applied sciences

MDPI

Article

A Machine Learning Multilayer Meta-Model for Prediction of Postoperative Lung Function in Lung Cancer Patients

Radomir Vešović [1,2], Milan Milosavljević [3], Marija Punt [4] and Jelica Radomirović [3,4,*]

[1] Faculty of Medicine, University of Belgrade, Dr Subotica 8, 11000 Belgrade, Serbia; rvesovic@gmail.com
[2] Clinic for Thoracic Surgery, University Clinical Center of Serbia, Pasterova 2, 11000 Belgrade, Serbia
[3] Vlatacom Institute of High Technology, Milutina Milankovica 5, 11070 Belgrade, Serbia; milan.milosavljevic@vlatacom.com
[4] School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, 11120 Belgrade, Serbia; marija.punt@etf.bg.ac.rs
* Correspondence: jelica.radomirovic@vlatacom.com

Abstract: The goal of this paper is to inform the machine learning community of our results obtained during the development of a system for the assessment of the postoperative lung function of patients suffering from lung cancer. The system is based on a new multilayer regression meta-model, which predicts individual postoperative forced expiratory volume in 1 s (poFEV1) for each patient based on preoperative measurements. The proposed regression models are especially trained to predict this key indicator for the 1st, 4th, and 7th day after surgery. Based on our knowledge, this is the first attempt to obtain poFEV1 in the most critical postoperative period of the first seven days. The high accuracy of the proposed predictive meta-model allows surgeons a number of key insights, starting with whether the patient is suitable for surgical intervention, and ending with the preparation of individualized postoperative treatment. It should be noted that the existing, widely used predictive models, based on functional segments (FC), Juhl-Forst, and Nakahara formulas, give two to three times worse results compared to the proposed new regression meta-model. Based on the SHapley Additive explanations (SHAP) value of the trained meta-model, it is possible to obtain a complete picture of the partial effects of each prognostic factor for each patient preoperatively on the outcome of the surgical intervention. In addition, the global model interpretation by SHAP values reveals some new interdependencies that were not known in medical circles until now. For instance, the influence of age and biomass index on the condition of the patient on the first day after surgery, or the constant significant influence of muscular support for inhalation in the entire seven-day follow-up period.

Keywords: machine learning; stacked learning; forced expiratory volume; SHAP; personalized treatment

Citation: Vešović, R.; Milosavljević, M.; Punt, M.; Radomirović, J. A Machine Learning Multilayer Meta-Model for Prediction of Postoperative Lung Function in Lung Cancer Patients. *Appl. Sci.* **2024**, *14*, 1566. https://doi.org/10.3390/app14041566

Academic Editor: Hariton-Nicolae Costin

Received: 15 January 2024
Revised: 7 February 2024
Accepted: 10 February 2024
Published: 15 February 2024

1. Introduction

Pulmonary resection is the standard procedure in the treatment of early-stage lung cancer [1,2]. Assessment of postoperative lung function of patients is one of the central problems of clinical surgical practice in this domain for several reasons:

1. Risk Stratification: The preoperative identification of patients who may be at higher risk for postoperative complications allows the medical team to implement appropriate measures to minimize complications and optimize outcomes.
2. Surgical Planning: Knowing the assessment of lung function helps the surgical team plan the surgical procedure effectively.

3. Informed Consent: Obtaining informed consent is crucial, as patients need to be aware of the possible postoperative challenges and complications they may face based on their individual lung function.
4. Predicting Postoperative Outcomes: This information helps patients prepare for the recovery process and facilitates appropriate postoperative care planning.
5. Risk-Benefit Evaluation: For some patients with compromised lung function, the risks associated with surgery may outweigh the potential benefits. Postoperative lung function assessment helps to make informed decisions regarding the appropriateness of surgery as a treatment option for individual patients.
6. Postoperative Care Planning: Knowing the postoperative lung function enables healthcare providers to plan for appropriate postoperative care. It helps determine the need for intensive care, ventilatory support, or physiotherapy during the recovery period.

In summary, the preoperative assessment of postoperative lung function is essential to identify high-risk patients, optimize surgical planning, set realistic expectations, and provide appropriate care before and after pulmonary resection. This approach improves patient safety, enhances outcomes, and helps ensure that the surgical procedure is performed with the highest level of personalized care [3,4].

Forced Expiratory Volume in 1 s (FEV1) is a measure widely used in pulmonary function tests to assess lung function [4]. The test involves taking a deep breath in and then exhaling as quickly and forcefully as possible to measure the volume of air expelled within that initial one-second time frame. FEV1 was expressed in liters [L] or as a percentage [%] of the predicted value for age, gender, and height, according to the European Community for Steel and Coal prediction equations [5]. It is essential to note that FEV1 is just one component of pulmonary function testing, and additional parameters like Forced Vital Capacity (FVC) and the FEV1%/FVC ratio are often considered together to gain a comprehensive understanding of lung health.

The predicted postoperative FEV1 (ppoFEV1) is today dominantly used in the assessment of postoperative lung function after pulmonary resection of patients suffering from lung cancer. For this purpose, various measurements and data that can be obtained preoperatively are used, such as the number of resected lung segments, quantitative computed tomography (CT), spirometry, or perfusion scintigraphy [4,6–8]. The common feature of all these methods is that they are tested for a prediction horizon of three to six months after lung resection [4]. Contrary to such predictions, ppoFEV1 for a time horizon of up to seven days would represent a key indicator of the surgery outcome since most cardio-respiratory complications are developed during that period. At the same time, that period coincides with the typical length of the patient's stay in the hospital, when various measures can be taken to reduce the risk of an undesirable outcome. Unfortunately, in the available literature, there are very few results related to this immediate postoperative period. In the paper [9], it was stated that the existing methods for calculating predicted poFEV1 significantly underestimate the loss of lung function in the immediate postoperative period of 6 days.

From these facts, the basic motivation and goal of our work follows: the synthesis of a machine learning (ML) system for predicting postoperative FEV1 within the first 7 days after undergoing pulmonary resection, individually for each patient. For this study, we collected necessary measurements and data from 79 patients who underwent pulmonary resection. Each patient is described with an initial 35 features, which include the patient's age, type of operation, cancer location, various spirometry measurements, the number of lung segments to be removed, as well as a series of measurements related to the mobility of the left and right hemidiaphragm, etc. In order to be included in the study, the presence of primary cancer was also necessary, as well as a complete assessment of the functional status and overall cardiorespiratory risk.

The proposed machine learning system is especially trained to predict postoperative FEV1 for the 1st, 4th, and 7th day after surgery. To our knowledge, this represents the first

published result that predicts postoperative FEV1 in the most critical postoperative period of the first 7 days. The high accuracy of the proposed predictive meta model, with a mean absolute error (MAE) ranging from 8% to 11% and mean absolute percentage error (MAPE) ranging from 14% to 23%, provides surgeons with a number of key insights, starting with whether the patient is suitable for surgical intervention, and ending with the preparation of individual postoperative treatment.

A considerable benefit for clinical practice is the doctor's precise insight into the impact of each input feature on the resulting ppoFEV1 individually for each patient before surgery. This insight can be obtained by calculating SHAP values based on the obtained regression model [10,11]. The additive nature of SHAP values and simple interpretation, in our opinion, provide great opportunities for significant improvement of the clinical practice of lung cancer treatment, especially in terms of more precise preoperative insight into possible risks, as well as in the personalization of the recovery procedure. It should be kept in mind that the validity of using SHAP values for these purposes is critically conditioned by the accuracy of the given predictive model.

Preliminary results of our research intended for medical circles, were published earlier [12]. The central question analyzed in [12] was the confirmation of the significant influence of inspiratory respiration muscles on postoperative prediction of the lung function in the first seven postoperative days. The goal of this paper is to provide the machine learning community with our results obtained during the development of the system [12]. The rest of this paper is organized as follows. In Section 2, the synthesis procedure of the regression prediction model for postoperative FEV1 is given. In order to achieve the highest possible accuracy, it was necessary to complicate the architecture of the predictor. Therefore, we examined entire classes of multi-layer meta-models, whose individual elementary blocks are basic machine learning regression models. The choice of architecture was dominantly conditioned by the fact that we had a training set of 79 instances with 35 initial features at our disposal. Section 3 presents the results of the experimental evaluation of the proposed prediction model, while Section 4 provides a detailed interpretation of the model based on SHAP values. In the concluding section, the conditions for the broader application of these results in clinical practice are commented upon.

2. Materials and Methods

Prediction of Postoperative FEV1 by Multilayer Regression Meta-Model

Let us denote with X, the vector of preoperative characteristics measured for each patient individually, and by y, the poFEV1 on the given day after pulmonary resection. At our disposal is a training set (X_i, y_i), $i = 1,2,\dots,N$ obtained by monitoring the past N patients who underwent this surgical operation. Then, within the framework of the machine learning methodology, the synthesis of a predictor

$$\hat{y}_i = F(X_i; \theta), \tag{1}$$

is performed by minimizing the selected criterion function

$$J(\theta) = \sum_{i=1}^{N} d(y_i, \hat{y}_i), \tag{2}$$

in the parameter space θ of the model F, where d is a suitably chosen distance measure between the actual and predicted values. Since FEV1 is a continuous variable, model (1) is of a regression type. The most commonly used distance measures d are:

$$d_{AE}(y_i, \hat{y}_i) = |y_i - \hat{y}_i|, \tag{3}$$

which gives the Mean Absolute Error (MAE) criterion J_{MAE},

$$d_{PE}(y_i, \hat{y}_i) = \left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\ [\%], \tag{4}$$

giving the Mean Absolute Percentage Error (MAPE) criterion J_{MAPE},

$$d_{SQ}(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2, \quad (5)$$

which gives the Mean Square Error (MSE) criterion J_{MSE}.

In this paper, we will examine the effectiveness of model F when it belongs to the stacked regressor class, [13,14]. Stacked regression, also known as stacked ensemble regression, is a powerful technique that combines the predictions of multiple base regression models to produce a more accurate and robust final prediction. It offers several advantages over using just a single base regression model:

1. *Improved Predictive Performance*: Stacked regression leverages the strengths of different base regression models by blending their predictions together. As a result, it can capture complex patterns and relationships in the data that individual models might miss, yielding a performance better than any single base regression models [12].
2. *Reduced Overfitting:* Base regression models may suffer from overfitting if they are too complex or trained on limited data. Stacking helps reduce overfitting by blending the predictions from multiple models, effectively smoothing out any model-specific noise and generalizing better to unseen data.
3. *Model Diversity:* To benefit from stacked regression, it is important to use diverse base regression models. Diversity can be achieved by training different models with varying algorithms, hyperparameters, or feature subsets. When combined, these diverse models contribute complementary information, leading to a more robust and reliable ensemble.
4. *Handling Nonlinearity:* Stacked regression is particularly effective at handling nonlinearity in the data. The individual base regression models might be limited in their ability to capture nonlinear relationships, but by combining them, the stacked ensemble can approximate more complex patterns.
5. *Adaptability:* Stacked regression can be applied to a wide range of regression problems, including those with high-dimensional data, outliers, and complex interactions among features. It can be adapted to different types of regression algorithms, such as linear regression, decision trees, support vector regression, or neural networks.
6. *Handling Model Biases:* Different base regression models may have their biases and limitations. Stacked regression can mitigate these biases by considering a variety of perspectives, leading to a more well-rounded and reliable final prediction.

A typical stacked regression architecture involves using multiple layers of regression models to make predictions. The two main components in this architecture are the base regressions and the meta-model, see Figure 1. The first layer comprises base regression models. The second layer consists of the meta-model. The input to the meta-model is the output of base regression models from the first layer and some selected input features. This architecture is directly generalized to a multi-layered architecture, in which there are multiple layers of individual regression models between the input training data and the output meta-model.

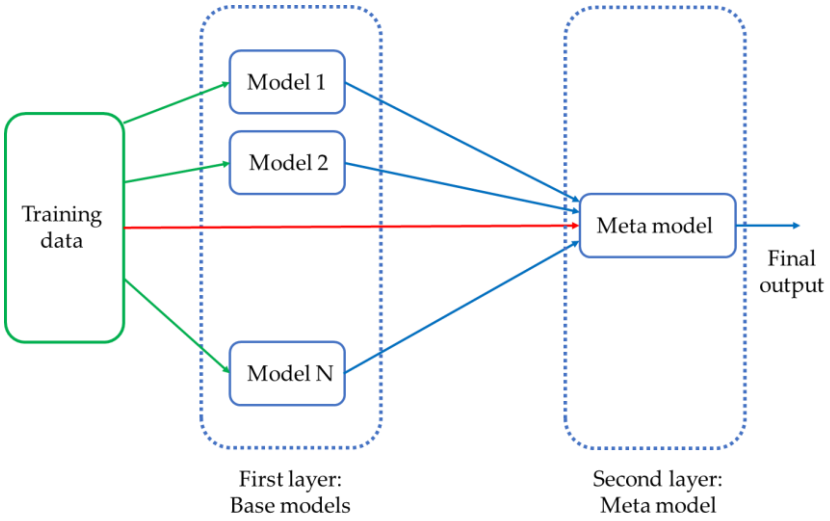

Figure 1. Stacked regression. The first layer comprises base regression models. The second layer consists of the meta-model. The input to the meta-model is the output of base regression models from the first layer (denoted by the blue arrows) and some selected input features (denoted by the red arrow).

Algorithms 1 and 2 formally show the stacked regression training algorithm with the simultaneous estimation of hyperparameters of individual base and meta-models. The output of the procedure is the average performance of stacked regression H_R over K *Testing sets*, (Step 2 in Algorithm 1), as well as the optimal hyperparameters θ_j^* (Step 1.2, Algorithm 1). Algorithm 2 shows separately the algorithm for training stacked regression H_R with additional L-fold cross-validation [15,16].

The entire Algorithm 2 actually corresponds to Step 1.1.a.3 of the algorithm from Algorithm 1. The goal of additional cross-validation is to utilize the entire training dataset efficiently and to get a more reliable estimate of the meta-model performance. The L-fold cross-validation process involves the following steps (see Algorithm 1 for a more detailed description):

Step 1: Divide the original *Training validation set* V into L equal-size subsets.

Step 2: For each fold l (where l ranges from 1 to L):

1. Use L-1 folds for training the base regressions and make predictions on the remaining fold l (so-called *out-of-fold predictions*).
2. Store these predictions as meta-features for fold l. These are the inputs for the meta-model in this specific fold.

Step 3: Once the predictions (meta-features) for all L folds have been obtained, combine them to create a new dataset, the *meta-training dataset*.

Step 4: Train the meta-model on the *meta-training dataset*, where the target variable is still the actual target values from the original training dataset.

A *meta-training dataset* consisting of the *out-of-fold predictions* formed in Step 2 prevents the leakage of target information into it, thus reducing the possibility of overfitting.

Algorithm 1: Training stacked regression with hyperparameter tuning and model performance evaluation and SHAP values calculation

---

Input: Training data $D = \{(X_i, y_i), i = 1,2, \ldots, N\}$, T—number of base models

Output: A stacked regression H_R

Randomly partition D into K equal size subsets $D = \{D_1, D_2, \ldots, D_K\}$

1: **for** $k = 1,2, \ldots, K$ **do**

Training set $\leftarrow D \setminus D_k$

Testing set $\leftarrow D_k$

1.1: **for** $i = 1,2,\dots,m$ **do**
$m = m_1 m_2 \cdots m_{T+1}$; m_j – number of distinct hyperparameters θ for j-th base model; m_{T+1} denotes the number of distinct hyperparameters for meta-model
a: Repeat K-1 times only for samples in the training set
a.1: *Training validation set* $\leftarrow$ K-2 subsets
a.2: *Testing validation set* $\leftarrow$ remaining subset
a.3: Train H_R on the *Training validation set* using hyperparameter θ_i
a.4: Test H_R on the *Testing validation set*
b: Record $J(\theta_i)$, the average performance of H_R over K-1 *Testing validation sets*
end for
1.2: Determine θ_j^*, where $j = \underset{i}{\operatorname{argmax}} J(\theta_i)$
1.3: Train the stacked regression H_R on the *Training set* using hyperparameter θ_j
1.4: Test the stacked regression H_R obtained in step 1.3 on the *Testing set*
1.5: Calculate SHAP values on the *Testing set*
end for
2: Return the average performance of stacked regression H_R over K *Testing sets*
3: Return the SHAP values of all K *Testing sets* (local feature importance for each observation in D)
4: Return the SHAP values averaged over all K *Testing sets* (global feature importance over entire D)

Since the complete training procedure for finding the optimal hyperparameters is extremely complex, a common simplification consists of omitting the finding of the optimal hyperparameters of the base and/or meta-models. In that case, Steps 1.1 to 1.3 in Algorithm 1 are omitted, while *Training validation* set V in Algorithm 2 is replaced by the entire *Training set* defined in Algorithm 1.

The stacked regression model with two layers can be directly extended to an arbitrary number of layers, in which each previous layer generates new features for the next layer of the models. In order to obtain an out-of-fold training set for the final meta-model, multi-layer models require an additional nested cross-validation loop for each additional layer. It is clear that this type of complexity of architecture and training can be practically feasible only in the case of sufficient computing resources and large training sets. Since our training data has only 79 instances, we decided on the simplest two-layer architecture from Figure 1, and the choice of default hyperparameters of all used models, both in the first and in the second layer of the architecture.

Algorithm 2: Extension of Step 1.1.a.3 of algorithm from Algorithm 1. Training stacked regression H_R with additional L fold cross validation

Input: *Training validation set*; hyperparameter θ_i, see a.3 in Algorithm 1
Output: H_R with hyperparameter θ_i, trained on *training validation set*
Randomly partition *Training validation set* V into L equal size subsets
V= $\{V_1, V_2, \dots, V_L\}$
a.3.1: **for** $l = 1,2,\dots,L$ **do**
Training set base model $\leftarrow V \setminus V_l$
Test set base model $\leftarrow V_l$
Training first level (base) models; T is the number of base models
a.3.1.1: Repeat for $t = 1,2,\dots,T$;
Train model h_{lt} on *Training set base model*
Construct a training set for second level meta-model
a.3.1.2: Get predictions $X_{meta\,l} = \{h_{l1}, h_{l2}, \dots, h_{lT}\}$ on *Test set base model*

end for

a.3.2: Training second level meta-model

Train a new stacked regression $H_R{}'$ on collection $\{X_{meta\,l}, y_l\}, l = 1,2,\dots,L$

a.3.3: Re-train first level base models

for $t = 1,2,\dots,T$ **do**

Train first level base models h_t on Training_validation set V

end for

a.3.4: Return $H_R(x) = H_R{}'(h_1(x), h_2(x), \dots, h_T(x))$

3. Experimental Evaluation

3.1. Collected Data and Feature Engineering

Data collection was confined to a one-year period during which we aimed to collect as diverse data as possible. Consequently, data of 79 patients who underwent surgery were collected. In order to be included in the study, patients had to fully cooperate during the measurement of diaphragmatic movements. The presence of a primary cancer diagnosis was also necessary, as well as a complete assessment of the functional status and overall cardiorespiratory risk. Movements of both hemidiaphragms were measured radiographically and ultrasonographically, along with muscle strength tests and respiratory function, preoperatively. In the initial stage of feature selection, we chose features that reflected the general medical state of the patient, including measurements related to diaphragm movement, as well as those necessary for prediction of FEV1 using traditional methods. Subsequently, domain experts performed additional filtration of the obtained feature set. This involved elimination of highly dependent features, as well as those already identified as lacking significant predictive value for FEV1. Further dimensionality reduction methods based on feature transformation (e.g., PCA) was not performed to maintain model interpretability. After data collection and feature selection, we were left with 25 most important variables, which were used to design the prediction model. Among them, 15 features were represented with absolute value, and the other 10 are represented with absolute and relative value, expressed as a percentage, see Table 1.

For the feature vector X_i, we selected 25 features, omitting ten features expressed in absolute measures, while keeping those same features expressed in relative values. Discarded features are marked in light ocher yellow in Table 1. This is a consequence of preliminary experiments, from which we concluded that the accuracy of the proposed meta-model was slightly higher when we choose features expressed in relative values. In order to have a complete insight into the feature engineering process, we retained the complete initial set of 35 features in Table 1.

Table 1. Feature description.

No.	Feature Label	Description	Type (Value Range)	Mean ± std
1	B	age	Integer [40, 78]	60.24 ± 7.31
2	E1	Type of the operation on the right lung	Categorical	
3	E2	Type of the operation on the left lung	Categorical	
4	K	BMI ($\frac{\mathrm{kg}}{\mathrm{m}^2}$)	Float [17.06, 35.25]	26.07 ± 1.84
5	O	Type of respiratory function	Categorical	
6	P	COPD (Chronic Obstructive Pulmonary Disease) index	Float [0.9210, 2.2166]	1.66 ± 0.22
7	S	Preoperative FEV1—Preoperative forced expiratory volume in the first second (L)	Integer [1570, 4350]	2656.96 ± 510.98

8	T	Preoperative FEV1%—Preoperative forced expiratory volume in the first second in [%]	Integer [45, 144]	94.53 ± 15.63
9	U	Preoperative VC—Preoperative vital capacity (L)	Integer [2040, 6780]	3968.99 ± 872.34
10	V	Preoperative VC %—Preoperative vital capacity in [%]	Integer [76, 148]	109.23 ± 15.87
11	W	Preoperative FVC—Preoperative forced vital capacity (L)	Integer [2010, 5960]	3754.05 ± 755.19
12	X	Preoperative FVC %—Preoperative forced vital capacity in [%]	Integer [73, 143]	107.86 ± 15.04
13	Y	Preoperative VCin—Preoperative vital capacity in inspiration (L)	Integer [2070, 6020]	3822.78 ± 794.46
14	Z	Preoperative VCin %—Preoperative vital capacity in inspiration in [%]	Integer [76, 143]	105.51 ± 14.20
15	AA	Preoperative FEV1%/FVC	Float [47.10, 97.36]	71.32 ± 9.41
16	CD	TLC—Total lung capacity (L)	Integer [4100, 9650]	6953.16 ± 1220.52
17	CE	TLC %—Total lung capacity in [%]	Integer [90, 160]	116.27 ± 14.41
18	CF	RV—Residual volume (L)	Integer [770, 5670]	2988.10 ± 805.40
19	CG	RV %—Residual volume in [%]	Integer [45, 292]	137.13 ± 34.03
20	CH	FRC (ITGV)—Functional residual capacity (L)	Integer [2090, 6200]	4227.72 ± 922.76
21	CI	FRC (ITGV) %—Functional residual capacity in [%]	Integer [78, 223]	133.67 ± 27.32
22	CJ	RV/TLC (% predicted)	Integer [48, 178]	109.49 ± 20.25
23	CK	FRC (ITGV) % (% predicted)	Integer [39, 151]	107.51 ± 18.53
24	CY	Mobility of the right hemidiaphragm measured radiographically (cm)	Float [0.9, 4.3]	4.16 ± 1.41
25	DB	Mobility of the left hemidiaphragm measured radiographically (cm)	Float [0.4, 4.70]	4.08 ± 1.39
26	EK	Mobility of the right hemidiaphragm measured by ultrasound (mm)	Float [47.9, 94.2]	68.25 ± 10.28
27	EL	Mobility of the left hemidiaphragm measured by ultrasound (mm)	Float [36.5, 95.0]	62.58 ± 11.10
28	FK	Preoperative PImax (cmH20)—Preoperative maximum inspiratory pressure (cmH2O)	Integer [26, 154]	83.38 ± 28.08
29	FL	Preoperative PImax %—Preoperative maximum inspiratory pressure in [%]	Float [35.75, 201.60]	109.63 ± 35.76
30	FS	Preoperative PEmax (cmH20)—Preoperative maximal expiratory pressure	Integer [43, 155]	102.22 ± 25.16
31	FT	Preoperative PEmax %—Preoperative maximal expiratory pressure in [%]	Float [46.16, 129.84]	92.19 ± 18.04

32	GA	Preoperative Snip (cmH20)—Preoperative „sniff" inspiratory pressure (cmH20)	Integer [20, 139]	86.53 ± 25.09
33	GB	Preoperative Snip %—Preoperative „sniff" inspiratory pressure in [%]	Float [24.61, 143.90]	91.83 ± 23.26
34	LU	The number of functional segments removed by the operation	Integer [1, 9]	3.41 ± 2.02
35	LV	The number of total functional segments in the lungs	Integer [14, 18]	17.23 ± 1.07

Features marked in ocher yellow are discarded in experiments

Postoperative FEV1 was measured on the first, the fourth, and the seventh day after the surgery. These variables were selected for output variable y in the model (1), see Table 2 and Figure 2. From these data, we can observe the typical dynamics of lung function recovery. Immediately after surgery, patients experience a decrease in lung function due to factors such as lung volume loss, anesthesia, pain, limited mobility, and the surgical trauma itself. By the fourth day, lung function tends to improve. By the seventh day post-surgery, the measured FEV1[%] should continue to improve. Pain management, early mobilization, and breathing exercises have a significant impact on helping patients regain lung function.

Table 2. Mean values and variances of measured postoperative FEV1 in [%] at the 1st, 4th, and 7th day after surgery.

	1st Postoperative Day	4th Postoperative Day	7th Postoperative Day
mean	44.68	50.95	58.01
std	14.07	15.80	14.78

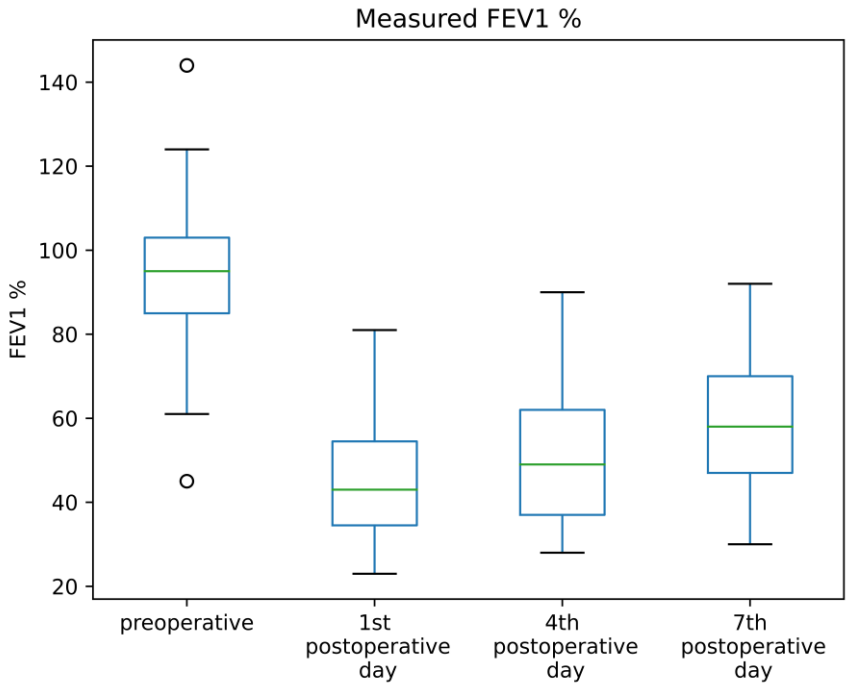

Figure 2. Box plots of preoperative and postoperative FEV1 in [%] measured at 1th, 4th, and 7th day after surgery.

3.2. Selection of Existing Methods for Comparison

In order to assess the accuracy of our model we compared it with existing methods of calculating predicted and postoperative FEV1 in [%], based on preoperative measurements. We limited ourselves to three basic methods: functional segments [17], Juhl-Frost [18], and Nakahara [19]:

$$ppoFEV1_{FC} = FEV1\left(1 - \frac{LU}{LV}\right)[\%], \quad (6)$$

where LU is the number of functional lung segments removed and LV is the total number of functional lung segments,

$$ppoFEV1_{Juhl-Frost} = FEV1\left(1 - S \cdot \frac{5.26}{100}\right)[\%], \quad (7)$$

where S is the number of lung segments removed,

$$ppoFEV1_{Nakahara} = FEV1\left(1 - \frac{n-a}{42-a}\right)[\%], \quad (8)$$

where n is the number of resected sub-segments in the lobe, that is, 6, 4, and 12 for the right upper, middle, and lower lobe and 10 for the left upper and lower lobe, while a is the number of sub-segments obstructed by the tumor.

3.3. Selection of Base and Meta-Models

Base models can be of different types and architectures, such as decision trees, random forests, support vector machines, neural networks, etc., [14,15,20]. The criterion for choosing base models depends on several factors:

Diversity: It's essential to choose base models that have different strengths and weaknesses. Models that make different types of errors or have varying biases can complement each other in the ensemble, leading to improved overall performance.

Performance: While diversity is important, base models should still demonstrate reasonable predictive performance on their own. Models that perform well individually are more likely to contribute positively to the ensemble.

Computational efficiency: Depending on the size of the data and available computing resources, the computational cost of training and forecasting with underlying models should be considered.

The criterion for choosing the meta-model includes [15,20]:

Performance: The meta-model should be chosen based on its ability to effectively combine the predictions from the base models. Common meta-models include linear regression, logistic regression, or more complex models like random forest or gradient boosting machines.

Complexity: Simpler meta-models are preferred over complex ones, as they are less prone to overfitting and can generalize better on new data.

Interpretability: Depending on the application, interpretability might be important. If the interpretability of the final model is a requirement, a meta-model that is more transparent and provides insights into how the ensemble makes predictions should be selected.

Taking into account the amount of our training data, as well as the choice of SHAP as an agnostic method for interpretation, the dominant criteria was reduced to diversity and performance for base models and performance for meta-models. We have included 12 basic models in the wider list: Lasso (Least absolute shrinkage and selection operator) [21], Extra Tree (Extremely randomized Trees) [22], Random Forest [23], LightGBM (A Highly efficient gradient boosting decision tree) [24], SVM (Support Vector Machine) for regression, [25], in two variants SVM Linear and SVM.RBF with linear and radial bases kernel, respectively, AdaBoost [26], KNN (K Neighbors Regressor), [27], MLP (Multi-Layer Perceptron), [28], with two variants of architecture (MLP1-three layer architecture, MLP2-two layer architecture), Ridge Regression, [29], XGBoost (eXtreme Gradient Boosting), [30] and LogisticReg (Logistic Regression) [31].

We chose Random Forest as the output meta-model, due to its good generalization properties and ability to model a wide class of nonlinear mappings. In addition, Random Forest has an advantage over a similar class of models, such as LightGMB, AdaBoost, and XGBoost, primarily in terms of robustness to the selection of initial parameters and the built-in bootstrapping mechanism. These advantages are important in the case of very short training sets. Furthermore, compared to Extreme Trees, Random Forest demonstrates superior precision in selecting subsets of features during the node-splitting process of growing individual trees. The final selection of Random Forest over competing models was the result of experimental evaluation.

Table 3 shows the performance (MAPE and MAE) of individual candidates for the baseline models obtained by 5-fold cross-validation and then averaged for all three prediction tasks: 1st, 4th, and 7th days after surgery. The baseline models were taken with default parameters, since the amount of data at our disposal did not allow hyperparameter optimization. All experimental results were obtained using the Scikit learn package [32], LightGBM package [33], and XGBoost package [34]. In order to get a clearer picture of the selected basic algorithms, we list some of the most important default parameters: Lasso (alpha = 1, max_iter = 1000), ExtraTrees (n_estimators = 100, criterion = gini, max_features = sqrt), RandomForest (n_estimators = 100, criterion = gini, max_features = sqrt), LightGBM (boosting_type = gbdt, n_estimators = 100, num_leaves = 31), SVM (kernel = linear, C = 1, tol = 0.001), SVM (kernel = rbf, C = 1, tol = 0.001, gamma = scale), AdaBoost (n_estimators = 50, leakage_rate = 1), KNN (n_neighbors = 5, metric = minkowski, p = 2), MLP (hidden_layer_size = {7,3,2}, activation = relu, solver = Adam, max_iter = 200), , MLP (hidden_layer_size = {3,2}, activation = relu, solver = Adam, max_iter = 200), Ridge (alpha = 1, solver = auto), XGBoost (see [34]), LogisticRegression (penalty = l2, solver = lbfgs, max_iter = 100). The light ocher yellow indicates nine models that were included in the final architecture. The omission of the MLP models was primarily motivated by the large number of free parameters compared to the available length of the training sets. Additionally, our results confirmed the already noted fact (see [35,36] that despite the significant progress in the application of deep neural networks for tabular data, they are still outperformed by tree-based models on many standard benchmarks. The last row of Table 3 shows the corresponding performance of the meta-model for the selected nine base regressors: Lasso, Extra Tree, Random Forest, SVM Linear, KNN, Ridge Regression, XGBoost, SVM RBF, and LogisticReg. As expected, the meta-model gave better results for both performances (MAPE and MAE) compared to any base model.

Table 3. Performance (MAPE and MAE) of individual candidates for the baseline models obtained by 5-fold cross-validation and then averaged for all three prediction tasks: 1st, 4th, and 7th days after surgery. The last row shows the corresponding performance of the meta-model, for the selected ten base regressors.

No.	Model	MAPE	MAE
1	Lasso	19.57	9.05
2	Extra Tree	20.27	9.55
3	Random Forest	20.44	9.66
4	LightGBM	20.96	9.80
5	SVM Linear	21.11	9.83
6	AdaBoost	20.90	9.93
7	KNN	22.46	10.23
8	MLP2	22.95	10.66
9	Ridge Regression	23.73	10.78
10	XGBoost	22.72	11.10
11	SVM.RBF	23.82	11.39
12	LogisticReg	23.89	11.73
13	MLP1	25.33	11.84

14	Meta-model	18.64	8.93

The baseline models used for meta-model are highlighted

To minimize the risk of overfitting we wish to highlight the taken measures:

- Selection of base regressors dominantly employing ensemble methods
- Selection of Random Forest as the final regressor
- Nested cross-validation (see Algorithm 2), ensuring the training of stacked regressor H_R, in an additional L-fold cross-validation. It is well known that nested cross-validation minimizes the amount of information leakage between the training and validation sets [37].

Table 4 shows the performance of the proposed meta-model, along with the performances of three standard models (Functional segments, Juhl-Frost, Nakahara), obtained with 5-fold cross-validation. In order to quantitatively measure the gain of the meta-model, we introduced the values $Gain\_MAPE$, i.e., $Gain\_MAE$ defined as the ratio of the respective performances of the compared model M with the meta-model:

$$Gain\_MAPE = \frac{J_{MAPE}(M)}{J_{MAPE}(meta\ model)}, \quad (9)$$

$$Gain\_MAE = \frac{J_{MAE}(M)}{J_{MAPE}(meta\ model)}, \quad (10)$$

Table 4. Performances of the proposed meta-model, along with performances of three standard models (Functional segments, Juhl-Frost, Nakahara).

	1st Day after Surgery		**4th Day after Surgery**		**7th Day after Surgery**	
	MAPE	**Gain_ MAPE**	**MAPE**	**Gain_ MAPE**	**MAPE**	**Gain_ MAPE**
Our meta-model	**19.36 ± 1.11**		**22.16 ± 2.79**		**14.40 ± 1.96**	
Functional segments	80.32 ± 6.00	4.15	58.14 ± 6.20	2.62	35.64 ± 2.81	2,47
Juhl-Frost	66.43 ± 4.20	3.43	46.42 ± 4.72	2.09	27.47 ± 3.60	1.91
Nakahara	76.28 ± 5.46	3.94	54.59 ± 5.33	2.46	32.58 ± 3.06	2,62
Average gain		3.84		2.39		2.21
	1st Day after Surgery		**4th Day after Surgery**		**7th Day after Surgery**	
	MAE	**Gain_ MAE**	**MAE**	**Gain_ MAE**	**MAE**	**Gain_ MAE**
Our meta-model	**8.24 ± 0.93**		**10.56 ± 0.87**		**7.98 ± 1.51**	
Functional segments	31.30 ± 2.44	3.80	25.40 ± 2.51	2.41	18.86 ± 1.01	2.36
Juhl-Frost	25.87 ± 1.79	3.14	20.03 ± 2.32	1.90	14.32 ± 1.64	1.79
Nakahara	29.57 ± 2.06	3.59	23.65 ± 2.40	2.34	17.11 ± 1.25	2.14
Average gain		3.51		2.18		2.10

From Table 4, we concluded that our meta-model gave significantly more accurate results for ppoFEV1[%] than traditional calculations. By analyzing the gains, we can conclude the following:

1. The proposed meta-model in the first postoperative day gave a gain of more than three times compared to any traditional method, both for MAPE and MAE criteria (average Gain_MAPE = 3.84, average Gain_MAE = 3.51). Among traditional prediction methods, Frost's model performed best, compared to which the meta-model has Gain_MAPE = 3.43, i.e., Gain_MAE = 3.14.
2. On the fourth and seventh postoperative days, the gain was over two times, more precisely an average Gain_MAPE = 2.39 for the 4th day and average Gain_MAPE = 2.21 for the 7th day, i.e., Gain_MAE = 2.18 for the 4th day and Gain_MAE = 2.10 for the 7th day. For these days, Frost's method proved to be the best of the traditional methods, giving Gain_MAPE = 2.09 and Gain_MAPE = 1.91, for the 4th and 7th day, respectively. The corresponding values for Gain_MAE are 1.90 and 1.79, respectively.
3. The superior advantage of the meta-model in the first postoperative day has the greatest clinical value since it can be used to project optimal recovery during the hospital stay, which usually ends after 7 days.

In evaluating the effectiveness of regression predictive models, residuals as the difference between predicted and actual values

$$e_i = y_i - \hat{y}_i \, , i = 1,2, \ldots , N \tag{11}$$

are also informative. Figures 3–5 show box plots of residuals for our meta-model and the three traditional methods, for the 1st, 4th, and 7th day after operation.

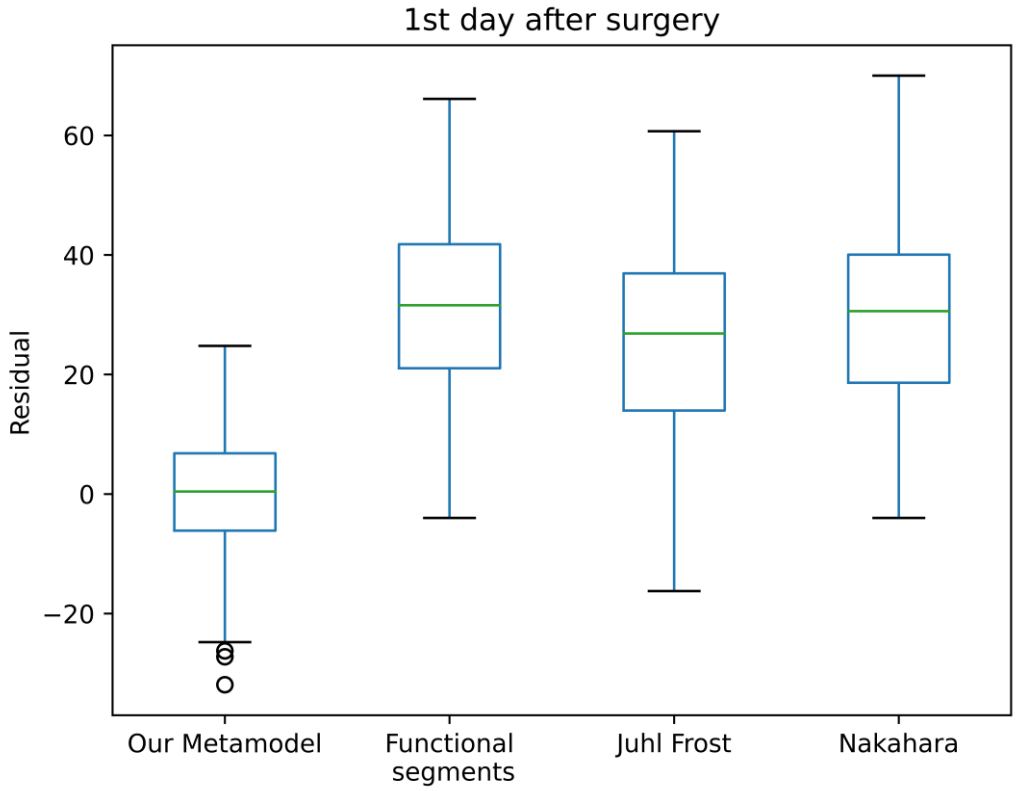

Figure 3. Residuals for the first postoperative day prediction.

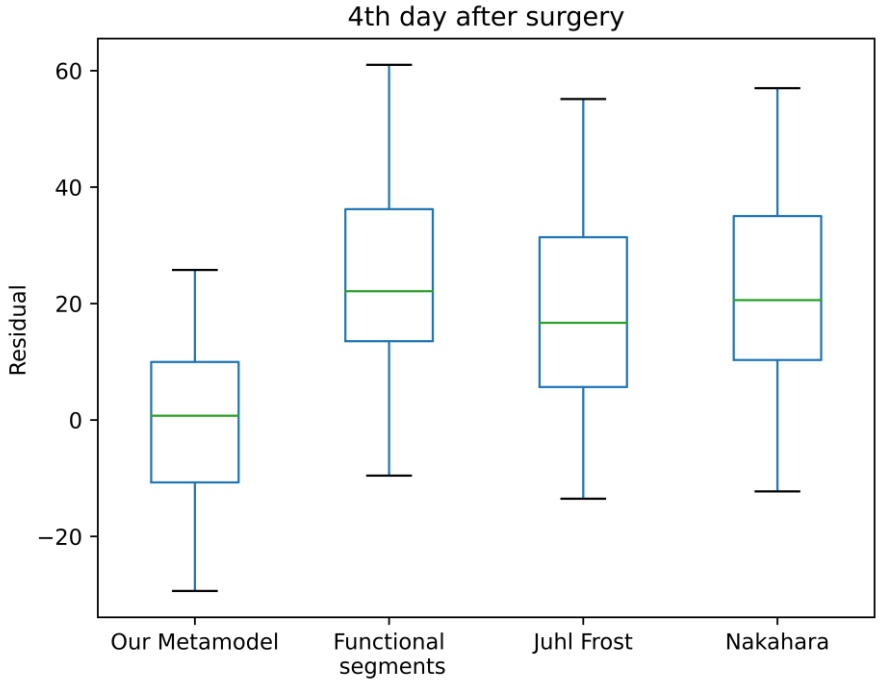

Figure 4. Residuals for the fourth postoperative day prediction.

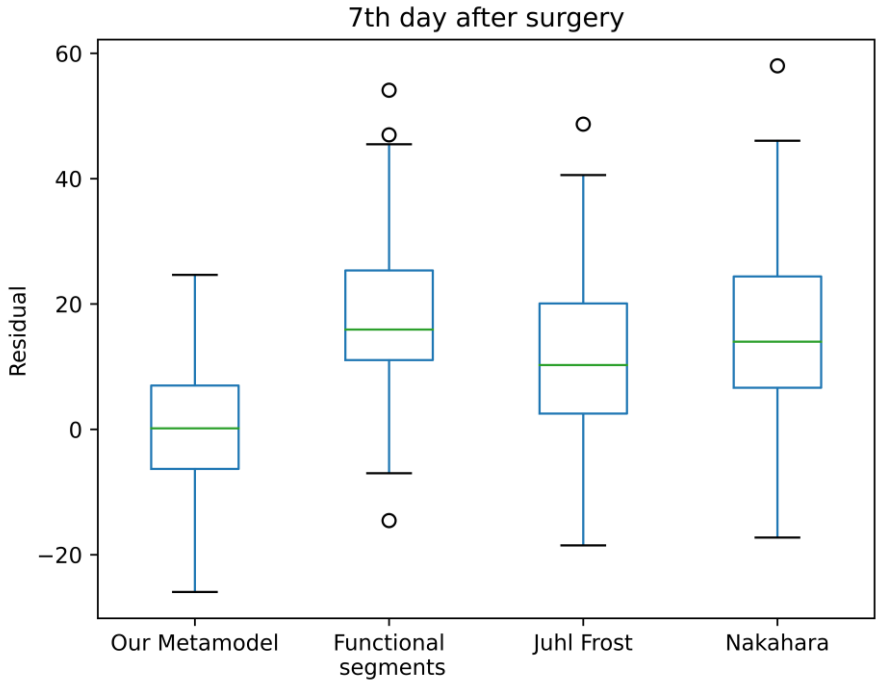

Figure 5. Residuals for the seventh postoperative day prediction.

These results show that, unlike traditional methods, the proposed meta-model gave residuals whose expected value is close to zero. This property is an important characteristic in well-fitted regression models that capture the underlying relationship between input and output variables. However, it is important to note that having zero mean residuals is not always a strict requirement for a good regression model. What is more important is that the residuals are centered around zero and do not exhibit any systematic patterns, such as trends or heteroscedasticity. Analyzing the box plots from Figures 3–5, we concluded that our meta-model has good properties in this respect as well.

4. Model Interpretation

Model interpretation is essential in the field of medicine to ensure patient safety, support clinical decision-making, identify biases, improve model performance, and foster trust between ML systems and healthcare professionals. It is a key component for the safe and effective application of the full ML potential in improving healthcare practices [38–40].

In order to demonstrate the effectiveness of the latest agnostic methods and tools for interpreting complex ML models, we opted for SHAP due to its good theoretical and practical properties. It is based on cooperative game theory by assigning an importance value for every sample to each feature [11]. In this context, agnostic means that SHAP can be applied to any black-box ML model without requiring knowledge of its internal architecture or parameters. These properties allow SHAP to have both local and global levels of model interpretation. Global explanation aims to provide insights into the overall behavior and trends of a ML model across the entire dataset. It highlights the general relationships between input features and model predictions, revealing which features tend to have more significant impacts on predictions on average. On the other hand, local explanation focuses on explaining the prediction made by the model for a specific instance or observation. It provides insights into why the model arrived at a particular decision for that individual case.

All experimental results were obtained using the package described in [41]. The procedure for calculating SHAP values for a given meta-model is presented in Algorithm 1. In Step 3, the algorithm returns SHAP values for each feature and each observation in the training data set. In Step 4, the algorithm calculates an average of SHAP values over the entire training data set for each feature.

4.1. Global Model Interpretation

Within the global interpretation, Figure 6 shows the SHAP value plot for all training data corresponding to the 1st postoperative day. The input for this plot is data from Step 3, Algorithm 1. Variables are ranked in descending order, while the horizontal location shows whether the effect of that value is associated with a higher or lower prediction. The vertical ordering of the points for each feature column should reflect the corresponding density of accumulation. The color shows whether that variable is high or low for that particular observation.

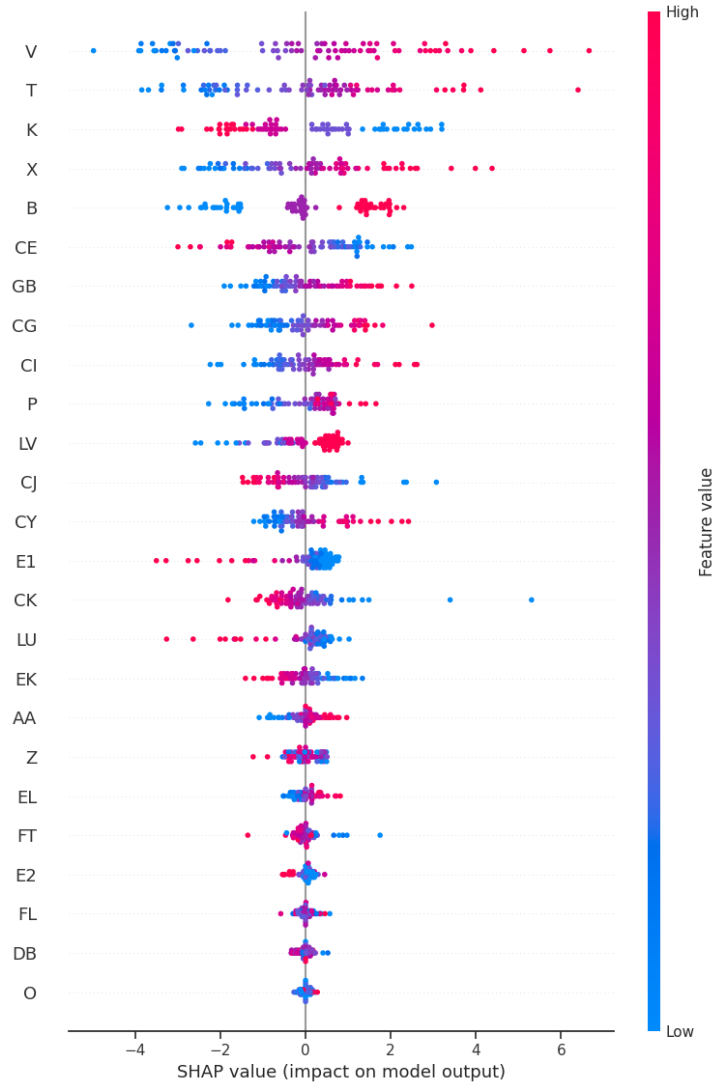

Figure 6. SHAP values for all training data corresponding to the 1st postoperative day. Variables are ranked in descending order, while horizontal location shows whether the effect of that value is associated with a higher or lower prediction. The vertical ordering of the points for each feature column should reflect the corresponding density of accumulation. Color shows whether that variable is high (in red) or low (in blue) for that observation.

Figures 7–9 show the absolute SHAP values averaged over the entire training set corresponding to the 1st, 4th, and 7th postoperative day, respectively. The input for this plot is data from Step 4, Algorithm 1. This bar chart gives us the feature importance, regardless of whether it has a positive or negative effect on the output of the ML model. The numerical values next to each bar indicate the percentage share of the given feature in the entire feature significance pool of the given ML model. The feature subset (V, T, X, K, B, P) is the most important for the patient's postoperative status on the 1st day after surgery. We took a limit of 5% to be in it. On the 4th and 7th day after surgery, these subsets are (V, T, X, K, E1, P), and (T, V, LU, X, P, E1), respectively. Features V (Vital capacity), T (Preoperative FEV1%), X (Forced vital capacity), and P (measure of Chronic Obstructive Pulmonary Disease (COPD) index) appear in all three sets as dominant factors of preoperative lung condition, which is to be expected. What is not intuitively expected is that on the first postoperative day, the features B (age) and K (Body Mass Index) also enter this set, which then lose their significance in the following days. This could be a sign for doctors to pay attention to these two features when preoperatively assessing the patient's future condition immediately after surgery. On the other hand, the appearance of features LU (the number

of functional lung segments removed) and E1 (type of surgical procedure) in the set of significant features on the 7th day after surgery, indicates their importance in the following postoperative period. The number of removed functional segments is a key parameter in classic prognostic systems, see Formulas (6)–(8). Its absence on the 1st and 4th day after surgery makes these classic methods inaccurate predictors of the patients' condition on those days.

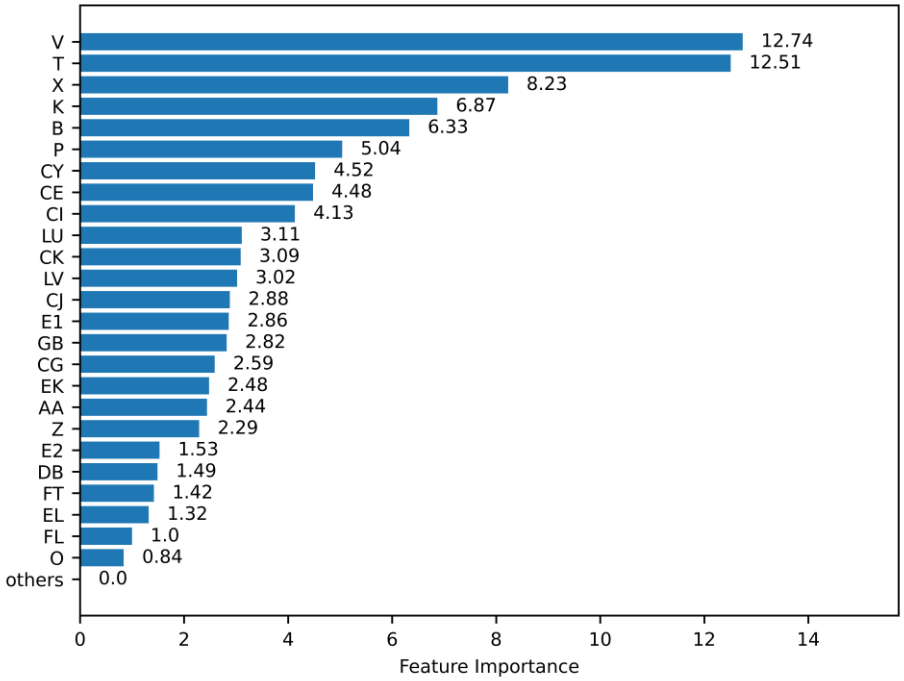

Figure 7. Absolute SHAP values averaged over the entire training set corresponding to the 1st postoperative day.

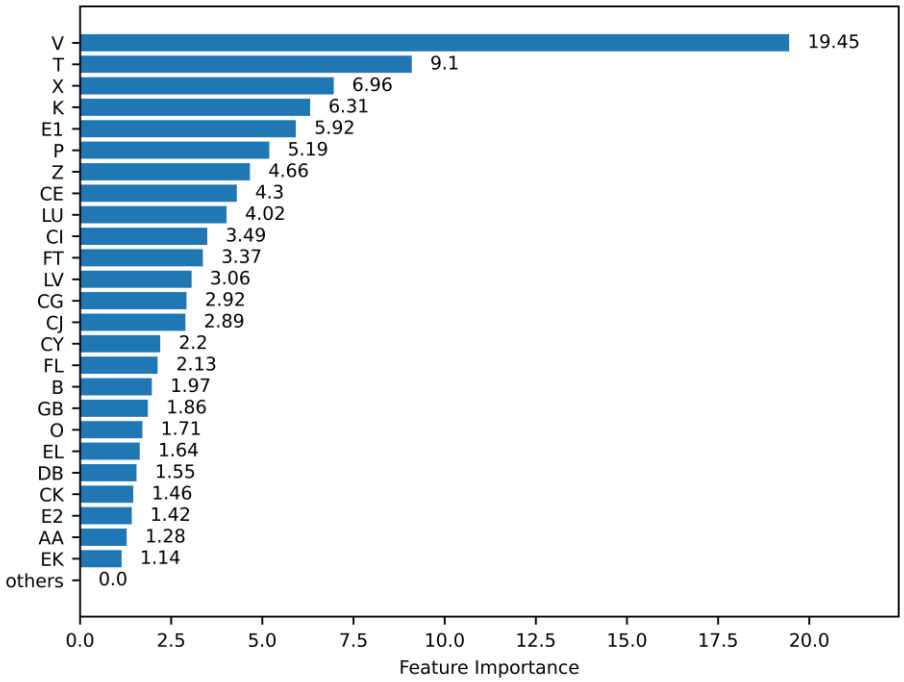

Figure 8. Absolute SHAP values averaged over the entire training set corresponding to the 4th postoperative day.

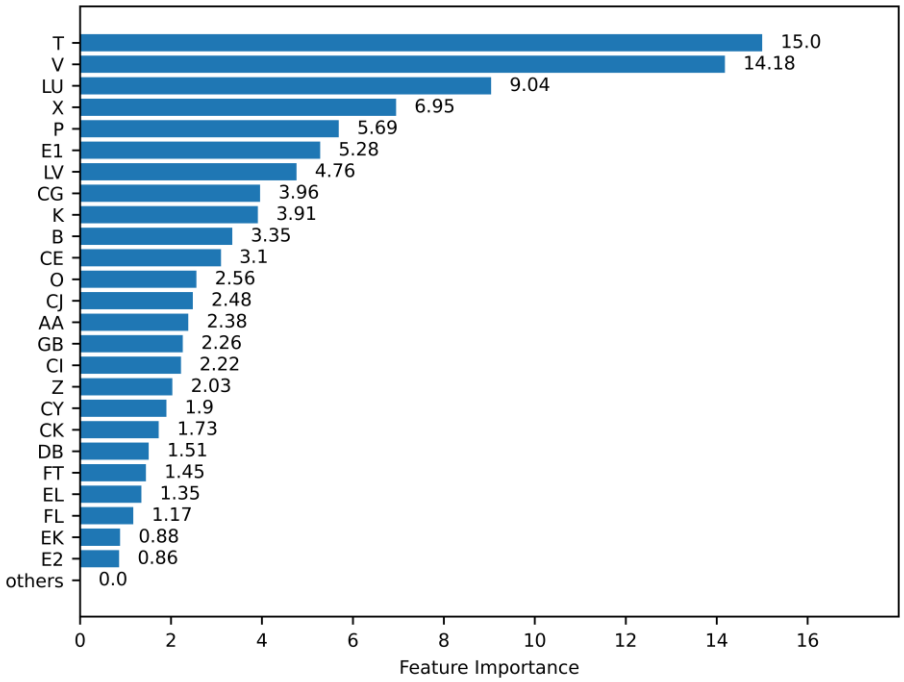

Figure 9. Absolute SHAP values averaged over the entire training set corresponding to the 7th postoperative day.

In order to illustrate the explanatory possibilities of SHAP, it was important to address the question of whether and how much the postoperative state of patients is influenced by the preoperative state of their muscular support for inhalation. The first step was the identification of relevant features. In the set of our measurements, these were features (CY, DB, EK, EL, FL, GB), see Table 1. The additivity of SHAP values allowed us to quantify the collective impact of this set of features by summing their individual absolute values. We previously showed that *Inspiratory respiration muscles* (IRM) has a significant impact in the seven-day postoperative period, with its impact being most significant on the first postoperative day [12]. This fact was not known and verified in modern medical practice.

We can get further global insight into the ML model based on the so-called SHAP dependence plots. They show the marginal effect of one or two variables on the predicted outcome of an ML model [37,40]. The variable that is found to have the greatest interaction with the analyzed variable is automatically included in the diagram. A typical form is given in Figures 9 and 10. Figure 9 shows the SHAP dependence plot for variable X (preoperative forced vital capacity). The most influential variable on X is variable Z (preoperative vital capacity in inspiration). The value of the variable Z is coded according to the color map given on the right side of the plot, with red corresponding to large and blue to small values.

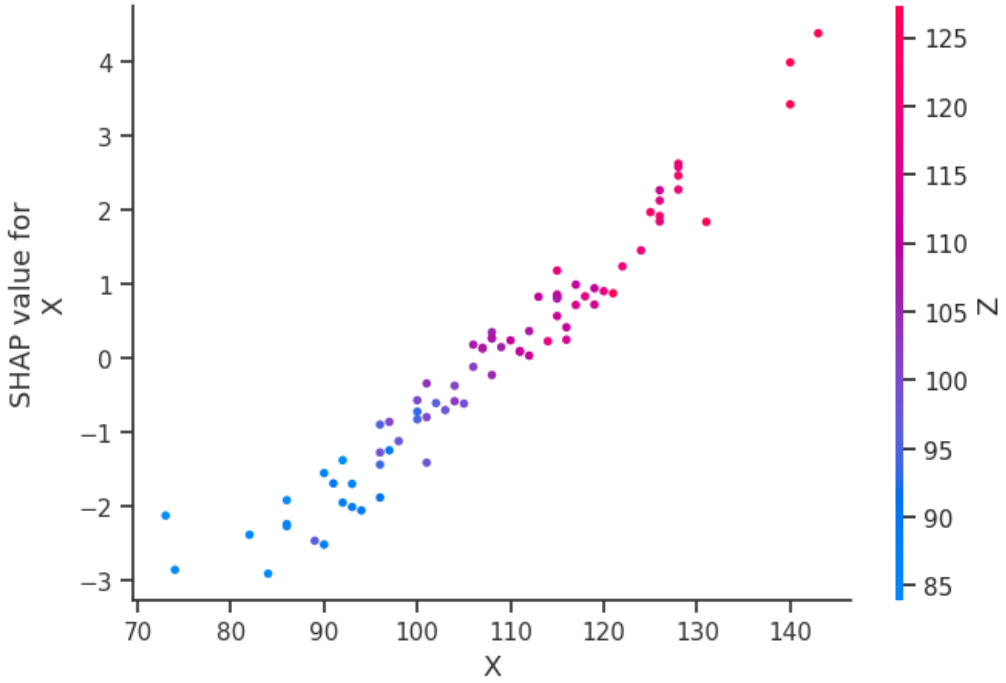

Figure 10. SHAP dependence plot for variable X—preoperative forced vital capacity. The most influential variable on X is variable Z—preoperative vital capacity in inspiration. The size of the variable Z is coded according to the color map given on the right row—red corresponds to large and blue to small values. One can clearly see the almost linear dependence of the size of X and the size of its influence on the output of the ML system.

Figures 10 and 11 show the SHAP dependence plot for variable P (Chronic Obstructive Pulmonary Disease index). The most influential variable on P is variable V (preoperative vital capacity). A rather linear dependence of P and the magnitude of its influence on the output of the ML system can be clearly seen, except for the interval PP $\in$ [1.6, 2] in which we observe the effect of saturation. As for the most influential variable V, its large values correspond to this saturation region, while its small values are associated with small values of P and its negative influence on the postoperative FEV1%. Namely, for $p <$ 1.5 its SHAP values are negative.

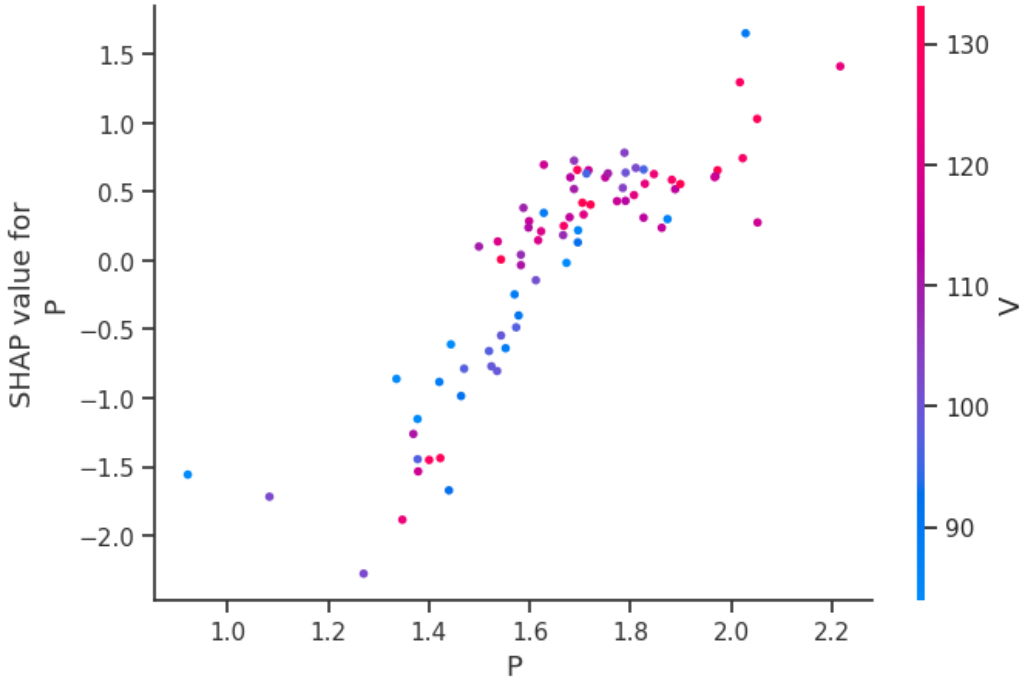

Figure 11. SHAP dependence plot for variable P—Chronic Obstructive Pulmonary Disease index. The most influential variable on P is variable V—preoperative vital capacity. The value of the variable V is coded according to the color map given on the right side of the plot—red corresponds to large and blue to small values. A rather linear dependence of P and the magnitude of its influence on the output of the ML system can be clearly seen, except for the interval PP∈[1.6,2] in which we observe the effect of saturation. As for the most influential variable V, its large values correspond to this region of saturation, while its small values are associated with small values of P and its negative influence on postoperative FEV1%. Namely, for $p < 1.5$ its SHAP values are negative.

4.2. Local Interpretability

The special value of the SHAP method is local interpretability, i.e., the possibility of obtaining the so-called individual SHAP value plot for each individual observation. In Figures 12 and 13, two such waterfall plots are given for two characteristic patients n1 and n2 from our training set, respectively.

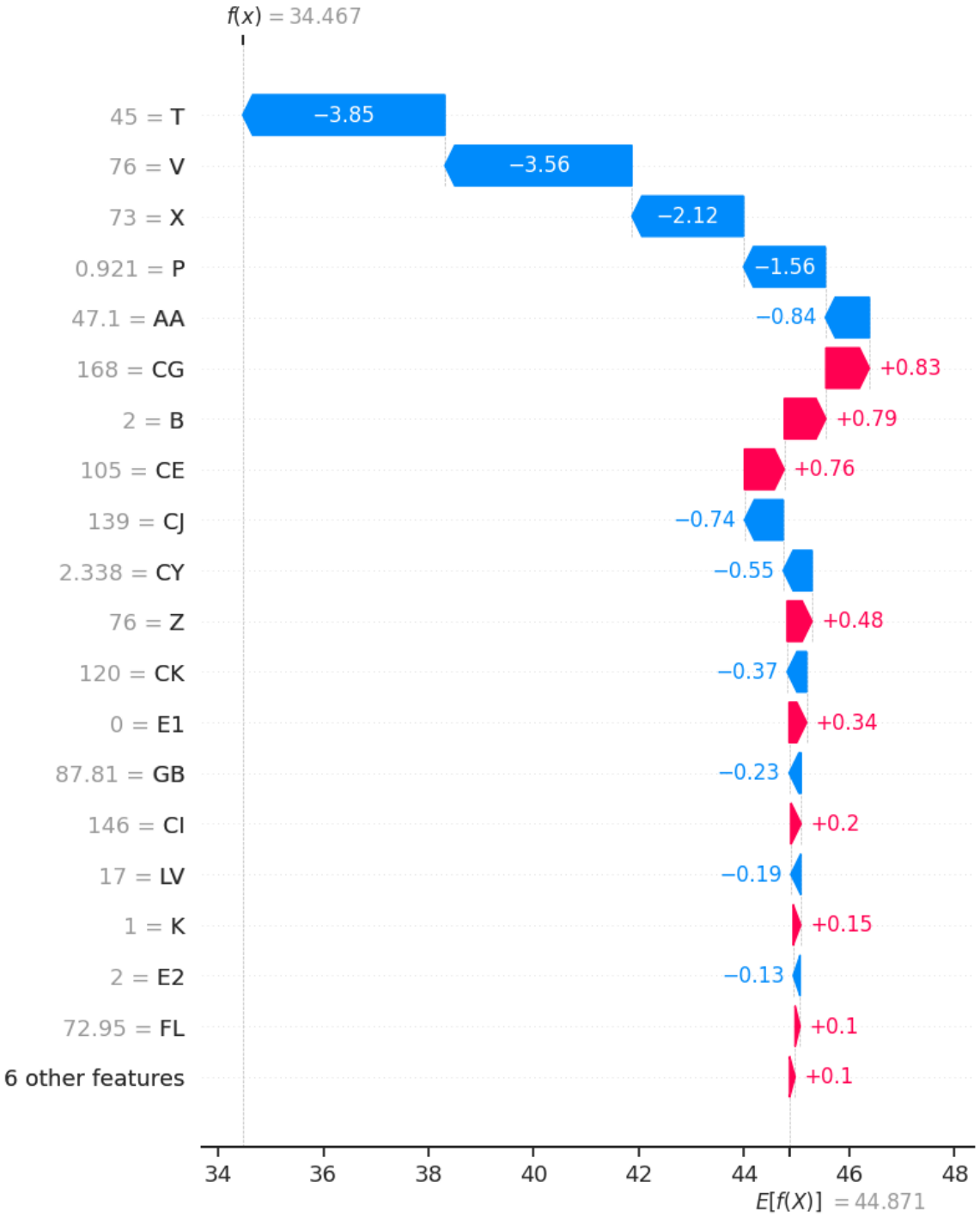

Figure 12. Individual SHAP value plot for patient n1. The base value of FEV1% = 44.871 at the bottom, is the average ML output for all observations. The model prediction for this patient is FEV1% = 34.467, as shown at the top.

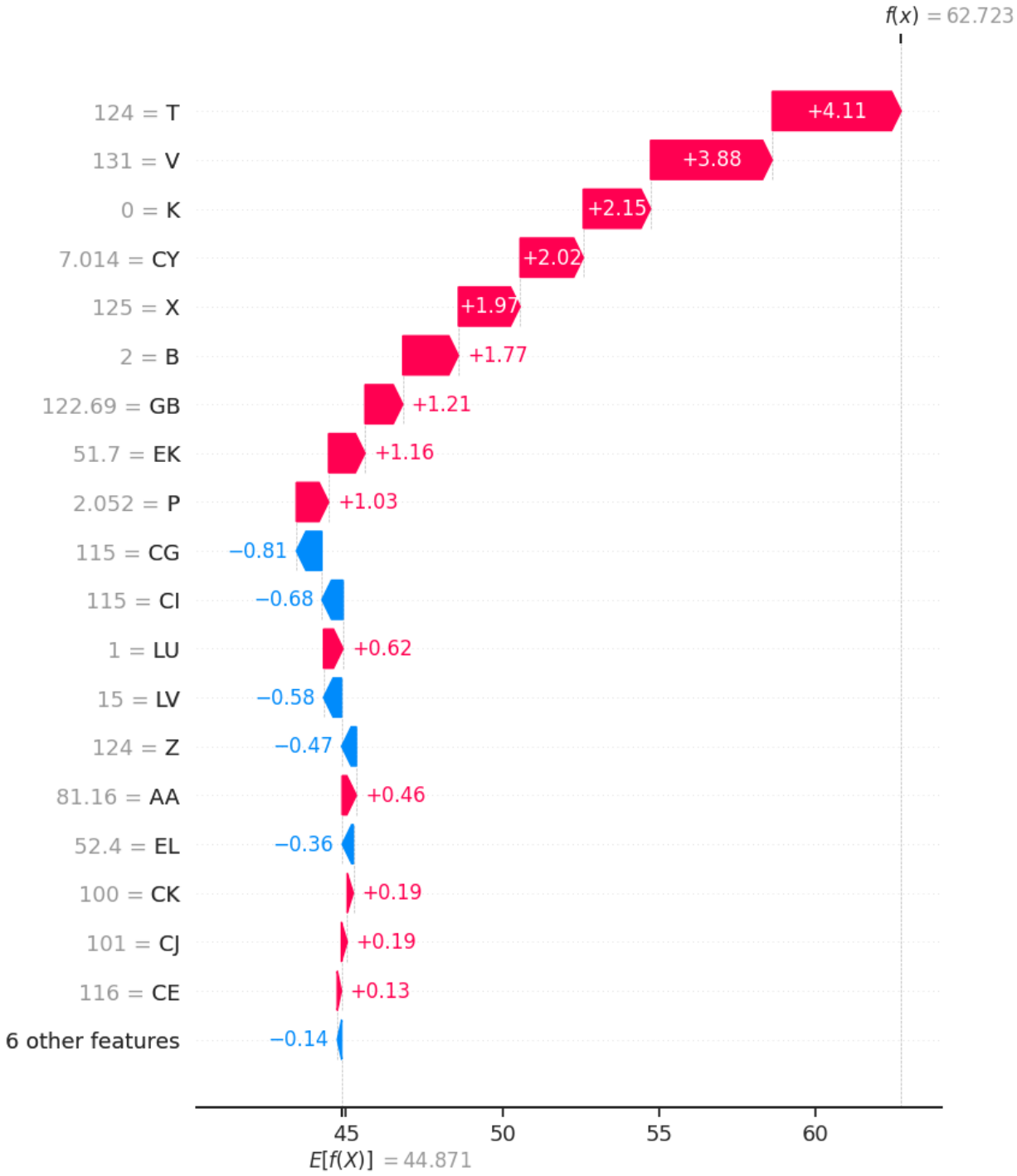

Figure 13. Individual SHAP value plot for patient n2. The base value of FEV1% = 44.871 at the bottom, is the average ML output for all observations. The model prediction for this patient is FEV1% = 62.723, as shown at the top.

The graphic shows the reason for obtaining a specific prediction for a given patient and his preoperative characteristics. To analyze Figure 12, start at the bottom of the waterfall chart and add (red) or subtract (blue) values to arrive at the final prediction. It starts with a base value of 44,871 at the bottom, which is the average ML output for all observations. This value can also be interpreted as the expected prediction of the ML model when there is no single preoperative measurement. The model prediction for this patient is 34,467, as shown at the top. On the left side of the figure, the input values of the preoperative variables for patient n1 are shown. Figure 12 clearly shows that patient n1 will have

a critically low FEV1% = 34,467 on the first day after surgery, primarily because his preoperative values for T, V, X, and P are quite low. For patient n2 from Figure 13, a rather high value of FEV1% = 62,723 is predicted on the first postoperative day. This predicted outcome is the result of high preoperative values for critical features T, V, CY, X, and GB. It is interesting to note that among these features is the feature CY (Mobility of the right hemidiaphragm measured radiographically), which also belongs to the composite feature IRM (inspiratory respiration muscles). As we have already established, IRM is the most responsible for the postoperative outcome on the first day after surgery.

5. Conclusions

The presented experimental results and the interpretation of the identified regression meta-model for predicting the postoperative condition of patients during the first week after lung cancer surgery imply the following facts:

1. It is possible for these purposes to design a multi-layer prognostic regression meta-model with sufficient accuracy even in the conditions of relatively small training sets with input features that are routinely collected in the preoperative period.
2. The accuracy of the proposed model far exceeds the accuracy of traditional prognostic models.
3. Global interpretation of the obtained meta-model using SHAP values showed several interesting new insights important for clinical practice, such as the role of IRM and BMI on the condition of patients on the first critical day after surgery.
4. It was demonstrated how, based on local interpretation of SHAP values, a more accurate picture of postoperative risk factors personalized for each patient is obtained. This interpretation is performed preoperatively, which in our opinion can contribute to a significant improvement of the surgery procedure itself, as well as more successful postoperative rehabilitation of patients.

The introduction of such consultative systems into clinical practice is associated with a number of additional production challenges, such as continuous updating of training sets, adaptation of the system to current changes in the health status of the relevant population, as well as continuous education of doctors in accepting this new technology.

Author Contributions: Conceptualization, R.V. and M.M.; methodology, M.M. and J.R.; software, J.R.; validation, R.V., M.M., and J.R.; formal analysis, R.V. and M.M.; investigation, R.V.; resources, R.V.; data curation, R.V.; writing—original draft preparation, R.V.; writing—review and editing, M.M. and M.P.; visualization, J.R.; supervision, M.M. and M.P.; project administration, M.M. and M.P.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the World Medical Association Declaration of Helsinki and in accordance with the relevant guidelines and regulations. This study protocol was reviewed and approved by the Ethics Committee of the University of Belgrade, Faculty of Medicine, approval number 29/XII-10. All subjects gave written informed consent before participation

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used and/or analyzed in the current study are available through the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: This work was supported by the Vlatacom Institute of High Technologies.

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

References

1. Howington, J.A.; Blum, M.G.; Chang, A.C.; Balekian, A.A.; Murthy, S.C. Treatment of stage I and II non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **2013**, *143*, e278S–313S.
2. Ettinger, S.D.; Akerley, W.; Bauman, J.R.; Bharat, A.; Bruno, D.S.; Chang, J.Y.; Chirieac, L.R.; D'Amico, T.A.; DeCamp, M.; Dilling, T.J. et al. Non–Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **2022**, *20*, 497–530.
3. Steyerberg, E.W. *Clinical Prediction Models*; Springer: New York, NY, USA, 2009.
4. Oswald, N.K.; Halle-Smith, J.; Mehdi, R.; Nightingale, P.; Naidu, B.; Turner, A.M. Predicting Postoperative Lung Function Following Lung Cancer Resection: A Systematic Review and Meta-analysis. *EClinicalMedicine* **2019**, *15*, 7–13.
5. Quanjer, P.H.; Tammeling, G.J.; Cotes, J.E.; Pedersen, O.F.; Peslin, R.; Yernault, J.C. Lung volumes and forced ventilatory flows. Report Working Party. Standardization of lung function tests. European Community for Steel and Coal. Official statement of the European Respiratory Society. *Eur. Respir. J. Suppl.* **1993**, *16*, 5–40.
6. Wyser, C.; Stulz, P.; Soler, M.; Tamm, M.; Muller-Brand, J.; Habicht, J.; Perruchou, A.P.; Bolliger, C.T. Prospective evaluation of an algorithm for the functional assessment of lung resection candidates. *Am. J. Respir. Crit. Care Med.* **1999**, *159*, 1450–1456.
7. Wu, M.T.; Pan, H.B.; Chiang, A.A.; Hsu, H.K.; Chang, H.C.; Peng, N.J.; Lai, P.-H.; Liang, H.-L.; Yang, C.-F. Prediction of postoperative lung function in patients with lung cancer: Comparison of quantitative CT with perfusion scintigraphy. *Am. J. Roentgenol.* **2002**, *178*, 667–672.
8. Cukic, V. Preoperative prediction of lung function in pneumonectomy by spirometry and lung perfusion scintigraphy. *Acta Inf. Med.* **2012**, *20*, 221–225.
9. Varela, G.; Brunelli, A.; Rocco, G.; Marasco, R.; Jimenez, M.F.; Sciarra, V.; Aranda, J.L.; Gatani, T. Predicted versus observed FEV1 in the immediate postoperative period after pulmonary lobectomy. *Eur. J. Cardiothorac. Surg.* **2006**, *30*, 644—648.
10. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; p. 30.
11. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28)*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 2016; Volume 2, pp. 307–318.
12. Vesovic, R.; Milosavljevic, M.; Punt, M.; Radomirovic, J.; Bascarevic, S.; Savic, M.; Milenkovic, V.; Popovic, M. The Role of the Diaphragm in Prediction of Respiratory Function in the Immediate Postoperative Period in Lung Cancer Patients Using a Machine Learning Model. *World J. Surg. Oncol.* **2023**, *21*, 393.
13. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259.
14. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64.
15. Aggarwal, C.C. *Data Classification Algorithms and Applications*, 1st ed.; CRC Press: Boca, Raton, FL, USA, 2015.
16. Stacking. StackingCVRegressor-mlxtend. Available online: https://rasbt.github.io/mlxtend/user_guide/regressor/StackingCVRegressor/ (accessed on 4 July 2023).
17. Brunelli, A.; Charloux, A.; Bolliger, C.; Rocco, G.; Sculier, J.-P.; Varela, G.; Licker, M.; Ferguson, M.K.; Faivre-Finn, C.; Huber, R.M.; et al. ERS-ESTS clinical guidelines on fitness for radical therapy in lung cancer patients (surgery and chemo-radiotherapy). *Eur. Respir. J.* **2009**, *34*, 17–41.
18. Juhl, B.; Frost, N. A comparison between measured and calculated changes in the lung function after operation for pulmonary cancer. *Acta Anaesthesiol. Scand. Suppl.* **1975**, *57*, 39–45.
19. Nakahara, K.; Monden, Y.; Ohno, K.; Miyoshi, S.; Maeda, H.; Kawashima, Y. A method for predicting postoperative lung function and its relation to postoperative complications in patients with lung cancer. *Ann. Thorac Surg.* **1985**, *39*, 260–265.
20. Zhou, Z.-H. Ensemble Learning. In *Encyclopedia of Biometrics*; Li, S.Z., Jain, A.K., Eds.; Springer: Boston, MA, USA, 2015; pp. 411–416.
21. Tibshirani, R. Regression Shrinkage and Selection via the lasso. *J. R. Stat. Society. Ser. B* **1996**, *58*, 267–288.
22. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42.
23. Braiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
24. Guolin, K.; Qi, M.; Thomas, F.; Taifeng, W.; Wei, C.; Weidong, M.; Qiwei, Y.; Tie-Yan, L. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: New York, NY, USA, 2017; Volume 30, pp. 3149–3157.
25. Drucker, B.; Kaufman, L.; Smola, J.; Vapnik, V. Support Vector Regression Machines. *Adv. Neural Inf. Process. Syst.* **1997**, *28*, 779–784.
26. Solomatine, D.; Shrestha, D. AdaBoost.RT: A boosting algorithm for regression problems. *IEEE Int. Conf. Neural Netw. Conf. Proc.* **2004**, *2*, 1163–1168.
27. Pedregosa, F.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
28. Schmidhuber, J. Annotated History of Modern AI and Deep Learning. *arXiv* **2022**, arXiv:2212.11279.
29. Hoerl, E.; Kennard, W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.
30. Tianqi, C.; Carlos, G. XGBoost: A Scalable Tree Boosting System. a*rXiv* **2016**, arXiv:1603.02754.

31. Tolles, J.; Meurer, J. Logistic Regression Relating Patient Characteristics to Outcomes. *JAMA* **2016**, *316*, 533–534.
32. Scikit-learn. Available online: https://scikit-learn.org/stable/ (accessed on 7 August 2023).
33. LightGBM Available online: https://lightgbm.readthedocs.io/en/stable/ (accessed on 5 February 2024).
34. XGBoost. Available online: https://xgboost.readthedocs.io/en/stable/ (accessed on 5 February 2024).
35. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *arXiv* **2022**, arXiv:2207.08815.
36. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90.
37. Scheffer, T. Error Estimation and Model Selection, Ph.D. Thesis, Technischen University at Berlin, School of Computer Science: Berlin, Germany, 1999.
38. Nardini, C. Machine learning in oncology: A review. *Ecancermedicalscience* **2020**, *14*, 1065.
39. Lu, S.C.; Swisher, C.L.; Chung, C.; Jaffray, D.; Sidey-Gibbons, C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Front. Oncol.* **2023,** *13,* 1129380.
40. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18.
41. Shap. Available online: https://shap.readthedocs.io/en/latest/ (accessed on 20 August 2023).