International Journal of Medical Informatics 196 (2025) 105809

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Tackling the small imbalanced horizontal dataset regressions by Stability Selection and SMOGN: a case study of ventilation-free days prediction in the pediatric intensive care unit and the importance of PRISM

Milad Rad [a,*], Alireza Rafiei [b], Jocelyn Grunwell [c,d], Rishikesan Kamaleswaran [c,e,f]

[a] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[b] Department of Computer Science and Informatics, Emory University, Atlanta, GA, USA
[c] Division of Critical Care Medicine, Children's Healthcare of Atlanta, Atlanta, GA, USA
[d] Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA
[e] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA
[f] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Synthetic data
Machine learning
Ventilation-free days
Data imbalance
Critical care

ABSTRACT

Objective: The regression of small imbalanced horizontal datasets is an important problem in bioinformatics due to rare but vital data points impacting model performance. Most clinical studies suffer from imbalance in their distribution which impacts the learning ability of regression or classification models. The imbalance once combined with the small number of samples reduces the prediction performance. An improvement in the trainability of small imbalanced datasets hugely improves the potency of current prediction models that rely on a small set of valuable expensive samples.

Materials and methods: A method called Stability Selection has been used to overcome the high dimensionality problem, which arises when the sample sizes are relatively small compared to the number of features. The method was used to improve the performance of the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN), an imbalance removal algorithm. To test the new pipeline, a small imbalanced cohort of pediatric ICU patients was used to predict the number of Ventilator-Free Days (VFD) a patient may experience for an admission period of 28 days due to respiratory illnesses.

Results: Our model demonstrated its effectiveness by overcoming label imbalance while predicting almost all the non-surviving patients in the test dataset using Stability Selection before applying SMOGN. Our study also highlighted the importance of Pediatrics Risk of Mortality (PRISM) as a powerful VFD predictor if combined with other clinical features.

Conclusion: This paper shows how a hybrid strategy of Stability Selection, SMOGN, and regression can improve the outcome of highly imbalanced datasets and reduce the probability of highly expensive false negative detections in severe acute respiratory disease syndrome cases. The proposed modeling pipeline can reduce the overall VFD regression error but is also expandable to other regressable features. We also showed the importance of PRISM as a strong VFD predictor.

1. Introduction

The challenge of working with datasets that contain only a small number of samples is not a new issue. Since the early days of biomedical research, the process of collecting and processing each sample taken from a cohort has been relatively expensive and time-consuming. This has forced the researchers to limit the sample collection process, which resulted in small sample sizes in many of the biomedical research studies.

In most biomedical research studies, each patient may be assigned a target label if diagnosed with a specific symptom. This label can be used for classification purposes if it represents discrete categories, or it may

* Corresponding author.
E-mail addresses: ghiasirad.milad@gatech.edu (M. Rad), alireza.rafiei@emory.edu (A. Rafiei), jocelyn.grunwell@emory.edu (J. Grunwell), r.kamaleswaran@duke.edu (R. Kamaleswaran).

https://doi.org/10.1016/j.ijmedinf.2025.105809
Received 11 October 2023; Received in revised form 13 December 2024; Accepted 22 January 2025
Available online 25 January 2025

fall along a continuous spectrum, which makes it a regression problem. For example, the target in a small cohort of ICU patients could be defined as either survival or non-survival. Developing a predictive model based on this ICU dataset might not be challenging in itself. The major challenge, however, is that among all the actual samples collected from the test subjects, only a small fraction of the patients may not survive. This results in a heavy imbalance in the label distribution. An example of such a label is Ventilation-Free Days (VFD) [1]. It indicates the number of days a patient has been without mechanical ventilation within a specific period. VFD indirectly represents the survival chance, as its lower value is associated with more severe conditions. If predicted correctly, the identification of patients who are more likely to encounter severe symptoms becomes easier, which helps the medical staff to monitor those patients closely and allocate resources more efficiently.

The datasets used in medical research are often not big enough to train a well-calibrated predictive model effectively. In most cases, ICU studies are conducted with a limited number of samples, or their goal is not to regress a continuous outcome but rather to classify a label.

In general, the main techniques to overcome the problem of imbalance can be summarized in three main categories [2]: 1) Over-sampling, 2) Under-sampling, 3) Hybrid of the two other categories. There are multiple techniques that are categorized in the hybrid methods like random over-sampling and under-sampling, informed under-sampling, and synthetic sampling with data generation [3]. Most of the techniques discussed in the literature center the solution on tackling the variation in sampling technique. For instance, Waikowski [4] tried to use feature selection to help resolve the imbalance in a variety of small sample datasets including protein structure and lymphography images.

Other researchers focused on over-sampling the minority label. Synthetic Minority Over-sampling Technique (SMOTE) [5,6], applied a simple minority replication class instance to introduce synthetic samples. Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) [7] has introduced an advance to this approach to handle the imbalance in regression problems by using random under-sampling with two over-sampling techniques of SmoteR [6] and introduction of Gaussian Noise. The same idea of synthetic over-sampling has also been used in [8] to improve the classification accuracy of fluid overload models utilizing Conditional Generative Adversarial Network (CTGAN) [9]. Another approach, called Deep Imbalanced Regression (DIR) [10], used deep learning to balance the target in large datasets using both label and feature distribution smoothing in deep imbalanced regression (DIR).

A gap in previous research in this field has been the challenge of enhancing regression techniques in the context of small, horizontal, and imbalanced datasets. Most of the methods previously described have either addressed this subject only superficially or have concentrated solely on one of its aspects. Ni et al. in [11] used five-gene signature to improve VFD prediction. Skeletal muscle was also used for this prediction, but they ignored vast number of features with much more data samples [12]. This underscores the need for more comprehensive and targeted approaches to dealing with this complex and multifaceted issue. In this study we aim to introduce a novel workflow to improve the regression accuracy of such datasets. Our case study is also the first to predict VFD using metabolites and gene expressions in pediatric patients.

2. Materials and methods

2.1. Dimensionality reduction

Biomedical data captured in an ICU typically consists of a mixture of demographics, visual recordings, and blood or other fluid sample readings by the medical staff. The dataset may also include metabolites, gene expression levels, and vital signs, along with gender, race, age, and other readable information. By integrating this information, a broad horizontal dataset is formed that contains more features than rows. This structure may lead to overfitting or noise within the modeling process due to the presence of features with low relevance to the target. As the number of features increases, the possibility of regression noise also rises, making the model more complex to analyze and potentially less effective.

To overcome this challenge and train an effective regression model, the first step in data preparation is to reshape the data to decrease the ratio of feature numbers (m) to the number of samples (n) from its initial horizontal shape, Eq. (1), and making it into a more squared format.

$$\frac{m}{n} >> 1 \tag{1}$$

Clinical datasets make possible the use of various feature selection algorithms. These can be categorized into three main classes of filter, wrapping, and embedded techniques [13]. Stability Selection [14], as one of the wrapping techniques in feature selection, is not a new feature selection approach but is very versatile and strong in ranking features by their relevance to the target value. It improves and enhances the existing methods by subsampling the features repeatedly and choosing the structures or variables that occur in a large fraction of the resulting selection sets [14]. If the probability of a subset of features being selected to remain in the set of features is called Π, and a regression model is being trained numerously by random sub-samplings, the set of stable features that occur in a large fraction of the selection set with the acceptable regularization λ and a minimum threshold $0 \leq \pi \leq 1$ parameters can be considered as a stable feature set, Eq. (2). Regularization parameter, λ, can represent any loss in the regression model, and its lower values implies less cost on the chosen features, meaning the lowest cost on the overall model. By choosing the top features after Stability Selection, the reduced dataset will have a more squared shape if $\frac{s}{n} \approx 1$, where s is the number of remaining stable features.

$$S^{stable} = k : max(\Pi) \geqslant \pi \tag{2}$$

2.2. Imbalanced distribution correction

Generally, the target in a small imbalanced dataset can be a left- or right-skew distribution [15]. Oftentimes, the minority population that is in the narrow part of the distribution contains important information, especially when it is associated with higher mortality or dangerous failures. If a model is unable to identify the patients with a higher possibility of developing life-threatening symptoms, its value diminishes.

The imbalance reduction techniques are limited in regression problems. In these problems there are just a few imbalance reduction techniques such as SmoteR, SMOGN, and DIR exist. Due to our focus on very small sample sizes, DIR was not applicable to this problem as it needs larger datasets to operate properly. SMOGN uses SmoteR with the addition of Gaussian noise. SMOGN can produce a more diverse set of samples compared to SmoteR. The performance of the imbalance reduction techniques in regression problems can be compared using continuous measures such as MAE (specifically when dealing with days as the model outcome).

SMOGN divides the dataset into two bins of minority and majority populations and applies oversampling and under-sampling to both. It is built on SmoteR with the addition of Gaussian noise. This makes SMOGN to be able to produce more diverse samples. If any feature selection algorithm is applied after this step, the synthetic samples will not have the correct distance to the minority population. Stability Selection feature set ensures that the KNN used in the SMOGN model will be trained using the features statistically more likely to reduce the regression loss. Stability Selection helps to capture the most important features that give identity to both the minority and majority populations, adding to model's generalizability. If the feature selection is applied after SMOGN, the feature selection would collect the important features based on the generated synthetic samples, resulting in a subset of features that do not reflect the behavior of the original dataset. Once applied correctly, SMOGN on its own is able to provide a smoother distribution of the target value in regression problems, Fig. 1 (a).

2.3. Modeling and Machine learning algorithms

This paper introduces the hybrid of Stability Selection and SMOGN to prepare an imbalanced horizontal small dataset for a regression model. This framework consists of three main steps. The first step uses the Stability Selection to find the most persistent features corresponding with the target value. This step is common in horizontally shaped datasets. In the second step, SMOGN is applied to help with flattening of the target distribution by synthetic minority oversampling along with majority under-sampling, Fig. 1 (b). The presence of SMOGN is critical to the prediction performance. Supplemental Fig. 1 shows the result of not applying SMOGN before the training of an XGBoost regression model. This model fails to correctly predict VFD specifically in the lower end of the spectrum.

3. Evaluation

3.1. Dataset

Two de-identified AirPICU datasets of our previous works are used in this paper [16,17]. The datasets contain 74 children having their metabolic signatures processed for 50 metabolites to study Pediatrics Acute Distress Respiratory Syndrome (PARDS) [18], Supplemental Table 1. In the other dataset, 52 of the children were analyzed for the expression levels of 608 genes from their airway fluid samples, Supplemental Table 2.

After pre-processing, 44 samples containing all demographics, gene expressions, and metabolite readings were filtered. These 44 samples contained all the necessary columns for the feature selection. Stability Selection was applied to the total number of features. Among the total of 663 features of genes, metabolites, and the main demographic features (sex/race/ethnicity/age), 30 features were found to pass threshold limit of 0.99 for Stability Selection, Supplementary Table 3, and the application of SMOGN reduced the samples to 37, by under-sampling the majority population and over-sampling the minority population, which helped smooth the left-skewed distribution of VFD to be more uniformly distributed. It also generated datapoints in the area that VFD was not available. A more uniform distribution reduces the bias in the model training process towards the majority population.

3.2. Case Study

Acute Respiratory Failure (ARF) is a major disease involving the sudden failure of any organ in the respiratory system [19] and affects children of all ages. Critically ill children are rapidly admitted to the Pediatric Intensive Care Unit (PICU), where they can receive advanced ventilation therapy, including invasive mechanical ventilation. Ventilators provide life-supporting care for these patients, providing an opportunity for the lungs to recover and return to spontaneous breathing. However, in a subset of these patients, the disease process may progressively deteriorate, leading to significantly poor outcomes, such as acute respiratory distress syndrome (ARDS) or death. Therefore, acute respiratory failure as a significant health condition requires continuous monitoring and clinical situational awareness.

The number of days a patient spends under ventilator is an indication of how severe the ARDS is. VFD can represent this severity, which is defined as the number of days a critically ill patient has been free from mechanical ventilation in a period of 28 days after admission into PICU. In general, provided that their components are well examined, VFD and other combinations of failure-free days can be a good predictor of survival in severe respiratory disease cases [20] since a lower VFD can signify a more severe health problem or even possibility of mortality.

The importance of VFD prediction is not limited to only this aspect. Due to its relative importance, VFD prediction and its significance have been studied previously, but none of them have mentioned the imbalance and high dimensionality in their dataset. There are also predictors correlating with VFD that were studied in another research. Skeletal muscle was recognized in [12] as a predictive agent in determining VFD, ICU days, and mortality in the elderly, while others like [21] studied the effect of red cell distribution to predict its behavior. There were also some classification models developed to predict the patient's mortality rate using deep learning while having VFD as one of its parameters [22]. Moreover, the impact of different drugs, protein concentrations, and specific genes have been analyzed in other studies [23–27]. Others like [28] analyzed the statistical behavior of VFD. Thus, there may be other helpful predictors in prediction of VFD. Different feature ranking algorithms are able to discover these features if present in the outcome of Stability Selection by finding the correlation between all the remaining features and the target.

3.3. Pipeline Performance

Various models were analyzed to demonstrate the effectiveness of the proposed approach in reducing the mean absolute error (MAE) in the modeling process using Leave One Out Cross Validation before preprocessing and normalizations. We reported the final average results. The initial analysis of comparing the different models demonstrated the power of XGBoost [29], as shown in Table 1, where MAE is mean absolute error, MAPE is mean absolute percentage error, MSE is mean squared error, RMSE is root of mean square error, and Corr. is the correlation between the prediction and actual values. Using grid search, XGBoost was fine-tuned and with the optimized hyperparameters (Table 2). For the rest of this research, XGBoost is used for further sensitivity analysis (https://github.com/ghiasirad/vfd_regression). We chose MAE to be the main assessment metric in VFD prediction with focus on more severe cases. Different problems may require some adjustment to MAE. [30] defined bMAE which is a more balanced MAE for visual imbalanced regressions, but MAE at its simplest form can reflect the clinical importance of the minority population in this paper better. VFD by definition is limited to 28 days and MAE with its interpretability and simplicity provides a more sensible measure and better clinical explainability. Moreover, MAE is less sensitive to outliers. Specially in an imbalanced population, where the minority population with lower values are of interest, choosing a ratio based performance measure may result in incorrect imputations.

The original non-enhanced dataset was used to evaluate the model. The results indicated a reasonable error across the VFD spectrum, with only one test sample having an error of more than 5 days as an acceptable threshold to determine the severity of the situation. While accurately predicting a high VFD population is crucial, understanding VFDs under 20 is particularly valuable in this context since it can predict the possibility of higher mortality. For this part of the spectrum, the model returned median errors of less than 5 days, as shown in Fig. 2 (b).

We should also point out that imbalance reduction in regression problems is still an ongoing area of research, unlike classification problems where there are more set of diverse techniques available to be used. However, we have demonstrated how SMOGN has improved the VFD prediction in our proposed framework in a very small set of samples by comparing the results with and without SMOGN application specifically. This allows us to overcome the extreme distribution imbalance. On the other hand, preventing overfitting in a generalizable sense is an ongoing active area of research. However, we believe that for the constraints defined through this critical care environment, we have demonstrated that the methods and the framework by which they are assembled allows us to medicate the challenges of overfitting when learning on extreme distribution imbalance.

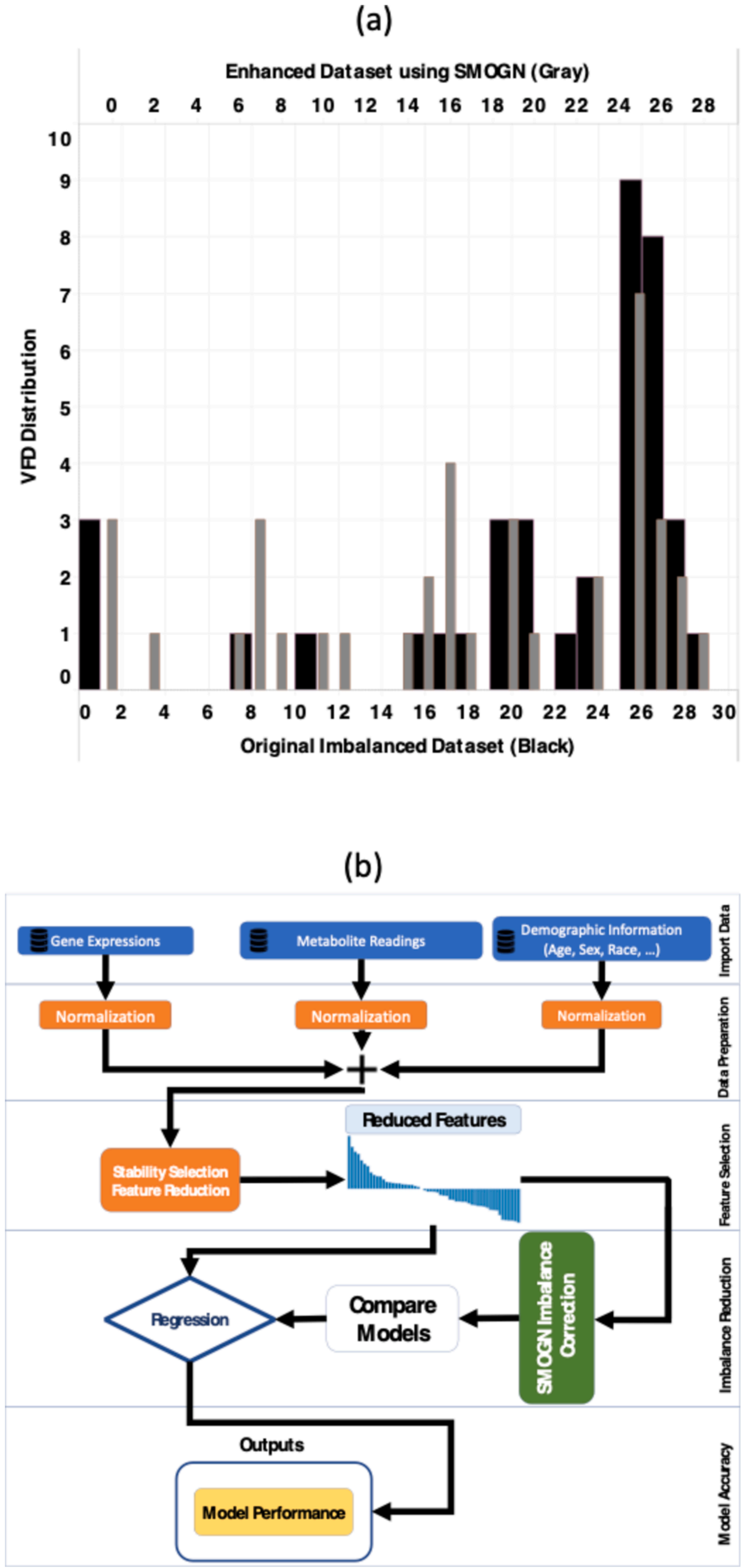

Fig. 1. a) The application of SMOGN on the highly imbalanced dataset of children admitted to ICU with the need to be connected to ventilators. b) The steps taken from feature reduction using Stability Selection to imbalance correction using SMOGN and finally to a regression model.

Table 1
Comparison between different models when trained using SMOGN enhanced dataset.

Regression model	MAE	MAPE	MSE	RMSE	Corr.
XGBoost	2.76	%14.6	21.45	4.63	0.91
Decision Tree	2.83	%12.6	21.07	4.59	0.95
Ridge	3.07	%17.1	18.39	4.29	0.91
Stochastic Gradient Descent	3.44	%17.1	20.43	4.52	0.92
Gradient Boosting Tree	3.78	%22.1	20.68	4.55	0.90
Random forest	4.08	%28.0	30.43	5.52	0.84
Extra trees	4.11	%26.9	34.65	5.89	0.79
Ordinary Least Squares	4.82	%35.2	43.73	6.61	0.87
Lasso	4.93	%35.5	35.59	5.97	0.79
K Nearest Neighbors (K = 5)	7.38	%48.9	77.0	8.77	0.51

Table 2
XGBoost algorithm details and metrics.

Property	Value
Booster	gbtree
Actual number of trees	17
Max trees depth	3
Eta (learning rate)	0.2
Alpha (L1 regularization)	0
Lambda (L2 regularization)	1
Gamma (Min loss reduction to split a leaf)	0
Min sum of instance weight in a child	1
Subsample ratio of the training instance	1
Fraction of columns in each tree	1

Error Property	Value
Min	-3.26
Max	18.40
Mean	3.16
Standard Deviation	3.99

4. Discussion

4.1. Feature importance

Stability Selection passes the top persistent features (30 in this study) to SMOGN. These features, which contain various genes, demographic, and clinical categorical features, are fed into an embedded feature importance ranking algorithm in the XGBoost regression model using their correlation to the target VFD values, Fig. 2(a). The association of PRISM and the IL1 gene family shows how these two features can play a significant role in detecting the mortality rate of pediatric respiratory diseases.

4.2. Stability selection importance in pipeline performance

The Stability Selection helps the data flow to be lightweight and more in shape with the highest relatable features. Stability Selection bootstraps the features and ranks the features more persistent to be in the top features higher. Therefore, by keeping the features that probabilistically are ranked higher, the noise is reduced, especially in the region of higher density. In the prediction of VFD, the majority population (VFD > 20) is more vulnerable to addition of noise if Stability Selection is not used, Fig. 3 (a). Addition of Stability Selection improves the model in this area and reduces the discrepancies, Fig. 3(b).

4.3. PELOD sensitivity analysis

One of the features recorded during admission in pediatric ICU is pediatric logistic organ dysfunction (PELOD) [31–33], which is a representation index for pediatric organ failure. PELOD can still be used in mortality prediction of children with ARF [34], but as PELOD contains non-respiratory organ failure signature [35] and its correlation with mortality, its presence as one of the features possibly would increase noise hence the risk of over-fitting in the model that is focused on just respiratory features. PELOD is thus excluded from the training features;

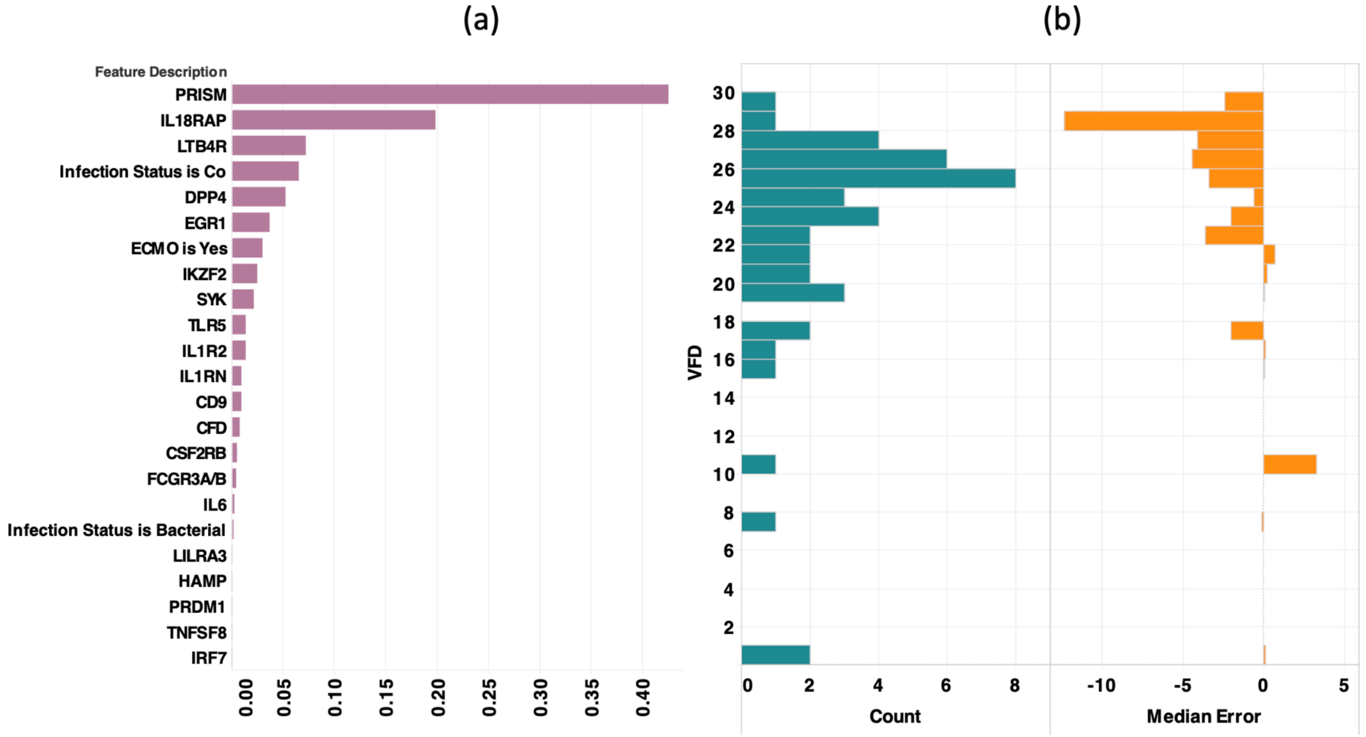

Fig. 2. a) The feature importance demonstrates the role of PRISM and IL1 gene family expression in the VFD prediction. b) The median prediction error distribution shows that lower predicted VFDs are in the range of 5 days of MAE.

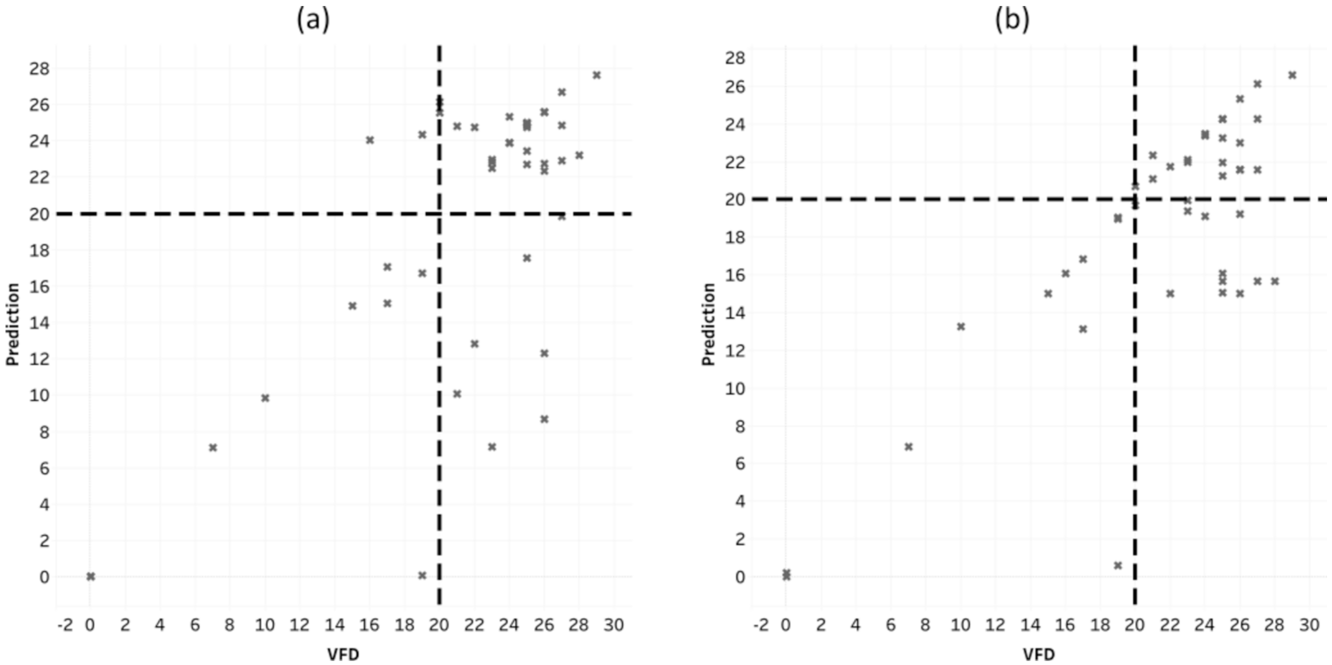

Fig. 3. a) The results of a trained and tuned XGBoost model with imbalance removal using SMOGN without Stability Selection feature reduction. b) The prediction accuracy of the developed model against the original VFD label shows high accuracy for the VFD $<$ 20 with SMOGN tied to Stability Selection.

however, due to its relevance to the mortality rate, a direct outcome of VFD, it is retained for sensitivity analysis, Although PRISM also has non-respiratory signatures, its dependence to clinical measurements convinced us to keep PRISM in the features as a method of clinical feature summarization. We performed sensitivity analysis on PELOD to analyze this parameter against the pipeline's prediction error. Prediction error was then compared against notable features of self-identified gender and race features. The first is designed to reflect the PELOD correlation to VFD, and the second one is aimed at studying the prediction error's sensitivity against those two categorical features.

4.3.1. Importance of PELOD as a Mortality Indicator

PELOD can be an important feature in predicting VFD. In general, organ failure can increase the chance of mortality, and, although it is not an index related to respiratory factors, the initial feature importance with PELOD in the features showed that it is the most important among all the features, which proved the conclusion in [34]. Fig. 4 (a) shows the prediction accuracy against the normalized PELOD values. This shows the correlation between PELOD and the prediction of VFD, although PELOD has been removed from predictive features. The higher PELOD values are associated with lower VFD, which is a combination of composite outcome of higher mortality rate.

For the rest of the study, we removed it from the features, and the pipeline was again trained and tested using the remaining features to test the absence of PELOD on the model. The accuracy measures as shown in Table 2 remained in the same range.

4.3.2. Prediction vs Gender/Race

VFD is not uniformly distributed across gender and race in the test dataset. From the data on gender and race collocated at the time of admission from the 44 admitted patients, the female population, despite having a higher average VFD, experienced a greater MAE across all racial groups following prediction (Fig. 4 (b)). One reason can be the patient's imbalanced distribution regarding this gender, which contained 19 out of 44 subjects in the study. On the other hand, the same challenge exists if the predictions are compared against each documented race. In the dataset, the 20 patients identified as black had lower average VFD, while their prediction error was much higher than the rest of the other categories (Fig. 4 (c)). This shows that the prediction gap observed among different genders also extends into different race categories.

4.4. Impact of PELOD and PRISM Presence on VFD Prediction

PRISM was developed from the physiologic stability index to summarize the multiple variables correlating with the mortality rate of patients in PICU into one indicator [36,37]. This feature is being recorded in the first 24 hours after admission. As such, it is a natural candidate to be a predictive feature, and accordingly, it was picked by the feature ranking algorithm.

Six different combinations of presence/removal of PRISM and PELOD were fed into the pipeline with XGBoost models tuned using Grid Search and a linear fit was performed on the VFD and its predicted values. The values used for PELOD and PRISM were recorded during the first 24 hours of admission and at this time, the VFD prediction was done for the first 28 days of admission. As expected, if PRISM is kept in the features, even without PELOD, the prediction remains in an acceptable range, as shown in Table 3 (model number 1). This set of features shows the best linear fit between VFD and predicted values with the highest R-squared measure for the fit, Fig. 5. The absence of PRISM, on the other hand, increases the different error indicators drastically (models 3, 4, and 5). If just PRISM is kept as the sole predicting feature (model No. 6), the model will fail to predict efficiently, predicting 26 out of the total 44 cases have 21 as their VFD.

4.5. IL1 Gene Family Importance in PARDS

The IL-1 family of cytokines are proteins that regulate the pro-inflammatory immune response to both infections and sterile inflammation [38] and their elevated plasma levels of IL-1ra may serve as a marker of early activation of the inflammatory cascade. For example, plasma IL-1ra levels are higher in children with PARDS than those without PARDS and are associated longer duration of mechanical ventilation, more death in ARF cases with mechanical ventilation, and PICU stay in overall.

Our framework was able to highlight the importance of IL1 gene

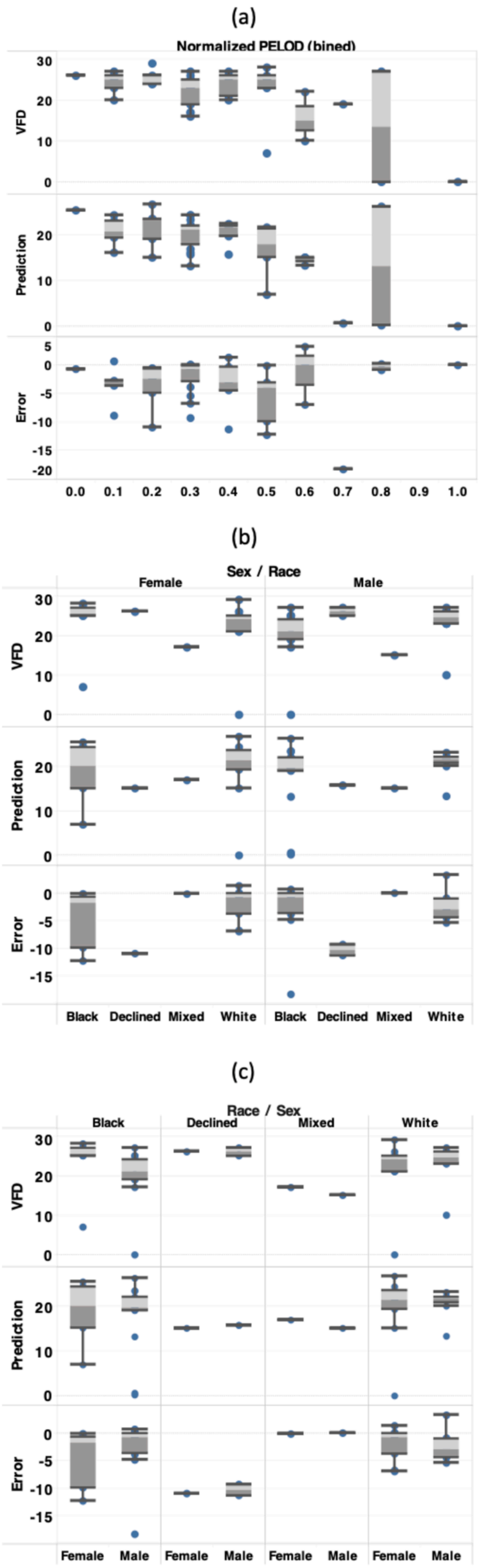

Fig. 4. a) The change in VFD prediction and median error against normalized PELOD values. Error maintains an acceptable range with PELOD change, while prediction values increase with PELOD. b) VFD, predictions, and median error of predictions against different recorded race categories with the same gender. c) VFD, predictions, and median error of predictions against different genders of the same race.

Table 3
The model prediction power for six different combinations of using PELOD and PRISM as predictors.

	Features	MAE	MAPE	MSE	RMSE	Corr.
1	All except PELOD	3.50	%15.7	29.47	5.43	0.77
2	All Features	3.24	%14.2	22.45	4.73	0.83
3	All except PRISM	4.42	%19.1	37.17	6.10	0.74
4	No PELOD/PRISM	6.62	%30.6	118	10.86	0.44
5	Just PELOD	5.71	%24.7	57.75	7.60	0.61
6	Just PRISM	4.98	%22.4	50.18	7.08	0.59

family in prediction of the VFD. We used this finding as another level of verification for the proposed approach. IL1 is a gene family associated with PARDS [39–41], hence its presence can determine whether the patient is going to have a low or high VFD value. It is expected that the patients with PARDS cases experience lower VFD. Therefore, we showed that IL1 as a PARDS indicator can also be a strong VFD predictor.

4.6. Limitations

We acknowledge that heterogeneity arising from diverse causes of acute respiratory failure, including cardiovascular and neuromuscular factors, and the varying VFD trajectories associated with these causes, is a study limitation. We also acknowledge that we were limited to use SMOGN for imbalance reduction as the imbalance reduction techniques are limited for regression problems specially for small datasets.

5. Conclusion

In this paper we have contributed to the problem of high imbalance in regression problems on small clinical datasets with high number of features. This issue can cause expensive errors that may impact the life expectancy of the patients in the ICU. Our approach is novel in its view at these types of problems explicitly to handle clinical features and improving regression accuracy of high-stake target values. We have suggested a novel hybrid pipeline to improve the regression in exceedingly small imbalanced horizontal datasets. The pipeline used Stability Selection for the first feature ranking and SMOGN to balance the dataset and smooth the imbalanced target. To test the proposed pipeline, we used ventilation-free days in a small cohort of imbalanced horizontal ICU patient’s dataset, with high VFD values for most cases and scattered small VFD values for a small group. The training of the XGBoost model proved the effectiveness of combining the Stability Selection and SMOGN to increase the performance in the pre-processing of such datasets, while giving only 3.16 days in average error, and less than 2 days in average error for the minority VFD distribution.

The study also highlighted the importance of PRISM as a powerful VFD predictor if combined with other clinical features and suggested that PELOD should not be considered a good VFD predictor when PRISM is present.

The adoption of this study can be streamlined with standardization of sampling process. The combination of both metabolite levels and gene expressions at the beginning of the ICU admission is the first step in deploying this new reliable approach that can help the medical professionals determine the severity of ICU admissions. The modeling approach can be unique by study but is reliable in small sample sizes.

Summary Table

What was known	· Tackling imbalance in categorical labels has been studied before · Although in the past few years it has been expanded to regression problems, small clinical datasets having extensive number of features have not been addressed.
What this study adds	Provides a new hybrid method that improves the regression performance in highly imbalanced small and horizontal datasets.

(*continued on next page*)

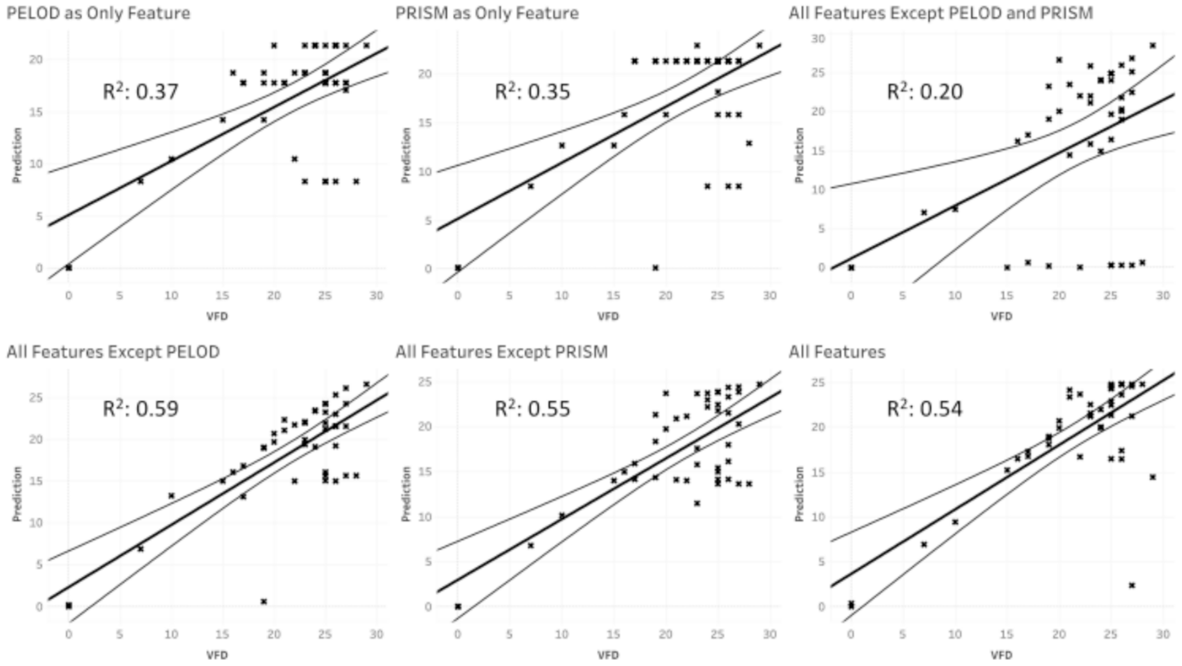

Fig. 5. Comparison of six different feature sets to train the XGBoost model showed keeping PRISM and removing PELOD improved the model's prediction performance with having a lower R-squared score.

(*continued*)

	· Proposes a hybrid of Stability Selection and SMOGN as a novel framework to improve the performance of regression models in such datasets. · Studies the importance of PELOD in VFD prediction. · Performs a combined sensitivity analysis on PRISM and PELOD. · Enables other clinical studies with limited sample numbers and the same characteristics to improve their prediction accuracy.

CRediT authorship contribution statement

Milad Rad: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Alireza Rafiei:** Writing – review & editing, Visualization, Validation, Methodology, Conceptualization. **Jocelyn Grunwell:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rishikesan Kamaleswaran:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

This work was supported by National Heart Lung and Blood Institute [grant number K23HL151897].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We want to thank Alena Alamgir from School of Materials Science and Engineering at Georgia Institute of Technology and Kirsten Cottrill at Emory University for the help in preparing this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2025.105809.

References

[1] L. Contentin, S. Ehrmann, B.G.-A.J. of, undefined 2014, Heterogeneity in the definition of mechanical ventilation duration and ventilator-free days, Atsjournals. OrgL Contentin, S Ehrmann, B GiraudeauAmerican J. Respir. Crit. Care Med. 2014•atsjournals.Org. (n.d.). https://www.atsjournals.org/doi/full/10.1164/rccm.201308-1499LE (accessed October 6, 2023).

[2] G. Haixiang, L. Yijing, J. Shang, … G.M.-E. systems with, undefined 2017, Learning from class-imbalanced data: Review of methods and applications, Elsevier. (n.d.). https://www.sciencedirect.com/science/article/pii/S0957417416307175 (accessed October 7, 2023).

[3] H. He, E.G.-I.T. on knowledge and data, undefined 2009, Learning from imbalanced data, Ieeexplore.Ieee.OrgH He, EA GarciaIEEE Trans. Knowl. Data Eng. 2009•ieeexplore.Ieee.Org. (n.d.). https://ieeexplore.ieee.org/abstract/document/5128907/ (accessed October 7, 2023).

[4] M. Wasikowski, X.C.-I.T. on knowledge and, undefined 2009, Combating the small sample class imbalance problem using feature selection, Ieeexplore.Ieee.OrgM Wasikowski, X ChenIEEE Trans. Knowl. Data Eng. 2009•ieeexplore.Ieee.Org. (n. d.). https://ieeexplore.ieee.org/abstract/document/5276797/ (accessed October 7, 2023).

[5] A. Fernández, S. Garcia, F. Herrera, N.C.-J. of artificial, undefined 2018, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, Jair.OrgA Fernández, S Garcia, F Herrera, NV ChawlaJournal Artif. Intell. Res. 2018•jair.Org. 61 (2018) 863–905. http://www.jair.org/index.php/jair/article/view/11192 (accessed October 7, 2023).

[6] L. Torgo, R. Ribeiro, B. Pfahringer, P.B.-P. conference on, undefined 2013, Smote for regression, SpringerL Torgo, RP Ribeiro, B Pfahringer, P BrancoPortuguese Conf. Artif. Intell. 2013•Springer. 8154 LNAI (2013) 378–389. Doi: 10.1007/978-3-642-40669-0_33.

[7] P. Branco, R.P. Ribeiro, L. Torgo, B. Krawczyk, N. Moniz, SMOGN: a pre-processing approach for imbalanced regression, Proceedings.Mlr.PressP Branco, L Torgo, RP RibeiroFirst Int. Work. Learn. with Imbalanced, 2017•proceedings.Mlr.Press. 74 (2017) 36–50. http://proceedings.mlr.press/v74/branco17a (accessed October 7, 2023).

[8] A. Rafiei, M. Ghiasi Rad, A. Sikora, R. Kamaleswaran, Improving mixed-integer temporal modeling by generating synthetic data using conditional generative adversarial networks: A case study of fluid overload prediction in the intensive care unit, Comput. Biol. Med. 168 (2024) 107749, https://doi.org/10.1016/J.COMPBIOMED.2023.107749.

[9] L. Xu, M. Skoularidou, … A.C.-I.-A. in neural, undefined 2019, Modeling tabular data using conditional gan, Proceedings.Neurips.CcL Xu, M Skoularidou, A Cuesta-Infante, K VeeramachaneniAdvances Neural Inf. Process. Syst. 2019•proceedings. Neurips.Cc. (n.d.). https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html (accessed October 7, 2023).

[10] Y. Yang, K. Zha, Y. Chen, H.W.-… on M. Learning, undefined 2021, Delving into deep imbalanced regression, Proceedings.Mlr.PressY Yang, K Zha, Y Chen, H Wang, D KatabiInternational Conf. Mach. Learn. 2021•proceedings.Mlr.Press. (2021). http://proceedings.mlr.press/v139/yang21m.html (accessed October 7, 2023).

[11] J.-X. Ni, Y.-B. Qian, Y.-W. Zhang, Identification and development of a five-gene signature to improve the prediction of mechanical ventilator-free days for patients with COVID-19, (n.d.). https://bioinfogp.cnb.csic.es/tools/venny/in- (accessed December 14, 2023).

[12] L.L. Moisey, M. Mourtzakis, B.A. Cotton, T. Premji, D.K. Heyland, C.E. Wade, E. Bulger, R.A. Kozar, Skeletal muscle predicts ventilator-free days, ICU-free days, and mortality in elderly ICU patients, Crit. Care. 17 (2013), https://doi.org/10.1186/CC12901.

[13] B. Mwangi, T.S. Tian, J.C. Soares, A review of feature reduction techniques in Neuroimaging, Neuroinformatics. 12 (2014) 229–244, https://doi.org/10.1007/S12021-013-9204-3.

[14] N. Meinshausen, … P.B.-R.S.S.S.B., undefined 2010, Stability selection, Acad. Meinshausen, P BühlmannJournal R. Stat. Soc. Ser. B Stat. 2010•academic.Oup. Com. (2009). https://academic.oup.com/jrsssb/article-abstract/72/4/417/7076513 (accessed October 8, 2023).

[15] H.S.- Biometrika, undefined 1955, On a class of skew distribution functions, JSTORHA SimonBiometrika, 1955•JSTOR. (n.d.). https://www.jstor.org/stable/2333389 (accessed October 8, 2023).

[16] J. Grunwell, M. Rad, … S.S.-C.C., undefined 2021, Machine Learning–Based Discovery of a Gene Expression Signature in Pediatric Acute Respiratory Distress Syndrome, Ncbi.Nlm.Nih.GovJR Grunwell, MG Rad, ST Stephenson, AF Mohammad, C Opolka, AM FitzpatrickCritical Care Explor. 2021•ncbi.Nlm.Nih. Gov. (n.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8208445/ (accessed October 7, 2023).

[17] J. Grunwell, M. Rad, S.S.-S. Reports, undefined 2021, Cluster analysis and profiling of airway fluid metabolites in pediatric acute hypoxemic respiratory failure, Nature.ComJR Grunwell, MG Rad, ST Stephenson, AF Mohammad, C Opolka, AM FitzpatrickScientific Reports, 2021•nature.Com. (n.d.). https://www.nature.com/articles/s41598-021-02354-4 (accessed October 7, 2023).

[18] K.E. Orloff, D.A. Turner, K.J. Rehder, The Current State of Pediatric Acute Respiratory Distress Syndrome, Https://Home.Liebertpub.Com/Ped. 32 (2019) 35–44. Doi: 10.1089/PED.2019.0999.

[19] H. Boren, J. Busey, … R.C.-… R. of R., undefined 1966, Therapy of Acute Respiratory Failure: A Statement by the Committee on Therapy, Atsjournals.Org. (n.d.). https://www.atsjournals.org/doi/pdf/10.1164/arrd.1966.93.3P1.475 (accessed October 9, 2023).

[20] N. Yehya, M.O. Harhay, M.A.Q. Curley, D.A. Schoenfeld, R.W. Reeder, Reappraisal of ventilator-free days in critical care research, Atsjournals.Org, N Yehya, MO Harhay, MAQ Curley, DA Schoenfeld, RW ReederAmerican J. Respir. Crit. Care Med. 2019•atsjournals.Org. 200 (2019) 828–836. Doi: 10.1164/rccm.201810-2050CP.

[21] T.M.N. Otero, D.D. Yeh, E.K. Bajwa, R.J. Azocar, A.L. Tsai, D.M. Belcher, S. A. Quraishi, Elevated red cell distribution width is associated with decreased ventilator-free days in critically ill patients, J. Intensive Care Med. 33 (2018) 241–247, https://doi.org/10.1177/0885066616652612.

[22] Z. Che, S. Purushotham, … R.K.-A. annual symposium, undefined 2016, Interpretable deep models for ICU outcome prediction, Ncbi.Nlm.Nih.GovZ Che, S Purushotham, R Khemani, Y LiuAMIA Annu. Symp. Proceedings, 2016•ncbi.Nlm. Nih.Gov. (n.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333206/ (accessed October 8, 2023).

[23] M. Badr, B. De Oliveira, K. Abdallah, … A.N.-J. of C., undefined 2021, Effects of methylprednisolone on ventilator-free days in mechanically ventilated patients with acute respiratory distress syndrome and COVID-19: a, Mdpi.ComM Badr, B Oliveira, K Abdallah, A Nadeem, Y Varghese, D Munde, S SalamJournal Clin. Med. 2021•mdpi.Com. (n.d.). https://www.mdpi.com/2077-0383/10/4/760 (accessed October 8, 2023).

[24] C.M. Hendrickson, J. Abbott, H. Zhuo, K.D. Liu, C.S. Calfee, M.A. Matthay, Higher mini-BAL total protein concentration in early ARDS predicts faster resolution of lung injury measured by more ventilator-free days, Journals.Physiology.OrgCM Hendrickson, J Abbott, H Zhuo, KD Liu, CS Calfee, MA Matthay, NHLBI ARDS NetworkAmerican J. Physiol. Cell. Mol. 2017•journals.Physiology.Org. 312 (2017) L579–L585. Doi: 10.1152/ajplung.00381.2016.

[25] F. Simonis, A. Neto, J. Binnekade, A.B.- Jama, undefined 2018, Effect of a low vs intermediate tidal volume strategy on ventilator-free days in intensive care unit patients without ARDS: a randomized clinical trial, Jamanetwork.Com. (n.d.). https://jamanetwork.com/journals/jama/article-abstract/2710774 (accessed October 8, 2023).

[26] E.D. Morrell, D. Shane O'mahony, B.J. Glavan, S. Harju-Baker, C. Nguyen, S. Gunderson, A. Abrahamson, F.R. Ii, G. Rona, R.A. Black, M.M. Wurfel, Genetic Variation in MAP3K1 Associates with Ventilator-Free Days in Acute Respiratory Distress Syndrome, Atsjournals.OrgED Morrell, DS O'Mahony, BJ Glavan, S Harju-Baker, C Nguyen, S GundersonAmerican J. Respir. Cell Mol. Biol. 2018•atsjournals. Org. 58 (2018) 117–125. Doi: 10.1165/rcmb.2017-0030OC.

[27] D.D. Yeh, E. Fuentes, S.A. Quraishi, J. Lee, H.M.A. Kaafarani, P. Fagenholz, K. Butler, M. DeMoya, Y. Chang, G. Velmahos, Early protein inadequacy is associated with longer intensive care unit stay and fewer ventilator-free days: a retrospective analysis of patients with prolonged surgical intensive care unit stay, J. Parenter. Enter. Nutr. 42 (2018) 212–218, https://doi.org/10.1002/JPEN.1033.

[28] D. Schoenfeld, G.B.-C. care medicine, undefined 2002, Statistical evaluation of ventilator-free days as an efficacy measure in clinical trials of treatments for acute respiratory distress syndrome, Journals.Lww.Com. (n.d.). https://journals.lww.com/ccmjournal/Fulltext/2002/08000/Strategies_for_blocking_the_systemic_effects_of.16.aspx (accessed October 8, 2023).

[29] T. Chen, T. He, M. Benesty, V.K.-… version 0.4-2, undefined 2015, Xgboost: extreme gradient boosting, Cran.Ms.Unimelb.Edu.AuT Chen, T He, M Benesty, V Khotilovich, Y Tang, H Cho, K Chen, R Mitchell, I Cano, T ZhouR Packag. Version 0.4-2, 2015•cran.Ms.Unimelb.Edu.Au. (2023). https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf (accessed October 8, 2023).

[30] J. Ren, M. Zhang, Yu. Cunjun, Z. Liu, Balanced mse for imbalanced visual regression, in: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7926–7935.

[31] S. Leteurtre, A. Martinot, A. Duhamel, F.P.-T. Lancet, undefined 2003, Validation of the paediatric logistic organ dysfunction (PELOD) score: prospective, observational, multicentre study, Thelancet.Com. (n.d.). https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(03)13908-6/fulltext (accessed October 8, 2023).

[32] S. Leteurtre, A. Duhamel, B. Grandbastien, J.L.-T. Lancet, undefined 2006, Paediatric logistic organ dysfunction (PELOD) score, Thelancet.Com. (n.d.). https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(06)68371-2/fulltext (accessed October 8, 2023).

[33] S. Leteurtre, A. Duhamel, J. Salleron, B. Grandbastien, J. Lacroix, F. Leclerc, PELOD-2: an update of the PEdiatric logistic organ dysfunction score, Crit. Care Med. 41 (2013) 1761–1773, https://doi.org/10.1097/CCM.0B013E31828A2BBD.

[34] F. Leclerc, A. Duhamel, V. Deken, C. Le Reun, J. Lacroix, S. Leteurtre, D. Biarent, R. Cremer, S. Dauger, M. Dobrzynski, G. Emériaud, S. Renolleau, M. Roque-Gineste, D. Stamm, N. Richard, I. Wroblewski, Nonrespiratory Pediatric Logistic Organ Dysfunction-2 score is a good predictor of mortality in children with acute respiratory failure, Pediatr. Crit. Care Med. 15 (2014) 590–593, https://doi.org/10.1097/PCC.0000000000000184.

[35] S. Leteurtre, A. Duhamel, B. Grandbastien, F. Proulx, J. Cotting, R. Gottesman, A. Joffe, B. Wagner, P. Hubert, A. Martinot, J. Lacroix, F. Leclerc, Daily estimation of the severity of multiple organ dysfunction syndrome in critically ill children, C. Can. Med. Assoc. J. 182 (2010) 1181, https://doi.org/10.1503/CMAJ.081715.

[36] M. Pollack, U. Ruttimann, P.G.-C. care medicine, undefined 1988, Pediatric risk of mortality (PRISM) score., Eur. Pollack, UE Ruttimann, PR GetsonCritical Care Med. 1988•europepmc.Org. (n.d.). https://europepmc.org/article/med/3048900 (accessed October 8, 2023).

[37] PRISM III: An updated Pediatric Risk of Mortality score : Critical Care Medicine, (n.d.). https://journals.lww.com/ccmjournal/abstract/1996/05000/prism_iii__an_updated_pediatric_risk_of_mortality.4.aspx (accessed November 6, 2023).

[38] C.A. Dinarello, Interleukin-1 in the pathogenesis and treatment of inflammatory diseases, Blood. 117 (2011) 3720–3732, https://doi.org/10.1182/BLOOD-2010-07-273417.

[39] M.K. Dahmer, M.W. Quasney, A. Sapru, G. Gildengorin, M.A.Q. Curley, M. A. Matthay, H. Flori, Interleukin-1 Receptor Antagonist Is Associated With Pediatric Acute Respiratory Distress Syndrome and Worse Outcomes in Children With Acute Respiratory Failure, Pediatr. Crit. Care Med. 19 (2018) 930–938, https://doi.org/10.1097/PCC.0000000000001680.

[40] L. Ducharme-Crevier, J. Lacroix, Interleukin-1 receptor antagonist and interleukin-1β: Risk marker or risk factor for pediatric acute respiratory distress syndrome? Pediatr. Crit. Care Med. 19 (2018) 993–995, https://doi.org/10.1097/PCC.0000000000001713.

[41] J.J.M. Wong, H.L. Tan, J. Zhou, J.H. Lee, J.Y. Leong, J.G. Yeo, Y.H. Lee, Large scale cytokine profiling uncovers elevated IL12-p70 and IL-17A in severe pediatric acute respiratory distress syndrome, Sci. Rep. 111 (11) (2021) 1–10, https://doi.org/10.1038/s41598-021-93705-8.