Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 156 (2019) 87–96

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Young Scientist Conference on Computational Science

Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients

Ksenia Balabaeva*, Sergey Kovalchuk

ITMO University, 197101, 49 Kronverksky pr., St Petersburg, Russia

Abstract

Chronic diseases are complex systems that can be described by various heteroscedastic data that varies in time. The goal of this work is to determine whether historical data helps to improve machine learning predictive models or is it more efficient to use the latest data describing the disease in particular moment in time. For simplicity we call features from the first group dynamic and features from the second one – static. We study the way both groups affect predictions quality and its interpretation. We set the experiments on data of chronic heart patients from Almazov Medical Research Center. From this data we extracted more than 300 features from patient comorbidity, anamnesis, analysis, etc. In terms of Chronic Heart Failure (CHF) modelling three different tasks have been selected: CHF identification as main diagnosis, CHF stage classification and diastolic blood pressure prediction. For each task several machine learning algorithms on three groups of features: static, dynamic and the whole feature set. The results show that, in general, models perform better on combination of temporal and non-temporal features.

Keywords: chronic heart failure; predictive modelling; complex systems modelling; machine learning; interpretable machine learning

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: kyubalabaeva@gmail.com

10.1016/j.procs.2019.08.183

1. Introduction

Improving quality of decision making in a healthcare, making it more personalized and accurate is a complex task. One of the approaches to address these challenges is machine learning algorithms and data analysis. However, when it comes to chronic diseases that can be defined as complex systems, new problems arise.

1.1. Chronic Heart Failure

This work is based on data of more than 40 000 chronic heart failure (CHF) patients provided by Almazov Medical Research Center. CHF is a progressive syndrome that results in a poor quality of life for the patient. It is a syndrome that develops as a result of abnormal ability of the heart to fill and/or birth occurring in a state of imbalance vasoconstrictor and vasodilating neurohormonal systems; accompanied by inadequate perfusion of organs and tissues of the body and manifested complex of symptoms: shortness of breath, weakness, heartbeat increased fatigue and fluid retention [1]. The prevalence of CHF in western countries is different. It ranges from 1 to 2 % in the general population, up to 10% in individuals over 70 years old [1]. Almost half (45%) of patients with CHF die from sudden cardiac death, death from heart attack or stroke that is much less common (less than 2%). That is why modelling progression of CHF is an important task.

We have conducted experiments on three different tasks: CHF identification, CHF stage prediction, blood pressure prediction. For each task we select groups of dynamic and static features and analyze their effect and contribution to the prediction results.

2. Dynamic system concept

Chronic diseases can be described as a complex process which is a series of states actions and parameters that change over time [2]. Formally, complex system consists of a tuple $P(t) = \langle X(t), E(t), Z(t) \rangle$ .Where $X(t)$ is a sequence of multidimensional continuous or discrete values, changing over time, $E(t)$ is a sequence of events, and $Z(t)$ is a sequence of states.

Sequence $E(t)$ consists of events $e_i = \langle t_i, c_i, A_i \rangle$, where $t_i \in T^{(E)}$ is an event occurrence time, $c_i \in C^{(E)}$ is an event class label, A_i – a dictionary of attributes with keys specific for particular class label $\mathrm{c_i}$. Sequence of states $\mathrm{Z(t)} \in \mathrm{Z_1} \times \dots \times \mathrm{Z_t}$ is defined for n different state series. This sequence includes planned or scheduled states and random states changing within a relatively short period of time.

In terms of chronic heart failure, a complex system could be illustrated with a Figure 1. Here we see a set of events spread in time for a particular CHF patients: inpatient and outpatient episodes, dates of analysis collection and blood pressure measurements.

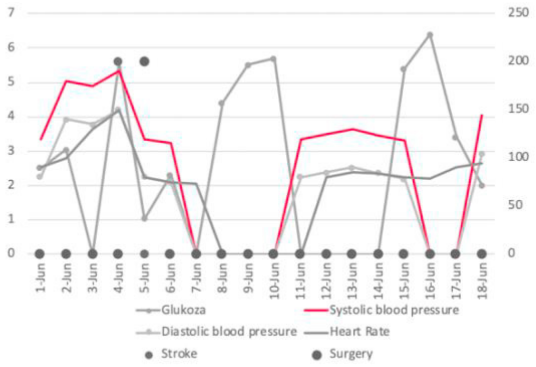

Figure 1Chronic heart failure as a complex system

From this picture we may conclude that patient records is a set of many various time series data. But these time series are not regular: they have different time spacing between moments of data collection. This time series irregularity is another challenge for chronic disease modelling, which we would like to address. But first, let's take a look at existing methods and approaches.

3. Related works

Various approaches to model stochastic processes that evolve over time exist, depending on task specification and available data. From decision rules, proportional hazards model to Markov decision processes, recurrent neural networks, dynamic bayesian networks, etc. [6] For instance, in 2013 the joint longitudinal modelling approach for rehospitalization problem have been explored on daily recorded blood pressure, weight and heart rate measurements [3]. There are also Dynamic Bayesian networks (DBNs) that offer an approach allowing the incorporation of the temporal and causal nature of medical domain knowledge as elicited from domain experts, thereby allowing for detailed prognostic predictions. One of the latest articles [4] is also dedicated to LSTM efficacy to predict the onset of CHF 15 months in advance using 12-month records. Moreover, such neural network architectures as GRU (RNN with gated recurrent units) achieves state-of-the-art quality level in medical predictions due to long temporal dependencies and missing time series values efficient treatment [5]. However, such models require tens and hundreds of thousand patient records and cannot be efficiently trained on datasets of less volume. Moreover, deep neural networks can hardly be interpretable.

Therefore, we would like to explore the efficiency of models that do not consider temporal effects in their structures but still can observe dynamic through the input data with temporal features. For each task 3 machine learning algorithms were trained: XGBoost, logistic regression, and random forest for classification and linear regression, decision tree and XGBoost for the regression task. We train models on subgroups of static and dynamic features, and then on their combination. The experiment results for CHF stage prediction are performed in the Table 1. For score calculation the F1 score with macro averaging is used. Then we analyze the feature importances and interpret the results applying SHAP method.

4. Data

For each of three tasks a single dataset was collected on the base of chronic heart failure patients from Almazov Medical Research Center (Table 1). The size of datasets varies due to the different occurrence of target variables.

Table 1 Datasets information

Task	Num. of patients	Num. of episodes
CHF diagnosis identification	165 353	290 679
CHF stage prediction	743	1 279
Blood pressure prediction	13 213	24 156

For input features we use demographic patient information (age, gender, etc.), anamnesis, comorbidity, blood and urine tests, blood pressure measurements, BSA, BMI and hospital processes data (number of episodes in the hospital, episode type, patient history duration etc.). Prediction in each task is made for the specific episode.

5. Experiments Settings

In order to model complex dynamic systems and make accurate predictions we need to identify the way that models can observe dynamical nature of data. As was discussed in related works section, existing approaches have several drawbacks. Therefore, we would like to study if machine learning models that do not observe any temporal feature relations internally will benefit from the input features engineered in a way to reflect temporal effects. In simple words, we select a group of non-temporal (static) features, describing only examined episode, for which a prediction is made, and temporal (dynamic) features, describing all previous patient records). To overcome the problem of inconsistent time spacing in data we use several aggregation functions (mean, median, standard deviation, max values, min values, etc.) for feature engineering in dynamic group (Figure 2).

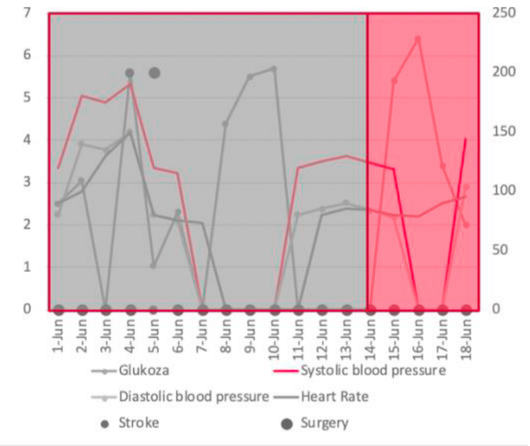

Figure 2 Temporal (grey area) and non-temporal (red area) features

For each prediction task we use the following pipeline. First, we preprocess raw data, calculate new features and divide the whole feature set into three groups: static, dynamic and mixed. Then we conduct experiments to select the best feature scaling method. After that, in order to reduce dimensionality and improve models performance we use forward feature selection algorithm. It filters features due to their positive or negative impact on the quality metrics and helps to improve the results. The next step is to train machine learning models on selected dataset, fine tune them and check performance using cross-validation. Finally, we interpret models predictions using SHAP (Shapley Additive exPlanations) algorithm [7]. The main concept is using game theory interpret complex machine learning model prediction. From this perspective all features are players whose goal is to contribute to the result, and their 'reward' is actual prediction minus the result from explanation model. SHAP is based on Shapley values from cooperative game theory that is formally defined as:

$$\phi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)] \qquad (1)$$

Where $\phi_i(v)$ – Shapley value , N – number of players (features), P_i^R – set of players with order, $v(P_i^R)$ – contribution of set of player with order, $v(P_i^R \cup \{i\})$ – contribution of set of players with order and player i.

For each task, we run the pipeline three times, using only static features, dynamic and their combination. The experiment results are provided in results and discussion section.

6. Results and Discussion

6.1. CHF identification

In this task we predict if the main episode diagnosis will be chronic heart failure. Therefore, such task may be considered as a binary classification, with label 1 if diagnosis is CHF and 0 – if not. For this task we had the largest dataset volume. The details of experiments are described in section 1.5. The only thing we need to add is that as classification models we used XGBoostClassifier, Logistic Regression, Decision Tree and Random Forest. And as a quality metrics we used F1-score with macro averaging, since classes were imbalanced.

One of the stages of our pipeline was dedicated to best scaling method selection. So, in Table 2 the results of selection are performed. We see that in this task only logistic regression boosted it's performance using standard scaler, which may be explained by the tree-type nature of XGB, Decision Tree and Random Forest.

Table 2 Scaling methods performance for CHF identification task

Model\Scaling Method	No scaling	Standard Scaler	Min Max Scaler	Max Abs Scaler	Robust scaler	Quantile Transformer Normal	Quantile Transformer Uniform
XGB	**1.0**	1.0	1.0	1.0	1.0	1.0	1.0
LogReg	0.808	**0.822**	0.809	0.809	0.820	0.794	0.791
Decision Tree	**0.999**	0.999	0.999	0.999	0.999	0.999	0.999
Random Forest	**0.999**	0.997	0.999	0.999	0.997	0.997	0.999

Concerning our main goal – comparison of static and dynamic feature groups performance in CHF we have the following results (Table 3). We may conclude that for all models the dynamic feature group outperformed the static one on both test and cross-validation sets. For majority of the ML models dynamic features set beats the mix of static and dynamic features except for logistic regression that slightly benefit from the use of all features on test set.

Table 3 Temporal and non-temporal features impact in CHF ident. task

Model\Scaling Method	Static Features		Dynamic Features		All features	
Model\Metrics	F1-score (test)	F1-score (cross-val.)	F1-score (test)	F1-score (cross-val.)	F1-score (test)	F1-score (cross-val.)
XGB	0.8679	0.6316	**1.0**	**0.9055**	0.9999	0.9038
LogReg	0.7873	0.6253	0.8124	**0.6676**	**0.8196**	**0.6676**
Decision Tree	0.8117	0.6320	**0.9998**	**0.9092**	0.9997	0.9092
Random Forest	0.8126	0.6313	**1.0**	**0.9093**	0.9999	0.9092

In terms of quality, the best model is XGB. Let's discover how its interpretation changed from using different temporal feature sets. In non-temporal feature set (Figure 3.1) the most important variables for XGB are patient age, gender and episode type. On the vertical axis are features sorted by the mean SHAP value. Each dot in the feature row determines single patient, the color represents the feature value and the x position is the impact of that feature on the model's prediction for that customer. It is interesting, that the older patient is - the more impact on model output the age has. The same happens to the gender: being a man, on average, increases the SHAP values. The episode type also can take 2 values (inpatient and outpatient type). Here we see that inpatient episode type, on average decreases the impact, while outpatient helps model to improve.

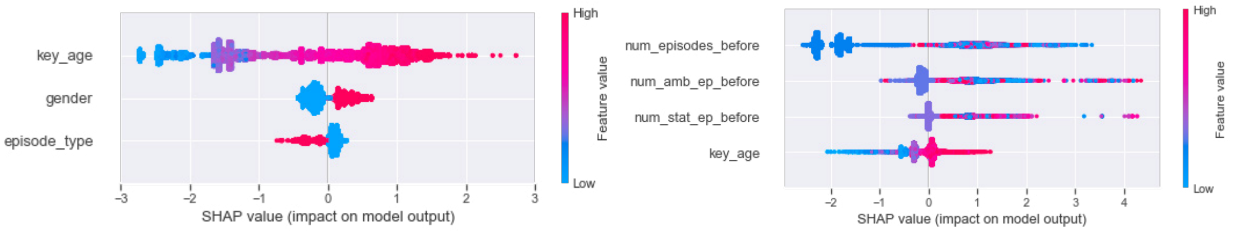

Figure 3 CHF identification task: XGB interpretation using SHAP static (3.1) and dynamic (3.2) feature subsets

For temporal feature set the most important features were number of episodes before and separate numbers of outpatient and inpatient episodes. Together with them the patient age is also informative its values have similar relation to Shap values, as in static group, however, in dynamic the dispersion is smaller so as the magnitude. Now let's analyze temporal and non-temporal subset. On figure 4 we see that top features are exact combination of previous plots. While dynamic features get higher positions in terms of impact and their distributions almost have not changed, the static features get down. What is more interesting is that episode type and gender binary variables are no more as strictly separated by type on the Shapley value.

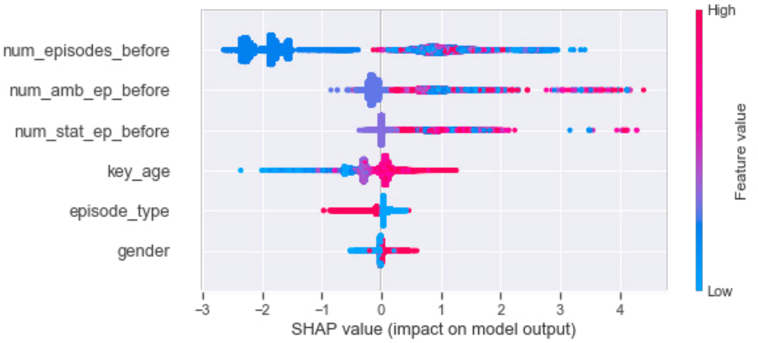

Figure 4 CHF identification task: XGB interpretation using SHAP all features

6.2. CHF stage prediction

According to heart remodeling severity, 4 stages of chronic heart failure can be identified:

- Stage 1. Initial stage of heart disease, when hemodynamics alone is not broken and there is asymptomatic dysfunction of the left ventricle. Encoded as 1.
- Stage 2 A. Clinically severe stage of heart disease with adaptive remodeling of the heart and blood vessels. Encoded as 2.
- Stage 2 B. Severe stage of heart disease with pronounced changes in hemodynamics in both circles of blood circulation and disadaptive remodeling of the heart and blood vessels. Encoded as 3.
- Stage 3. The final stage of heart damage. Pronounced changes in hemodynamics and irreversible structural changes in other organs. Encoded as 4.

Therefore, it is a multiclassification task. The target labels are also disbalanced with the most popular of 2B stage. As algorithms we test XGB, LogReg, Random Forest, Decision Tree and compare the results using F1-score with macro averaging.

Scaling selection experiments are provided in the table 4, The min max scaler and power transformer almost doubles the performance of logistic regression and Quantile transformers help XGB and Random Forest to improve, however decision tree results remain constant with or without scaling.

Table 4 Scaling methods performance for CHF stages classification (mixed features)

Model\Scaling Method	No Scaling	Standard Scaler	Min Max Scaler	Max Abs Scaler	Robust Scaler	Quantile Transformer Normal	Quantile Transformer Uniform	Power Transformer Yeo Johnson
XGB	0.4436	0.4436	0.4436	0.4436	0.4436	**0.4485**	**0.4485**	0.4436
LogReg	0.2757	0.4504	**0.4974**	0.4847	0.2692	0.4657	0.4844	0.4936
Decision Tree	**0.4151**	0.4151	0.4151	0.4151	0.4151	0.4151	0.4151	0.4151
Random Forest	0.5425	0.5425	0.5425	0.5425	0.5425	**0.5592**	**0.5592**	0.5425

Temporal experiments results are displayed in the Table 5. As in the previous task, for multiclassification static features set is the least valuable for prediction quality and all models benefit on cross-validation set from combination of temporal and non-temporal features.

Table 5 Temporal and non-temporal features performance in CHF stages classification task

Model\Scaling Method	Static Features		Dynamic Features		All features	
Model\Metrics	F1-score (test)	F1-score (cross-val.)	F1-score (test)	F1-score (cross-val.)	F1-score (test)	F1-score (cross-val.)
XGB	0.4128	0.3853	0.4838	0.4459	**0.5096**	**0.4516**
LogReg	0.4915	0.4332	**0.5565**	0.4653	0.532	**0.4901**
Random Forest	0.4721	0.3599	0.6062	0.5841	**0.6287**	**0.5901**
Decision Tree	0.3938	0.3461	**0.4044**	0.4012	**0.4044**	**0.4236**

For interpretation analysis we use random forest (RF) predictions as the most accurate model. From the static feature group (figure 5.1) RF output mainly depends on several heart diseases as main diagnosis, from which the main one is CHF. More than that, a high impact model gets from mean and minimal values of systolic blood pressure and mean and maximal values of diastolic blood pressure. However, the second place takes minimal value of PLT. From dynamic group (figure 5.2) the highest impact to the model output gives max previous CHF stage and the standard deviation of previous stages, which is obvious, since it is highly correlated with the target variable. Other important features are connected with previous blood and urine tests.

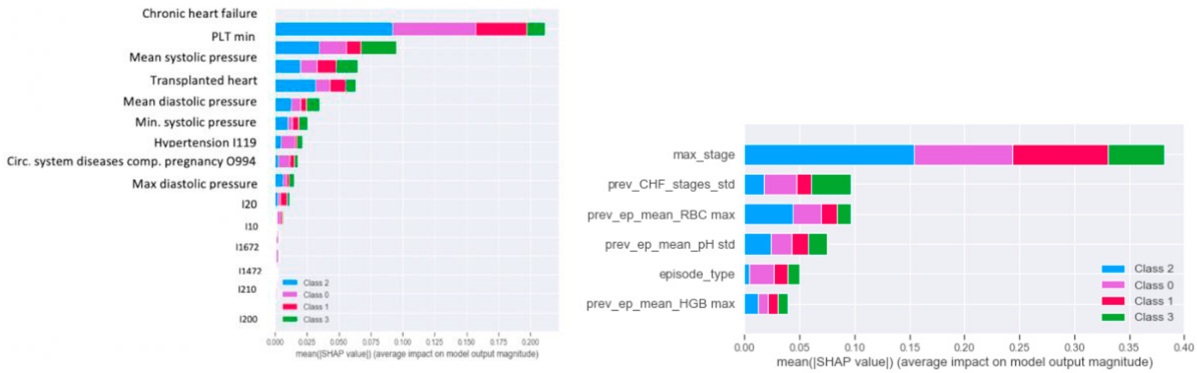

Figure 5 CHF stage prediction: RF interpretation using SHAP static (5.1) and dynamic (5.2) feature subsets

The combination of temporal and non-temporal features helps to get the highest results for RF. On the figure 6, we see that the first place is taken by the dynamic feature (maximal previous CHF Stage), followed by the CHF as a main diagnosis current episode statistics on blood tests and pressure measurements, main diagnosis, as well as some previous statistics.

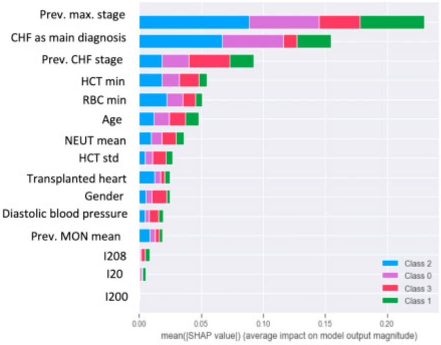

Figure 6 CHF stage prediction: RF interpretation using SHAP on all feature subsets

6.3. Blood pressure prediction

Blood pressure is one of the important indicators in chronic heart failure diagnosis and treatment. There are 2 types of blood pressure systolic and diastolic. For simplicity, we take only systolic pressure as the target variable. Normally systolic blood. Pressure should take values between 90 and 120 mmHg, however it may grow up to 180 mmHg and higher.

Therefore, systolic pressure prediction is a regression task, and we want to test our temporal feature subset on this task as well. As a regression algorithm we examine XGBoostRegressor, linear regression, decision tree and random forest. To compare the performance, we use RMSE metrics. But first, let's discuss the results of scaling methods.

Among all scaling methods (table 6) a significant improve gets only linear regression. On opposite, other tree type algorithms may gain from scaling only one thousandth RMSE improve, which is insignificant in terms of physical value of a target variable.

Table 6 Scaling methods performance for blood pressure prediction

Model\Scaling Method	No Scaling	Standard Scaler	Min Max Scaler	Max Abs Scaler	Robust Scaler	Quantile Transformer Normal	Quantile Transformer Uniform	Power Transformer Yeo Johnson
XGB	12.816116	**12.813315**	12.817549	12.817775	12.885244	13.103674	13.12458	12.873367
Linear Regression	17.558071	17.558154	17.557937	17.559313	17.557920	**14.185189**	1.099512e+13	15.742082
Decision Tree	12.431480	**12.431199**	**12.431199**	**12.431199**	**12.431199**	12.513183	12.513183	**12.431199**
Random Forest	**11.648463**	11.654806	11.654425	11.656689	11.649284	11.892355	11.91343	11.696527

Unlike in the previous tasks, the temporal features experiments in this section lead to the opposite conclusion. For blood pressure prediction the static feature set is more valuable (table 7): all models perform better on static features on the train set. However, on cross validation set, the combination of static and dynamic features gives more accurate results, but the difference is not significant. These results can be explained by the high volatility of blood pressure and that the latest information is more valuable than the past.

Table 7 Temporal and non-temporal features performance in blood pressure prediction task

Model\Scaling Method	Static Features	Dynamic Features	All features

Model\Metrics	RMSE (test)	RMSE (cross-val.)	RMSE (test)	RMSE (cross-val.)	RMSE (test)	RMSE (cross-val.)
XGB	**12.8527**	13.2622	18.5047	18.9644	12.9070	**13.2313**
LogReg	**13.5186**	13.665	17.7776	18.4615	13.5202	**13.6639**
Decision Tree	**12.4167**	**12.6285**	17.5632	18.292	**12.4167**	**12.6285**
RF	**11.7161**	12.2154	17.8906	18.7772	11.8224	**12.2413**

Concerning interpretation of the non-temporal feature set (figure 7), the highest impact on the output has mean diastolic pressure, which is highly correlated with diastolic pressure. We also see that higher values of diastolic pressure, on average, lead to the bigger increase in output value. Minimal and maximal values of diastolic pressure are also important. Another significant features are patient age, LPVP and NEUT from blood tests. In dynamic feature set, the most significant feature is episode type, age, gender and number of previous episodes of different types. In the combined feature set (figure 8) top important features are primarily from the static group.

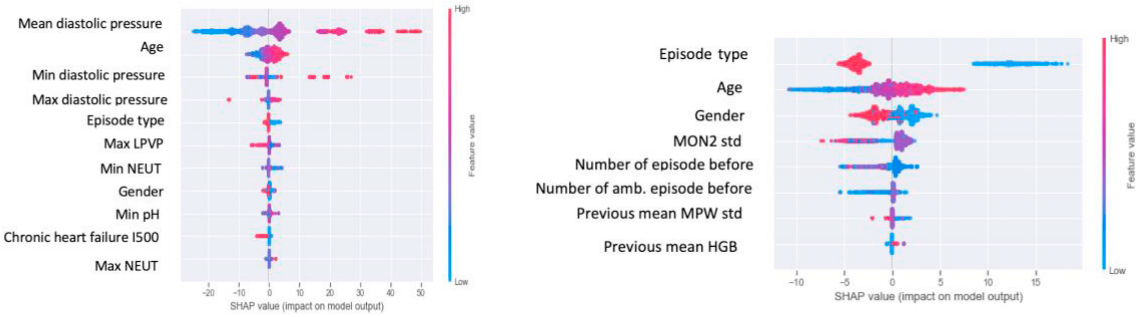

Figure 7 Blood pressure task: Decision tree interpretation on static (7.1) and dynamic (7.2) feature subsets

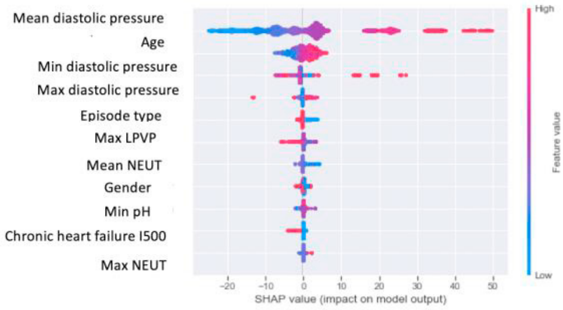

Figure 8 Blood pressure task: Decision tree interpretation on all features

To illustrate how this interpretation works on the level of a single patient, we provide the figure 9. It shows that the baseline systolic pressure value is equal to 125.8 mmHg, but the model prediction for this patient is 105.5. To understand why, we look at the blue features that negatively affected model output and pine ones, that affected it positively. Since the impact of blue features is higher, primarily due to the mean and minimal diastolic pressure values, than the impact, the model output is lower than the expected value.

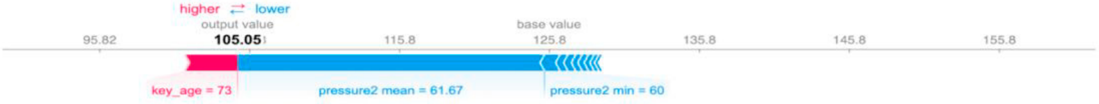

Figure 9 Blood pressure task: Decision tree interpretation on all features

7. Conclusion

In general, we may conclude that in binary and multi classification tasks, where target variables do not change in a high rate, ML models can boost the quality of performance using dynamic features from the past. However, when it comes to regression task with highly volatile targets, the latest static information might be more valuable. In order to understand the reasons lying behind the regression results the further experiments can be conducted. For instance, to analyze other methods for past data aggregation and different features engineering on other tasks.

Considering temporal and non-temporal features impact on the model interpretability, we see that if the dynamic feature improves model's quality, correlates with the target variable and is one of the most significant variables inside the dynamic set, it will likely remain its place in combination feature set. However, the magnitude of impact and its effect on particular target values may change. This effect might be connected with multicollinearity between features.

One more supplementary conclusion may be drawn from the scaling experiments results. First of all, we should say, that there is no universal scaling method that can affect all models' performances in the same way. Logistic regression performance can be significantly improved with scaling, due to its linearity. Moreover, even though the algorithms based on trees rely on the order of feature values, sometimes scaling may help to improve them too (Quantile Transform).

Finally, in the present paper we discussed the problem of complex dynamic systems modelling illustrated by example of chronic heart failure disease. We've analyzed how heterogeneous dynamic and static features affect the prediction quality on three different tasks. The experimental results confirm that the highest quality, on average, for classification task is achieved using the mix of static and dynamic features and for the regression task the static sample is more efficient, which might be due to the higher volatility of target values.

Comparing to other state of the art methods, we should say that the undermined approach deals with irregular multidimensional time series data. Other methods (ex. RNN) can use only heterogenous time-series that have a constant period between data collection, which is not the case of chronic diseases. The next step is to computationally compare the examined approach with other dynamic modelling methods.

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement #19-11-00326

References

[1] "Kardiologiia" **58 (S6)** (2018), http://webmed.irkutsk.ru/doc/pdf/hf.pdf

[2] Krikunov A, Bolgova E, Krotov E, Abuhay T, Yakovlev A, Kovalchuk S. (2016) "Complex data-driven predictive modeling in personalized clinical decision support for acute coronary syndrome episodes." *Procedia Computer Science* **80 :** 518-529.

[3] G.V. Ramani. (2013) "A Joint Survival-Longitudinal Modelling Approach for the Dynamic Prediction of Rehospitalization in Telemonitored Chronic Heart Failure Patients." *Statistical Modelling* **13(3)** : 179– 198

[4] S. Mallya. (2019), "Effectiveness of LSTMs in predicting Congestive Heart Failure onset.", preprint

[5] P.A. Deus. (2018), "Recurrent Neural Networks for Multivariate Time Series with Missing Values." *Scientific Reports* **8** : 6085

[6] A.J. Marcel (2008), " Dynamic Bayesian networks as prognostic models for clinical patient management." *Journal of Biomedical Informatics*, **41(4)**: 6085

[7] Lundberg, Lee. (2017) "A Unified Approach to Interpreting Model Predictions", URL: https://arxiv.org/pdf/1705.07874.pdf