Bloody Forecast: Daily Blood Demand Prediction Using Various Modeling Approaches

Martina Dankova
Dept. of Biomedical Engineering
FEEC, Brno University of Technology
Brno, Czechia
238521@vutbr.cz

Stanislava Koskova
Transfusion and Histic Treatment Dept.
University Hospital Brno
Brno, Czechia
koskova.stanislava@fnbrno.cz

Filip Plesinger
Supervisor
Dept. of Medical Signals
Institute of Scientific Instruments of the CAS
Brno, Czechia
fplesinger@isibrno.cz

***Abstract*—Sufficient blood supply is critical not only for scheduled surgeries, but also for emergency medical interventions. In our study, we focus on predicting the daily blood demand separately for two blood types: A+ and O-, based on data from the Transfusion and Tissue Department of University Hospital Brno. The dataset consisted of data on blood demand from 2021 to 2024 and was extended by data regarding non-working days, national and school holidays, seasons, and influenza epidemics. The performance of various prediction models was measured using the normalized Mean Absolute Error (nMAE), which reflects the average prediction error relative to the average daily blood demand. When tested on data from 2023, the best performance was achieved by linear regression models, with a nMAE of 26% for A+ and 50% for O-, indicating lower predictability for blood types with smaller populations. Interestingly, models for different blood types use different features, as the demand for individual blood types depends on different factors. Despite relatively high nMAE values, the models still outperformed a "qualified guess" approach based only on historical averages.**

***Index Terms*—Blood demand, computational modeling, machine learning, feature selection**

I. Introduction

Blood supply management is crucial for ensuring adequate blood supply for transfusion centers. Predicting blood demand is a challenging task, as it is influenced by a variety of factors such as patient needs, seasonal variations, public holidays, and more. Accurate forecasting of blood requirements is essential to ensuring that blood donations are efficiently managed.

This study applies several computational models to predict blood demand in a transfusion center. Among these models, linear regression was employed to analyze the relationship between the blood demand and various factors including holidays, seasons, influenza epidemic and other relevant predictors. The goal is to develop a model that can be directly applied by transfusion centers to predict future blood requirements and ensure optimal blood supply.

Predicting the amount of donated blood is a well-established research problem and has been addressed in several studies using various predictive models. Studies focus on forecasting the total amount of blood donations over time, especially weekly (e.g. [1], [2]). However, the prediction of blood demand by specific blood types remains a less frequently explored task. Notably, one study [3] predicted daily blood demand with great performance, but the prediction was made for larger volumes of blood and it did not separate data by blood type. This gap shows the need for more detailed models that can predict daily blood demand for individual blood types.

II. Method

All steps of the analysis — including data preprocessing, feature extraction, model training, and evaluation — were conducted in Python 3.10 [5]. The implementation relied on libraries including Pandas [4] and Numpy [6] for data handling, Scikit-learn [7] and Statmodels [8] for machine learning models, and Matplotlib [9] for visualization.

A. Dataset

For creation of the models, dataset describing blood demand in University Hospital Brno was used. The data were collected in course of whole 3 years (2021, 2022 and 2023) and first 9 months of 2024. Dataset contained information about daily blood demand for individual blood types.

B. Used features

All features used for training were created based on properties of given days and influenza epidemic status in the Czech Republic and in the South Moravian Region. The derived features are grouped into four thematic categories.

The first category focuses on working and non-working days. The isWeekend feature indicates whether a given day is a weekday or a weekend, while isHoliday marks Czech national holidays. The isFreeDay feature combines these two, distinguishing between working and non-working days. The SchoolHoliday feature captures information about school holidays, including autumn breaks, Christmas, spring breaks, end-of-semester breaks, Easter holidays, and summer vacations. Additionally, the freeInRow feature reflects the number of consecutive non-working days a given day belongs to. The last three features in this category — freedaysCurrentWeek, freedaysPreviousWeek, and freedaysNextWeek — indicate the number of non-working days in the current, previous, and upcoming week, respectively.

The second category is based on influenza data from the National Institute of Public Health [10]. These features capture the number of patients tested for influenza both across the Czech Republic (CR_test_3weeksago) and in the South Moravian Region (JMK_test_3weeksago), as well as the number of positive test results (CR_pos_3weeksago, JMK_pos_3weeksago). The feature JMK_function_3weeksago indicates whether influenza data for the South Moravian Region is available, as hospitals in Brno did not report data every week. Since the models are supposed to be used to predict blood demand two weeks ahead and National Institute of Public Health is publishing the statistics with a slight delay, the values of these features must be from three weeks before the prediction day.

The last two categories include features that provide information about the specific day of the week and the corresponding calendar month.

C. Features selection, model training and testing

After preparing the dataset, forward feature selection process was manually implemented. Model started with an empty feature list, and in each iteration, the feature that improved the model's performance the most was added. This process continued until performance began to deteriorate. Statmodels was used for training the models and calculating the p-values to assess the statistical significance of individual features. The models were trained using data from the full years of 2021 and 2022, as well as the first three quarters of 2024. They were then tested on data from the entire year of 2023. The performance of the models was evaluated using the normalized mean absolute error (nMAE), which compares the mean absolute error with the average daily demand.

D. Comparison with other modeling approaches

In addition to linear regression, more advanced machine learning methods were also implemented. The used forward neural network consisted of a single hidden layer with 8 and 10 neurons for A+ and O-, respectively, determined through manual hyperparameter optimization. Neural network was using the ReLU activation function and the ADAM optimizer. The implemented random forest algorithm had 50 estimators and unlimited maximal depth.

Furthermore, we prepared a Naive model to represent an informed estimate made without the use of any machine learning approach. Naive model simply averages the values from previous years (aligned by the days of the week).

III. Results and Discussion

A. Models' performance

Using the set of prepared features and a basic linear regression approach, blood demand can be estimated with normalized Mean Absolute Error of $26.2 \pm 22\%$ for the A+ blood type and $50.4 \pm 45\%$ for the O- blood type. The relatively high standard deviations indicate that the model's accuracy varies significantly between individual days. The naive model achieved $33.5 \pm 27\%$ nMAE for A+ and $61.5 \pm 60\%$ nMAE for O-, meaning that although the linear regression model has a relatively large error, especially for O-, it still outperforms 'qualified guess.' The comparison between the naive model and linear regression is shown in Fig. 1. This comparison also reveals that the blood type with higher average daily demand is more predictable, despite both datasets being the same size.

The results of all the models tested are displayed in Table I showing, that neither of these methods achieved better performance as linear regression. This could be caused by the limited dataset, with only 1,004 training samples, which may not have been sufficient for these more complex models.

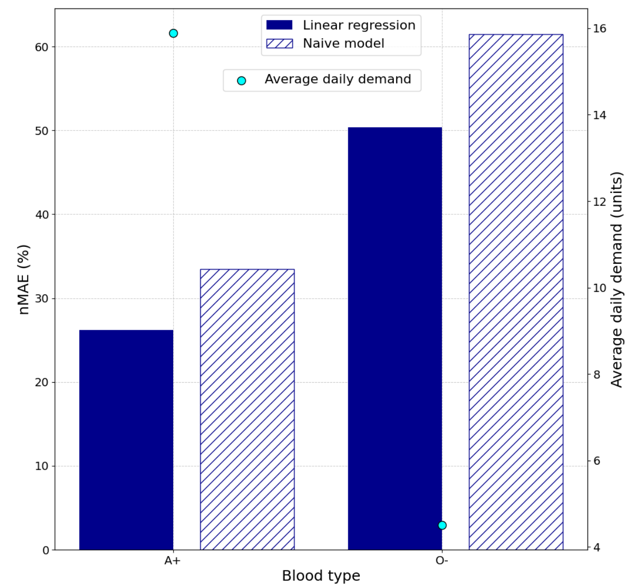

Fig. 1: Comparison of error of the naive model and linear regression model (bar chart). The figure also includes information on the average daily expenditure, represented by the dots.

TABLE I: Comparison of different models: normalized MAE and standard deviation for individual blood types

Model	Linear regression	Neural Network	Random Forest	Naive model
A+	$26 \pm 22\%$	$28 \pm 23\%$	$29 \pm 25\%$	$34 \pm 27\%$
O-	$50 \pm 45\%$	$53 \pm 45\%$	$57 \pm 47\%$	$62 \pm 60\%$

The prediction performance is also visualized over time, as shown in Fig. 2 and as a kernel density estimate (KDE) plot of predicted and real values in Fig. 3 and 4 .

B. Automatically selected features

The linear regression models for different blood types not only differ in performance but also in features they use. The overview of selected features is shown in table II. Both models selected several features from the category focused on non-working days, which corresponds with the fact that planned surgeries are not scheduled on weekends and national holidays. There is a significant decrease in December, probably due to the Christmas period. Higher numbers of positive influenza cases also lower blood demand presumably due to surgeries cancellation. Additionally, A+ demand appears to be lower

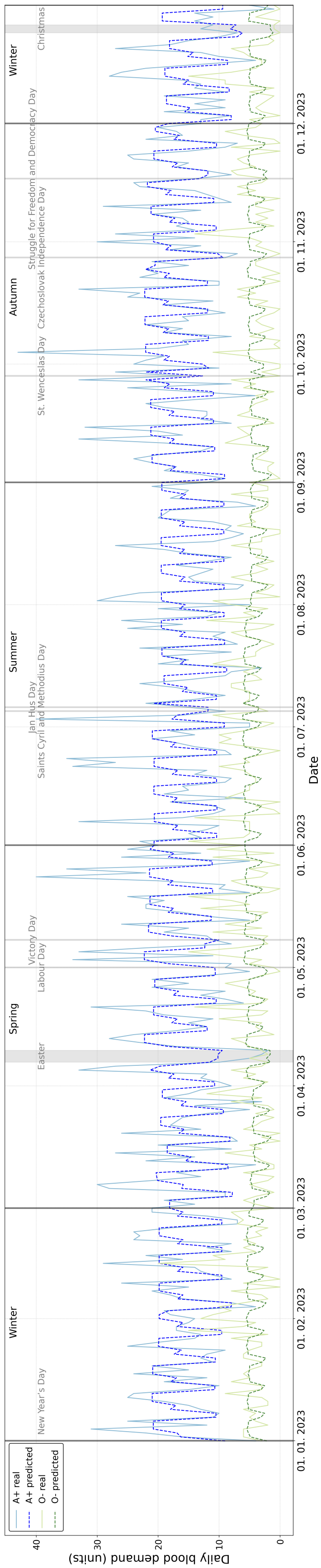

Fig. 2: Predicted and real values for both blood types over time

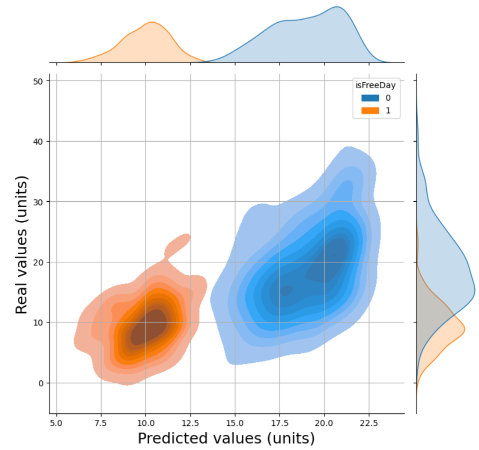

Fig. 3: A KDE plot comparing predicted and real values of blood demand for A+, with the values colored based on the classification of the day as either a working or non-working day.

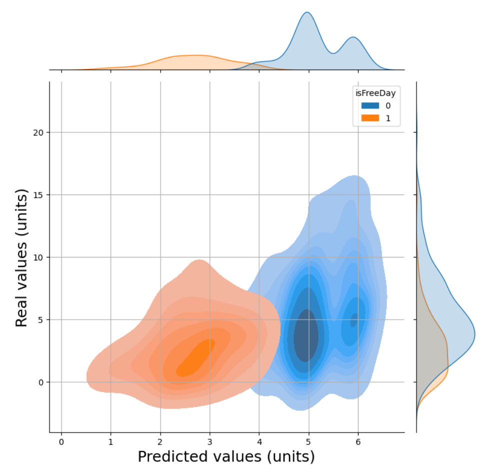

Fig. 4: A KDE plot comparing predicted and real values of blood demand for O-, with the values colored based on the classification of the day as either a working or non-working day.

on Mondays and Tuesdays, likely because patients are not admitted to the hospital over the weekend, and therefore surgeries are not performed on those days. In particular, all significant features for A+ have negative coefficients.

However, for O-, the coefficients for June and July are positive. According to the experience of Tissue and Histic Department, this could be related to a higher rate of car accidents or sports-related injuries during these months. O- is the universal donor type [11] and is commonly given to patients of all blood types in urgent cases with a risk of bleeding out.

C. *Impact of COVID-19 pandemic on training data*

A significant portion of the training dataset was collected during the COVID-19 pandemic. To find whether including 2021 in training is beneficial or not, some models were retrained on a dataset with 2021 data excluded. When tested on the 2023 dataset, the linear regression model for A+ and

TABLE II: Overview of selected features for individual models

Feature category	Feature	A+	O-
	Intercept	21.39	5.02
Days off	isWeekend	-2.2	
	isHoliday	-0.84	
	SchoolHoliday	-1.55	-0.33
	isFreeDay	-8.78	-1.92
	freeinRow	0.36	-0.38
	FreedaysCurrentWeek	0.31	0.07
	FreedaysPreviousWeek	-0.08	
	FreedaysNextWeek		
Influenza	JMK_pos_3weeksago	0.53	
	JMK_test_3weeksago		0.06
	CR_pos_3weeksago	-1.04	
	CR_test_3weeksago	0.44	
	JMK_function_3weeksago		-0.20
Day of the week	Monday	-3.00	
	Tuesday	-3.89	
	Wednesday		0.25
	Thursday		
	Friday		
	Saturday		0.61
	Sunday		
Calendar month	January		
	February		
	March	-1.71	-0.83
	April		
	May	0.64	0.46
	June		0.74
	July	0.02	0.91
	August		
	September		-0.56
	October	1.03	
	November		
	December	-2.17	-0.13

The number in each cell represents the linear regression coefficient for the corresponding feature, while the color of the cell indicates the p-value of the feature. Yellow means p-value under 0.05 representing statistically significant features. Gray means features that are not significant but still are improving model's performance. White-colored features were not selected during forward feature selection. The first row represents the intercept, which is the baseline value of the linear regression model.

O- achieved performances of 26.1% and 50.9%, respectively. In comparison to the results obtained using the full training dataset (26.2% for A+ and 50.4% for A-), it can be concluded that excluding COVID-affected data does not substantially alter the performance of the linear regression model. However, the performance of naive models deteriorated more significantly when the ”COVID years” were excluded. When the model was trained using data from 2022 and 2024, it resulted in a nMAE of 37% for A+ and 67% for O-. If the model relied solely on 2022 data, the performance was even worse: 40% for A+ and 75% for O-. This comparison suggests that, despite the distortion caused by the COVID-19 pandemic in the 2021 data, they still contribute to improving the models' performance.

IV. Conclusion

We developed predictive models to estimate the daily blood demand for two specific blood types, A+ and O-. The models trained using linear regression demonstrate how the demand for each blood type depends on different features, with varying performance for each blood type. Linear regression not only outperformed more sophisticated models, but is also the most practical choice for the given task. Its simple implementation and interpretability of results make it an ideal tool for use in the Transfusion and Tissue Department, where it can effectively serve to predict blood demand in practice.

Acknowledgment

This research was supported by the project RVO:68081731 by the Czech Academy of Sciences.

The data used in this study were provided by the Transfusion and Tissue Treatment Department, University Hospital Brno. We would like to express our sincere gratitude for their support and cooperation.

References

[1] PLESINGER, Filip, Stanislava KOSKOVA, Eniko VARGOVA, martina ADAMCOVA, Jan PAVLUS, Gabriela KOPECKOVA, Radovan SMISEK a Hana LEJDAROVA. Holiday Hemoglobin: How the Vacations Affect Blood Donations across Diverse Urban Sites [online]. In: . 2024, 2024-12-1, - [cit. 2025-03-11]. Available at: doi:10.22489/CinC.2024.075

[2] LI, Xiaofei, Xinyi DING, Helong GUO a Xiao ZHANG. Improved neural network for predicting blood donations based on two emergent factors. Transfusion Clinique et Biologique [online]. 2023, 30(2), 249-255 [cit. 2025-03-14]. ISSN 12467820. Available at: doi:10.1016/j.tracli.2023.01.006

[3] WANG, Yajie, Wei ZHANG, Quan RAO, Yiming MA, Xinyi DING, Xiao ZHANG a Xiaofei LI. Forecasting demands of blood components based on prediction models. Transfusion Clinique et Biologique [online]. 2024, 31(3), 141-148 [cit. 2025-03-14]. ISSN 12467820. Available at: doi:10.1016/j.tracli.2024.04.003

[4] W. McKinney, “Data Structures for Statistical Computing in Python,” Proc. 9th Python Sci. Conf., pp.56–61, 2010.

[5] G. Van Rossum et al., “Python 3 Reference Manual,” Nature, vol. 585, no. 7825, pp. 357–362, 2009.

[6] C. R. Harris et al., “Array programming with NumPy,”Nat. 2020 5857825, vol. 585, no. 7825, pp. 357–362, Sep. 2020.

[7] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., 2011.

[8] S. Seabold and J. Perktold, “Statsmodels: Econometric and Statistical Modeling with Python,” in Proceedings of the 9th Python in Science Conference, 2010.

[9] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007.

[10] STÁTNÍ ZDRAVOTNÍ ÚSTAV. Zprávy o chřipkové aktivitě, hlášení a výsledky laboratorních vyšetření [online]. [cit. 2025-03-14]. Available at: https://szu.gov.cz/temata-zdravi-a-bezpecnosti/a-z-infekce/ch/chripka/zprava-o-chripkove-aktivite-hlaseni-a-vysledky-laboratornich-vysetreni

[11] ČESKÝ ČERVENÝ KŘÍŽ. Jak darovat krev [online]. [cit. 2025-02-13]. Available at: https://web.archive.org/web/20090624015829/http://www.cckpraha1.cz /krev.htm