Vasculitis Outcome Prediction Using Machine Learning and Federated Learning

Kamil Woźniak[1,+,✉], Tadeusz Satława[1,+], Krzysztof Wójcik[2,+], Krystyna Milian[1], Sabina Lichołai[3,4], Tomasz Gubała[1], Grzegorz Biedroń[5], Katarzyna Wawrzycka-Adamczyk[6], Stanisława Bazan-Socha[2], Anna Masiak[7], Michał Chmielewski[7], Barbara Bułło-Piontecka[8], Alicja Dębska-Ślizień[8], Hanna Storoniak[8], Magdalena Krajewska[9], Hanna Augustyniak-Bartosik[9], Radosław Jeleniewicz[10], Maria Majdan[10], Katarzyna Jakuszko[11], Marcin Milchert[12], Marek Brzosko[12], Joanna Kur-Zalewska[13], Witold Tłustochowicz[13], Marta Madej[14], Anna Hawrot-Kawecka[15], Eugeniusz Kucharz[16], Piotr Głuszko[17], Krzysztof Bonek[18], Małgorzata Wisłowska[18], Joanna Miłkowska-Dymanowska[19], Anna Lewandowska-Polak[20], Joanna Makowska[20], Joanna Zalewska[21], Jacek Musiał[2], Jose Sousa[1] and Maciej Malawski[1]

[1]Sano - Centre for Personalised Computational Medicine, Kraków, Poland
[2]Prof A. Szczeklik 2nd Chair of Internal Medicine, Department of Allergy, Autoimmunization and Hypercoagulation, Jagiellonian University Medical College, Kraków, Poland
[3]Division of Molecular Biology and Clinical Genetics, Jagiellonian University Medical College, Kraków, Poland
[4]Academic Computer Centre Cyfronet, AGH University of Science and Technology, Kraków, Poland
[5]Department of Rheumatology and Immunology, Jagiellonian University Medical College
[6]Center for the Development of Therapies for Civilization and Age-Related Diseases, Jagiellonian University Medical College, Kraków, Poland
[7]Department of Internal Medicine, Connective Tissue Diseases and Geriatrics, Medical University of Gdansk, Gdańsk, Poland
[8]Department of Nephrology, Transplantology and Internal Diseases, Medical University of Gdansk, Gdańsk, Poland
[9]Division of Nephrology, Transplantology and Clinical Immunology at 4th Military Clinical Hospital, Department of Non-Procedural Clinical Sciences Faculty of Medicine, Wroclaw University of Science and Technology, Wrocław, Poland
[10]Department of Rheumatology and Connective Tissue Diseases, Medical University of Lublin, Lublin, Poland
[11]Department of Nephrology and Transplantation Medicine, Wroclaw Medical University, Wrocław, Poland
[12]Department of Rheumatology, Internal Medicine, Diabetology, Geriatrics and Clinical Immunology with the Gastroenterology Unit, Pomeranian Medical University in Szczecin, Szczecin, Poland
[13]Military Medical Institute, National Research Institute, Warszawa, Poland
[14]Department of Rheumatology and Internal Medicine, Wroclaw Medical University, Wrocław, Poland
[15]Department of Internal Medicine and Metabolic Diseases, Medical University of Silesia, Katowice, Poland
[16]Department of Internal Medicine, Rheumatology and Clinical Immunology, Medical University of Silesia, Katowice, Poland

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

[17]Pratia MCM Kraków, Kraków, Poland
[18]Department of Rheumatology, National Institute of Geriatrics, Rheumatology and Rehabiitation, Warszawa, Poland
[19]Department of Pneumology, Chair of Internal Medicine, Medical University of Lodz, Łódź, Poland
[20]Department of Rheumatology, Medical University of Lodz, Łódź, Poland
[21]Department of Rheumatology and Connective Tissue Diseases, Ludwik Rydygier Collegium Medicum in Bydgoszcz of the Nicolaus Copernicus University in Torun, Bydgoszcz, Poland
k.wozniak@sanoscience.org
[+]These authors contributed equally to this work

Abstract. Predicting the clinical outcomes in patients with antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) remains a challenge and early identification of patients who are at risk of severe disease course is crucial. To address this, we applied machine learning (ML) and federated learning (FL) techniques to the POLVAS dataset – the largest multicenter clinical database of vasculitis cases in Poland and one of the largest AAV datasets in Europe. Our goal was to predict the key outcomes: increased risk of death and the need for renal replacement therapy (RRT), an independent risk factor of death. We also analysed the significance of individual input features for the predictive capabilities of our models. Furthermore, we compared the performance of centralized model with FL models, which allow for data privacy preservation – a key factor given the highly sensitive medical data involved. We achieved a prediction performance of 0.86 AUC for RRT prediction, and 0.81 AUC for death prediction using a centralized approach and 0.86 weighted mean AUC for RRT prediction and 0.80 for death prediction with the FL approach. The presented results show that FL can effectively predict the risk of RRT and mortality in AAV patients, addressing privacy concerns without compromising the accuracy.

Keywords: AAV, Federated Learning, Machine Learning, Vasculitis

1 Introduction

ANCA-associated vasculitides (AAV), are a group of multisystem, life-threatening rare diseases in which inflammation mainly affects small and medium-sized vessels. There are three main types of AAV, two of which are predominant, namely granulomatosis with polyangiitis (GPA) and microscopic polyangiitis (MPA). The third and the rarest form of AAV is eosinophilic granulomatosis with polyangiitis (EGPA). They are characterized by the presence of antineutrophil cytoplasmic antibodies against neutrophil cytoplasmic enzymes – proteinase 3 (PR3), usually presenting as cANCA pattern when detected by immunofluorescence technique; and myeloperoxidase (MPO), usually presenting as pANCA pattern when detected by immunofluorescence technique. GPA, the most common type of AAV, characterized mainly by granulomatous inflammation of multiple organs, is usually associated with anti- PR3. MPA, on the other hand, is in most cases associated with anti-MPO.

The course of these diseases is currently difficult to predict, posing to a clinician a problem of prediction – at the time of diagnosis – of the future progression of the disease, including its pace. The disease frequently takes an undulating course with

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

consecutive exacerbations and remissions leading eventually to irreversible damage of various organs. Some patients present with only localized lesions involving the skin and/or single organs. In others, a systemic disease with multiorgan involvement, sometimes with a life-threatening course, develops early on. Difficulties in predicting the course of the disease are also due to incomplete understanding of its etiology. There are several hypotheses, some suggesting that genetic factors play an important role, while others assume that AAVs may be related to environmental factors such as viruses or bacterial infections. It has been shown [1] (for example), that *Staphylococcus aureus* nasal colonization is approximately three times more frequent in patients with GPA compared to healthy controls. Optimal as well as personalized treatment schemes for AAV patients are still under development.

The main objective of the presented study was, therefore, to apply a range of data analysis, machine learning (ML) and Federated Learning (FL) techniques to the Polish Vasculitis Consortium (POLVAS) cohort [2], in order to create predictive machine learning models, perform their in-depth analysis of accuracy and generalizability. These models are applied to patient's clinical data available at the time of diagnosis, in order to predict the risk of (a) developing renal impairment requiring RRT (renal replacement therapy - mainly haemodialysis and peritoneal dialysis), and (b) death. The second objective was to analyse which health-related factors increase the likelihood of death.

2 Methods

2.1 Data

Like other rare diseases, vasculitides pose problems for computational data analysis approaches due to the difficulty in involving an adequate number of patients. Fortunately, the Polish Vasculitis Consortium (POLVAS) has assembled a multi-center AAV database, forming the Polish vasculitis registry. POLVAS data is gathered from 12 centers in Poland. Adult patients diagnosed with AAV by participating centers between years 1990 and 2020 were included, with their clinical and laboratory data collected in the POLVAS registry [2,3,4]. We only included individuals who meet the American College of Rheumatology classification criteria for granulomatosis with polyangiitis (GPA), or microscopic polyangiitis (MPA) as well as GPA, MPA, and EGPA nomenclature proposed by the 2012 Chapel Hill Consensus Conference (CHCC 2012) [5]. Specific organ involvement, disease relapse, and disease remission were defined according to the Birmingham Vasculitis Activity Score, version 3. However, as data was collected retrospectively, the exact score for each individual patient was not available. We included patients with AAV regardless of disease severity and excluded patients with a documented diagnosis of AAV who were lost to follow-up at the time of data collection, patients younger than 18 years of age, and pregnant women.

Figure 1 depicts the distribution of data over centers and classes, i.e. renal therapy status and death status. The plots clearly indicate a high imbalance of data distribution across classes and centers. This imbalance is stronger in case of death status, where only 11% of patients have positive status, and only 8/12 centers have positive cases. In the case of renal therapy status 23% of patients have positive status and again only 8/12 centers have positive cases.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

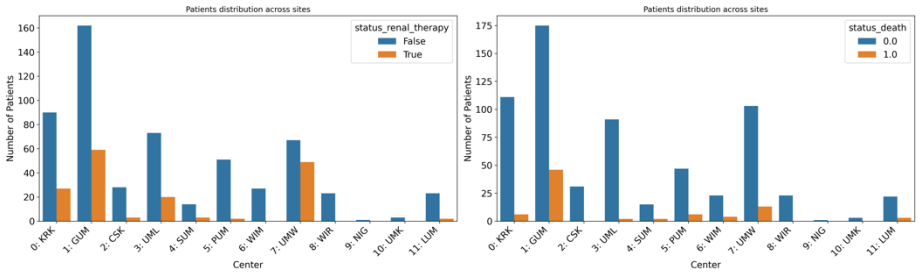

Fig. 1. Patients distribution across centers and classes.

2.2 Ethics Declaration

The data used in the presented work was collected with approval of the Bioethical Committee of the Jagiellonian University, followed by similar approvals issued by appropriate bioethical committees of all remaining POLVAS consortium participating centers. All participants have provided their written informed consent.

2.3 Pre-processing

In such a diverse and maintained by different centers dataset it is inevitable that some values are incorrect or missing. Therefore, ensuring the quality of the dataset used for modelling was an important part of the whole process.

In total, there was data for 747 patients, however multiple checks reduced this number. Two records have been removed due to improbable relation between first symptoms date and date of death (death prior to symptoms). Additional 15 rows were removed due to date of diagnosis before date of first symptoms. Another 2 patients were removed due to missing both diagnosis and first symptoms dates. Finally, one observation which was removed had more than 43 years between onset of symptoms and diagnosis and was treated as an outlier (all other patients had no more than 28 years with vast majority less than 4 years). Finally, 727 observations have been selected for the modelling stage.

Before training the ML model, the feature engineering was applied. Few features were derived from dates of disease progression. However, there were cases when those values (periods between dates) could not be calculated due to the presence of missing values. Although selected method, LightGBM [6], automatically handles missing data, in some cases it was beneficial to impute it. Therefore, missing dates were imputed using the mean difference between diagnosis and first symptoms dates for observations having both dates (444 days). Then, for 13 rows with 1st symptoms but without the diagnosis date, the latter was calculated as 444 days after first symptoms. Accordingly, the date of first symptoms have been set to 444 before diagnosis for 15 patients; cases when both dates were missing have been removed as described earlier. For 10 patients there were missing value of total number of exacerbations, and these were replaced with 0 value. 21 patients had information about the number of exacerbations but missing time between diagnosis to exacerbation – for these cases median value stratified by number of exacerbations was used. In the opposite case (reported time of diagnosis to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

exacerbation with 0 exacerbations; 11 such cases) number of exacerbations was set to 1.

Two numerical columns (CRP level in mg/l and creatinine level in mg/dl) have been first cleaned from outliers. Values less than 0 or greater than 1000 mg/l (for CRP) or 80 mg/dl (for creatinine) were treated as missing. Then those values have been set using IterativeImputer from sklearn library.

The analysis was performed on a retrospective database of 727 POLVAS patients diagnosed with AAV selected in the above manner. The demographic and clinical characteristics of the patients included in the study, organ involvement, antibody pattern as well as some laboratory data was collected and used in this study.

2.4 Developed Models and Input Selection

When designing the training procedure for the predictive models, an important factor from the clinical point of view was to discern information which could be provided to the physician performing the diagnosis. Preferably, the models should be able to communicate increased risk of RRT or death of the patient under examination, early enough for the physician to adjust therapy in order to reverse the unfavourable trajectory of the disease.

A more detailed discussion of the relative importance of specific information for prediction accuracy is provided in the Feature Significance Analysis section. The parameters used to inform predictive models included: specific organ involvement (musculoskeletal, skin, eye, ear/nose/throat, respiratory, heart, gastrointestinal tract, renal, urinary, central nervous system, neurological), detected antibody type (cANCA/anti-PR3 or pANCA/anti-MPO), CRP and creatinine levels (for risk of death prediction only), patient's age at different stages (first symptoms, diagnosis, first exacerbation), and the perceived speed of disease progression (time elapsed between selected stages). Additionally, interactions between pairs of organs (for example, eye and skin, or heart and CNS), and – based on medical expert knowledge – three- and four-organ groups, were included as well. In the RRT risk prediction model, creatinine level was explicitly excluded, since sufficiently elevated creatinine levels already call for renal replacement therapy, and the goal was to detect high-risk cases prior to that episode, when more aggressive drug treatment may still prevent the need for RRT. It is important to note that the POLVAS registry currently provides only the highest creatinine level recorded for each patient (we discuss this limitation in the Discussion section).

In the death risk prediction model application of the RRT procedure was also disregarded as a parameter – in order to base the model's predictive power only on earlier symptoms and tests.

Although data collected by the POLVAS consortium are centralized, being aware of the rarity of this scenario, in addition to the centralized models we also trained the predictive models in federated mode. Federated learning enables the development of prediction models based on data from multiple medical centers while preserving patient privacy by keeping data localized. This approach enhances predictive accuracy through diverse datasets, helps overcome data silos, and mitigates privacy concerns associated with centralized data aggregation [7]. The centralized and federated strategies were optimized independently from each other.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

2.5 Centralized Models

In the model construction procedure, we employed a range of ML tools, including: LightGBM [6], scikit-learn [8], scikit-optimize [9] and pandas [10]. We evaluated a number of different approaches to building the ML models (including logistic regression, support vector machines, random forests), obtaining the best results for the gradient boosting algorithm. We used a gradient boosting algorithm implementation called LightGBM. We performed the hyperparameter tuning procedure using the Bayesian optimization method [11] [12] and applied 10-fold cross-validation to assess generalization of models using Receiver Operating Characteristic (ROC) curve as a measure of the model's performance [13]. To prevent model overfitting [14], we used double (nested) cross-validation. We also examined the calibration of the model to assess the bias of the predicted probabilities.

2.6 Federated Models

To develop our classification models, we adapted the method and implementation presented in [15], which consists of training horizontal federated XGBoost model with a 1-layer convolutional neural network (CNN) representing local learning rates. Instead of using XGBoost, we applied LightGBM model to make the results comparable with the developed earlier centralized scenario. Additionally, we utilized W&B platform for finetuning optimal hyperparameters values and to optimize the training strategy. We trained multiple versions of the federated models and observed performance variability, with the results differing based on the specific center where they were evaluated. This observation highlighted the need for careful evaluation of the results. To assess the generalizability of the developed models, we performed training using Leave One (Center) Out (LOO) cross validation approach, using data from all but one center for training and testing on data from the remaining one. We applied stratified split with proportions: 72% of data were used for training, 13% for validation and 15% for testing, ensuring that at least one positive and negative case is reserved for validation and testing if possible.

We used the same metric, AUC, as in centralized scenario for assessing model performance. The final metrics for the federated models were calculated as a weighted mean, weighted by the test size, collected for centers with available positive samples (otherwise AUC is 0).

Based on the observation of high variability of results, in addition to preparing the globally optimized federated model, we also developed dedicated versions of the models, specifically optimized for each individual center. Besides finetuning hyperparameters (max_depth, max_bin, min_child_samples, num_leaves, learning_rate and CNN learning_rate), in this scenario we additionally allowed a selection of centers from which data were used for training. Figure 2 presents the results of optimizing training parameters for the seventh center, a plot from W&B platform, in this particular case, the model achieving the best performance was trained using data from centers: 1, 3, 4, 5, 11. The method is implemented in PyTorch using Flower framework [16].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

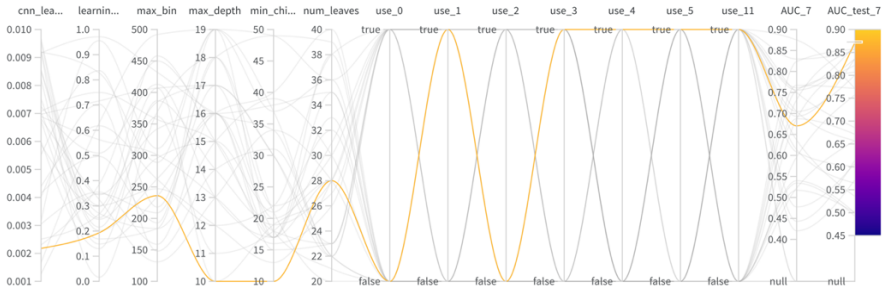

Fig. 2. Finding optimal training parameters for the federated model dedicated to the seventh center.

2.7 Rule-Based Federated Learning

Associative rules mining algorithms extract patterns from the data in the form of $F_l \wedge F_m \rightarrow C_n$ logical rules (F_l, F_m stand for specific ranges of values of the features l, m and C_n stands for class n). These algorithms, although currently somewhat overshadowed by methods based on optimization and gradient descent such as neural networks or gradient boosting, have their advantages such as high interpretability and less data needed to obtain useful results.

Among them is RIPPER [17], a robust algorithm for imbalanced datasets. It iteratively builds rules for the positive (usually minority) class and instances that do not satisfy these rules automatically default to the negative class. Algorithm runs until all positive cases are covered or the threshold number of iterations is reached. The method selects the features and their values as rule conditions that best separate positive cases from negative cases (based on FOIL's information gain [18]):

$$\text{gain} = p[\log\frac{p}{p+n} - \log\frac{P}{P+N}] \tag{1}$$

where p is the number of positive examples covered by the rule, $p + n$ is the number of all examples covered by the rule, P is the number of positive examples in the whole set, $P + N$ is the number of all examples in the set. This is followed by a pruning step, for which a separate pruning set is used from the training data, during which unnecessary conditions are removed from the rules, making the rules more compact and at the same time preventing overfitting by creating too specific rules. Each rule obtained in this way is added to the ruleset, and the cases covered by it are not taken into account when creating further rules.

Most rule-based methods emerged at a time when there was less demand for, and fewer opportunities to train, distributed models. There has been previous work attempting to train RIPPER model in a less localized way. For example, in one approach, the model was trained centrally, with RIPPER models acting as weak learners in an ensemble [19]. In another work [20], the authors presented a similar approach to the one we adopt below. This work does not explicitly position its solution as federated learning. Its authors used Hadoop framework and, as in the federated approach, they constructed local models. Then, the models in the form of rules were sent not to the central server,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

but to other local sites and validated there. The rules were then filtered and merged on this central server. The model demonstrated high prediction accuracy. The disadvantages of this approach, however, are the process of validating local models, the need to coordinate complex communication between local sites, high latency of server updates and increased risk when it comes to maintaining privacy. It also requires high level of trust between sites.

Here, we present a related approach as a proof of concept of the appropriateness of applying similar rule-based approach to small biomedical datasets. We chose RIPPER because it has been shown to achieve high prediction quality, often outperforming other similar methods, and is robust in handling imbalanced datasets since it prioritizes learning rules for the minority class first. RIPPER is also characterized by high degree of transparency, producing concise inference rules that are easy to interpret.

In this work, we trained RIPPER using a federated approach as follows: given data from multiple centers, we simulated local rule mining at each center and then sent these rules to a central server. The set of aggregated rules then created a global model. To classify a new case, each rule is checked to see if the case satisfies its conditions, and the rule applies. The matched rule determines the class to which the case belongs. If no rule covers the case, the default rule that assigns it to the majority class applies. As with the global LGBM model, here we also used the LOO approach, where each center was treated as a test set in a given iteration and the results were averaged. Again, we used Flower framework to simulate a federated environment.

2.8 Feature Significance Analysis

The second objective of our study was to gain better insight, by means of computational analysis, into which health-related factors increase the likelihood of an unfavorable course of vasculitis.

To this end, we used the optimization procedure based on 10-fold cross-validation and reapplied it to the entirety of the available data – the whole cohort of POLVAS patients. This enabled us to harness the full breadth of input data in order to gain the most accurate – given the circumstances – insight into the significance of individual features from the prediction perspective. Again, the hyperparameter tuning and the training of the best model was repeated twice: once for the prediction of the renal replacement therapy risk, and once for the prediction of the risk of death.

The measure of significance of a particular feature was computed as sum of information gained by the split on that feature for all trees in the model ensemble. It measures the amount of information that is gained (meaning how much entropy is reduced) after performing a split on feature F [21]:

$$\text{information gain}(T, F) = \text{entropy}(T) - \text{entropy}(T, F) \tag{2}$$

$$\text{entropy}(T, F) = \sum_{i}^{F_{splis}} \frac{|T_i|}{|T|} \text{entropy}(T_i) \tag{3}$$

$$\text{entropy}(T) = -p \log_2 p - (1-p) \log_2 (1-p) \tag{4}$$

where $|T_i|$ is the number of samples in subset T_i, $|T_i|$ is the number of all samples and p is probability of a positive outcome in set T.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

3 Results

3.1 Centralized Results

The model training and evaluation procedure (described in the Methods section) was performed twice: once for the RRT risk prediction model, and once for the risk of death prediction model. In both cases we obtained results which we consider promising. The outcome is a set of twenty models, ten per each clinical question – Table 1 presents the mean value and standard deviation for the ROC AUC measure across all ten trained models, with the best hyperparameters selected for each. Two ROC AUC charts in Figure 3 present the performance of all ten models in predicting each clinical event.

Table 1. Mean Area Under the ROC Curve (ROC AUC) measure, and standard deviations computed across all ten models (for each of the two clinical questions).

Model	Mean ROC AUC for 10 folds	Std deviation for ROC AUC
RRT prediction	0.86	0.04
Death prediction	0.81	0.08

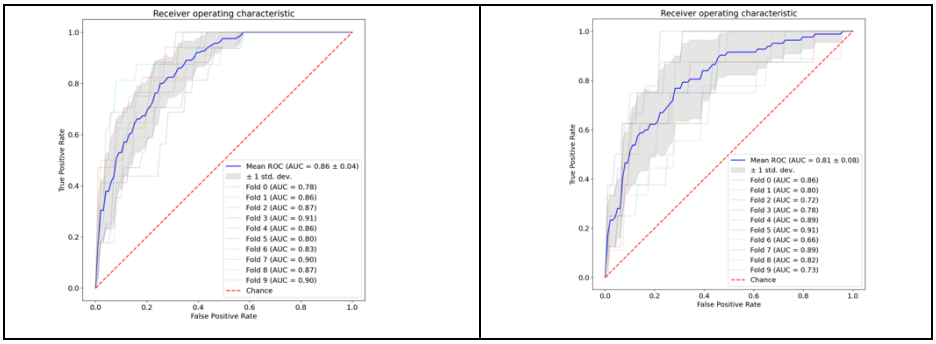

Fig. 3. ROC plot for all 10 folds (models) developed per each event prediction (need for RRT in the diagram on the left; death in the diagram on the right).

As described in the Methods section, we retrained final models, using the same cross-validation scheme, on the full dataset, in order to perform feature significance analysis. The 20 most significant features per model are presented in Figure 4.

When comparing the relevance of symptoms computed by our models (when trained on an optimal set of parameters) with clinical knowledge, it seems that inflammatory (CRP) exponents are important. For both models, this information proved to be relevant to the course of the disease. Furthermore, features related to age and progress of the disease are also important in both models. For the model which predicts the risk of RRT, features related to renal inflammation are significant, whereas the death risk prediction model takes into account involvement of other organs as well.

Information about the medical care unit (center) where the given patient had been treated was also significant, which is caused by slight specialization among the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

POLVAS centers regarding the cases which are treated at any particular center (for instance, some complex cases are transferred to a small subset of centers which have better experience in dealing with specific disease progression scenarios).

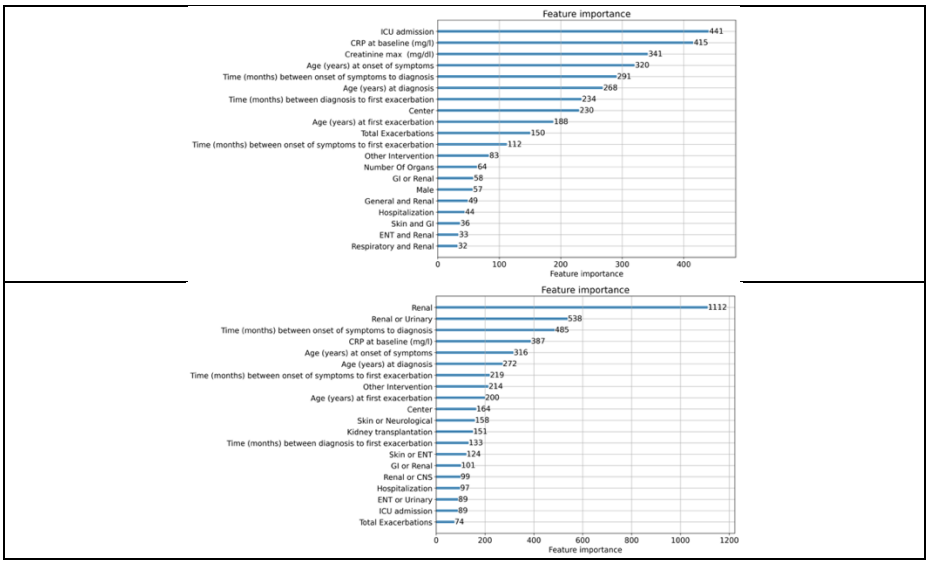

Fig. 4. Twenty features (columns in the input clinical data set used for training the models) evaluated as most significant for prediction of the risk of the clinical event in question – need for RRT in the bottom diagram and risk of death in the top diagram. Significance is computed using the sum of gained information formula (presented in Feature Significance Analysis section).

3.2 Federated Results

We optimized the global federated model and dedicated federated models for individual centers, predicting death and RRT. The obtained results are presented below.

Global Federated Models. Table 2 presents the performance achieved by the global federated models, evaluated using a LOO cross-validation approach. The mean performance metrics are derived from eight experiments, each corresponding to centers containing both positive and negative samples. The remaining 4 centers, which only contain negative cases, were excluded from the computation of the mean. The reported mean values are weighted according to the size of the test set for each individual center.

Table 2. Weighted mean test AUC and standard deviation test AUC from LOO experiments.

Model	Experiments count	Weighted mean test AUC from LOO experiments	Weighted std test AUC from LOO experiments
RRT prediction	8 x 20	0.76	0.11
Death prediction	8 x 20	0.61	0.08

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

The model developed to predict the need for renal therapy obtained mean weighted AUC equal to 0.76. The model predicting death achieved 0.61 mean weighted AUC. Notably, the performance of these federated models, indicating generalizability of the model to data from external centers is substantially lower compared to the centralized versions.

Dedicated Federated Models. The observation of high variability in the performance of federated models, when evaluated on local test sets, motivated us to further investigate by training additional dedicated federated models, optimized for individual centers. Table 3 presents results of these experiments, comparing three training strategies. The first and second strategy uses entire available training data, first optimizes local validation AUC, second global validation AUC, while the third strategy uses data only from specific centers, selected during optimization processes and optimizes local validation AUC. The model predicting the need for renal therapy achieved weighted mean test AUC of 0.83 when trained with entire training data and improved to 0.86 when trained with data from the selected centers. Similarly, the model predicting mortality achieved weighted mean test AUC of 0.77/0.78 when trained on the entire dataset and increased to 0.80 when trained using the center-specific data subset.

Table 3. Weighted mean test AUC and standard deviation test AUC of dedicated federated model.

Model	Experiments count	Training strategy	Weighted mean test AUC of dedicated federated models	Weighted std test AUC of dedicated models
RRT pred.	8 x 20	Using full dataset, optimizing local validation AUC	0.83	0.09
	1 x 20	Using full dataset, optimizing global validation AUC	0.83	0.11
	8 x 30	Using data only from selected centers, optimizing local validation AUC	0.86	0.12
Death pred.	8 x 20	Using full dataset, optimizing local validation AUC	0.77	0.20
	1 x 20	Using full dataset, optimizing global validation AUC	0.78	0.17
	8 x 30	Using data only from selected centers, optimizing local validation AUC	0.80	0.18

Dedicated Rule-Based Classifier. The results obtained with the rule-based federated RIPPER using LOO cross-validation approach are shown in Table 4. The resulting average ROC AUC values from the 8 experiments are 0.71 for the need of renal therapy prediction and 0.64 for death prediction. These results are comparable to those obtained using the LGBM model in a similar manner.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

Table 4. Results obtained using federated RIPPER.

RIPPER Model	Experiments count	Training strategy	Weighted mean test AUC of dedicated federated models
RRT prediction	8	LOO-CV	0.71
Death prediction	8	LOO-CV	0.64

4 Discussion

To the best of our knowledge, this is the first analysis which applies machine learning and especially federated learning techniques to answer clinical questions using data from a multi-center vasculitis registry of this size and quality. POLVAS is currently the largest AAV registry with this much information collected for each individual in Central Europe. We posed a question as to whether, based on the available data, we could predict the course of the disease and, in particular, the risk of RRT due to end-stage renal failure, and death, in centralized and federated approach. We also identified a group of parameters that may be relevant for this outcomes analysis.

A machine learning approach in the scope of disease course prediction has been presented by Lezcano-Valverde et al. [22] among others. In their work, they presented a mortality prediction model for rheumatoid arthritis using data from two independent cohorts. Of the variables analysed, age at diagnosis, erythrocyte sedimentation rate, and number of hospital admissions exhibited the highest predictive value. This model (by the authors' own admission) appears to be overestimated but could be validated and improved with an external data set [22]. The same disease was referred to by Lin et al. [4] In their work they present a model which aims to predict the severity of the disease course using clinical data. Using a machine learning approach, they obtained AUC=0.758 in the context of exacerbation of the disease [23]. A different approach was presented by Tsujitani et al. [24] In their study, they presented a bootstrapping-based neural network model to estimate the survival function. This approach is notably different from the standard Cox proportional hazards model used in this situation, and the model was tested using data from patients with primary biliary cirrhosis, with longitudinal data available [24]. When comparing our approach with the work presented above, we note, first of all, that the model is based on both clinical data and laboratory studies. We apply the machine learning approach and federated learning approach, and moreover, the disease addressed by our research has not yet been the subject of existing models. The results of experiments obtained with federated learning approach demonstrate that while generalizability of the global federated model might not be satisfactory to be implemented in medical practice, the performance of optimized models for individual centers may be sufficient. The improvement in the weighted mean AUC for both renal therapy prediction and mortality prediction suggests that tailoring the model to the specific characteristics of the data from each center can lead to more accurate predictions. This may reflect underlying heterogeneity in patient populations, clinical practices and data quality across different centers, especially in scenarios with very limited datasets. Exploratory analysis of the distribution of the most important features for both problems revealed significant variation across centers. A generalized federated model

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

may not fully capture the nuances of individual centers. Further research, with more thorough formal statistical uncertainty measurements, is needed to investigate the generalizability of this approach across more diverse datasets and its implications for other clinical prediction tasks.

In addition, given the specifics of the problem, i.e. the size and imbalance of the data, and the domain in which the explainability of the model's decisions is extremely desirable, we conducted the experiment using the RIPPER method in a federated way. The results obtained can be considered satisfactory in comparison to those achieved by LGBM, as shown in Table 4. RIPPER performed worse at predicting RRT, but slightly better at predicting death. We can observe the same pattern as in the case of LGM: the classification problem of renal therapy prediction proved to be easier than death prediction but in case of RIPPER, this is less evident. Although the values of the metric evaluating the predictive quality of the model obtained in this way are lower than those of the best local LGBM models, this is acceptable given the minimal complexity – and even the naivety – of the aggregation method used. Therefore, we can nevertheless positively assess the applicability of such a solution to this type of problem and medical data, where explainability and privacy preservation are top priorities and data are scarce. RIPPER results could probably have been better if a more in-depth analysis of the various parameter values had been carried out. We have not done any hyperparameter tunning which should increase the quality of predictions. Another area for improvement is the way the rules are aggregated. Nonetheless, presented approach shows promise and, since it has not yet been comprehensively explored, we leave the development of federated interpretable rule-based methods for future work.

The most important limitation of the presented study is the lack of available longitudinal data regarding when an exacerbation or adverse incident may have occurred, which would help the model take into consideration the temporal relations between laboratory tests, symptoms and exacerbation events. In the cases where the clinical parameter value refers to the worst value measured over the entire course of the disease, we abandon using such values for prediction, especially when such “worst-case” values are obtained for patients immediately prior to the predicted event. For instance, with regard to blood creatinine level, we had to remove the information as a prognostic marker for renal involvement, since the values recorded in the database for patients undergoing RRT are usually measured directly before ordering the RRT procedure. Therefore, using the worst recorded level of creatinine would reduce the trained models’ predictive performance for early diagnosis scenarios. If time-series data could be collected for key laboratory test results, or indeed even more laboratory parameters were recorded in the input data (for instance: troponins, acute-phase proteins or complement factors), the models could be improved further. We are therefore considering – on the basis of results obtained thus far – a prospective study of newly recruited AAV patients in order to test whether even more precise models could be developed using a similar ML approach when finer-grained time-series characteristics are added to input data and to test against other methods, such as deep learning models.

We consider that the application of knowledge provided by the trained models, especially after further validation, definition of actionable decision thresholds and evaluation on external datasets, could be incorporated into vasculitis treatment strategies through early identification of patients with an elevated risk of exacerbation. This, in turn, might result e.g. in enrolling such individuals in a more intensive monitoring

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

scheme of blood marker identified as most important by feature importance analysis, coupled with a more aggressive pharmacological therapy, effectively reversing the unfavourable prognosis of the disease progression.

In conclusion, we developed computational predictive models based on 727 patient records from the POLVAS registry, coming from twelve-center study which is representative of the entire Polish population. These models, when supplied with clinical information describing a specific patient case, predict the risk of renal replacement therapy and death respectively. To compare the performance of different training approaches and address the common challenge of distributed datasets, we trained the models using centralized and federated learning methods. With centralized learning approach we achieved AUCs of 0.86 and 0.81 for RRT and death prediction, respectively. With federated models, we achieved weighted mean AUC of 0.86 and 0.80 for RRT and death prediction, respectively. Our findings suggest that while general federated models might not fully capture the nuances of individual centers, optimizing dedicated federated models might reflect better underlying heterogeneity in patient populations, clinical practices and other factors, leading to improved performance. Additionally, we performed feature importance analysis, which provides valuable insights into the nature and progress of AAV vasculitis and tested the utility of the highly explainable rule-based method for this type of problems, with encouraging results.

Author contributions. T.S., S.L., T.G., K.Woz. and K.M. were responsible for conceptual work, design of the study, and methodological development; T.S. and K.Woz. were responsible for implementing and testing the centralized machine learning models; K.M. and K.Woz. were responsible for implementing the federated models; M.Mal., J.Mus., J.S. were responsible for conceptual work and scientific leadership; K.Woj. and S.L. provided clinical consultations; K.Woj. was responsible for data curation; K.Woj., G.B, K.W.A., S.B.S., A.M., M.C., B.B.P., A.D.Ś., H.S., M.K., H.A.B., R.J., M.Maj., K.J., M.Mil., M.B., J.K.Z., W.T., M.Mad., A.H.K., E.J.K., P.G., K.B., M.Wis., J.M.D., A.L.P., J.Mak., J.Z., J.Mus. were responsible for data collection; K.M., K.Woj., K.Woz. , S.L., T.G and T.S. were responsible for preparing the manuscript.

Acknowledgments. This publication is partly supported by the European Union's Horizon 2020 research and innovation programme under grant agreement "Sano" No. 857533, and by the "Sano" project, carried out within the framework of the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13, co-financed by the European Regional Development Fund. The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This work was supported by a grant from Polish National Science Center UMO-2018/31/B/NZ6/03898 (to Jacek Musiał).

Disclosure of Interests. The authors declare no competing interests.

References

1. Salmela, A., et al.: Chronic nasal staphylococcus aureus carriage identifies a subset of newly diagnosed granulomatosis with polyangiitis patients. Rheumatol. **56**(6), 965–972 (2017)
2. Musiał, J., Wójcik, K.: Polish vasculitis registry: Polvas. Pol. Arch. Intern. Medicine **127**(1), 71–72 (2017)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17

3. Wójcik, K., et al.: Subphenotypes of ANCA-associated vasculitis identified by latent class analysis. Clin. Exp. Rheumatol. **39**(S129), 62–68 (2021)
4. Wójcik, K., et al.: Clinical characteristics of Polish patients with ANCA-associated vasculitides-retrospective analysis of polvas registry. Clin. Rheumatol. **38**(9), 2553–2563 (2019)
5. Jennette, J.C.: Overview of the 2012 revised international chapel hill consensus conference nomenclature of vasculitides. Clin. Exp. Nephrol. **17**(5), 603–606 (2013)
6. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS 2017), pp. 3149–3157 (2017)
7. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**, 12598 (2020)
8. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
9. Head, T., et al.: scikit-optimize/scikit-optimize. Zenodo. (2021)
10. McKinney, W.: Data structures for statistical computing in python. In: Proc. 9th Python Sci. Conf., pp. 56–61 (2010)
11. Turner, R., et al.: Bayesian optimization is superior to random search for machine learning hyperparameter tuning. In: Proc. NeurIPS 2020 Competition Track, PMLR **133**, pp. 3–26 (2021)
12. Bergstra, J., et al.: Algorithms for hyper-parameter optimization. In: Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS 2011), pp. 2546–2554 (2011)
13. Mandrekar, J.N.: Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. **5**, 1315–1316 (2010)
14. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. **11**, 2079–2107 (2010)
15. Ma, C., et al.: Gradient-less federated gradient boosting tree with learnable learning rates. In: Proceedings of the 3rd Workshop on Machine Learning and Systems (EuroMLSys '23), pp. 56–63. ACM, New York, NY, USA (2023)
16. Beutel, D.J. et al.: Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390 (2020)
17. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning (ICML '95), pp. 115–123. Morgan Kaufmann Publishers, San Francisco, CA, USA (1995)
18. Ali, K., Brunk, C., Pazzani, M.: On learning multiple descriptions of a concept. In: Proceedings of the International Conference on Tools with Artificial Intelligence (ICTAI 1994), pp. 476–483. IEEE, Los Alamitos, CA, USA (1994)
19. Khang, V.H., Anh, C.T., Thuan, N.D.: Detecting fraud transactions using RIPPER algorithm combined with ensemble learning model. Int. J. Adv. Comput. Sci. Appl. **14**, 336–345 (2023)
20. Govada, A., Thomas, V.S., Samal, I., Sahay, S.K.: Distributed multi-class rule-based classification using RIPPER. In: Proc. IEEE Int. Conf. Comput. Inf. Technol. (CIT 2016), pp. 303–309. IEEE (2017)
21. Kubat, M.: An introduction to machine learning. Springer International Publishing (2015)
22. Lezcano-Valverde, J.M., et al.: Development and validation of a multivariate predictive model for rheumatoid arthritis mortality. Sci. Rep. **7**, 10189 (2017)
23. Lin, C., et al.: Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. PLoS ONE **8**(8), e69932 (2013)
24. Tsujitani, M., Sakon, M.: Analysis of survival data having time-dependent covariates. IEEE Trans. Neural Netw. **20**(3), 389–394 (2009)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_17