. The application of temporal architectures or neural networks designed for sequential time-series analysis could potentially enable earlier detection of clinical deterioration.

Despite the limited predictive performance of PCA-based models, principal component analysis itself proved valuable from an exploratory and interpretative perspective. Identification of biological axes associated with inflammatory response, cardiorenal dysfunction, and thrombotic disturbances may provide a basis for further phenotyping studies in COVID-19 patients. In the future, similar approaches could potentially be used to identify clinical subtypes differing in prognosis or treatment response.

The models developed in the present study were retrospective and were not designed for use at hospital admission. Nevertheless, future models restricted to variables available at admission or within a predefined early time window could be evaluated as support tools for triage and hospital resource management during crisis situations. Before clinical implementation, such models would require prospective assessment and external validation in independent populations, pandemic waves, and healthcare settings.

Additionally, the use of explainable artificial intelligence methods such as SHAP highlights the possibility of developing models that are not only predictive but also clinically interpretable. Analysis of the contributions of individual biological parameters may help identify pathophysiological patterns associated with the hospital course. In this context, machine learning models may serve not only prognostic purposes but also exploratory and hypothesis-generating functions.

5. Conclusions

The XGBoost model trained on raw predictor variables achieved the best performance for modeling length of hospital stay in patients with non-severe COVID-19. SHAP analysis identified CRP, D-dimers, IL-6, procalcitonin, and AST as the variables with the largest contributions to model predictions, supporting an association between inflammatory and coagulation-related processes and hospitalization duration. Although PCA enabled the identification of clinically interpretable biological patterns, dimensionality reduction resulted in reduced predictive performance. These findings suggest that preservation of the original predictor information is advantageous for machine learning modeling, while PCA may be more useful as an exploratory tool for biological interpretation than as a preprocessing strategy. Because aggregated laboratory measurements covered the entire hospitalization, the present results represent retrospective global relationships and do not establish admission-time prognostic performance.

Author Contributions: Conceptualization, M.R. and J.K.; Methodology, M.R.; Software, M.R.; Validation, M.R., J.J. and J.B.-K.; Formal Analysis, M.R.; Investigation, J.J.; Data Curation, M.R.; Visualization, M.R.; Writing – Original Draft Preparation, M.R.; Writing – Review & Editing, J.J., J.B.-K. and J.K.; Supervision, J.K.; Project Administration, J.K.; Funding Acquisition, J.J. and J.B.-K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. All research-related expenses and article processing charges (APC) were covered from private funds of the authors.

Conflicts of Interest:
The authors declare no conflict of interest.

References

1. Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics*, *13*(15), 2582. https://doi.org/10.3390/diagnostics13152582
2. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202
3. Barbour DL. Precision medicine and the cursed dimensions. *NPJ Digit Med*. 2019;2:4. Published 2019 Feb 4. doi:10.1038/s41746-019-0081-5
4. XGBoost Documentation. (n.d.). XGBoost documentation. Retrieved May 9, 2026, from https://xgboost.readthedocs.io/en/release_3.2.0/
5. SHAP Documentation. (n.d.). SHAP documentation. Retrieved May 9, 2026, from https://shap.readthedocs.io/en/latest/
6. Rutledge, A. C., Choi, Y.-H., Karp, I., Bhayana, V., & Stevic, I. (2021)., Biochemistry tests in hospitalized COVID-19 patients: Experience from a Canadian tertiary care centre., Clinical Biochemistry, 95, 41–48., https://doi.org/10.1016/j.clinbiochem.2021.05.008
7. Zanza C, Romenskaya T, Manetti AC, et al. Cytokine Storm in COVID-19: Immunopathogenesis and Therapy. *Medicina (Kaunas)*. 2022;58(2):144. Published 2022 Jan 18. doi:10.3390/medicina58020144
8. Ebinger J, Wells M, Ouyang D, et al. A Machine Learning Algorithm Predicts Duration of hospitalization in COVID-19 patients. *Intell Based Med*. 2021;5:100035. doi:10.1016/j.ibmed.2021.100035
9. Mahboub B, Bataineh MTA, Alshraideh H, Hamoudi R, Salameh L, Shamayleh A. Prediction of COVID-19 Hospital Length of Stay and Risk of Death Using Artificial Intelligence-Based Modeling. *Front Med (Lausanne)*. 2021;8:592336. Published 2021 May 4. doi:10.3389/fmed.2021.592336
10. Olivato, M., Rossetti, N., Gerevini, A. E., Chiari, M., Putelli, L., & Serina, I. (2022). Machine learning models for predicting short-long length of stay of COVID-19 patients. *Procedia Computer Science*, *207*, 1232-1241.
11. Amiri P, Montazeri M, Ghasemian F, et al. Prediction of mortality risk and duration of hospitalization of COVID-19 patients with chronic comorbidities based on machine learning algorithms. *Digit Health*. 2023;9:20552076231170493. Published 2023 Jun 6. doi:10.1177/20552076231170493
12. Orooji, A., Shanbehzadeh, M., Mirbagheri, E., & Kazemi-Arpanahi, H. (2022). Comparing artificial neural network training algorithms to predict length of stay in hospitalized patients with COVID-19. *BMC Infectious Diseases*, *22*(1), 923.