Mathematical Biosciences and Engineering

MBE, 20(7): 13398–13414.
DOI: 10.3934/mbe.2023597
Received: 17 April 2023
Revised: 16 May 2023
Accepted: 22 May 2023
Published: 12 June 2023

http://www.aimspress.com/journal/MBE

Research article

A non-linear SVR-based cascade model for improving prediction accuracy of biomedical data analysis

Ivan Izonin[1,*], Roman Tkachenko[2], Olexander Gurbych[1], Michal Kovac[3], Leszek Rutkowski [4,5,6] and Rostyslav Holoven[7]

[1] Department of Artificial Intelligence, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine
[2] Department of Publishing Information Technologies, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine
[3] Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovak Republic
[4] Systems Research Institute of the Polish Academy of Sciences, Warsaw, Poland
[5] AGH University of Science and Technology, Krakow, Poland
[6] Information Technology Institute, University of Social Sciences, Lodz, Poland
[7] Department of System Design, Ivan Franko National University of Lviv, Lviv, Ukraine

* **Correspondence:** Email: ivanizonin@gmail.com; Tel: +380988889687.

Abstract: Biomedical data analysis is essential in current diagnosis, treatment, and patient condition monitoring. The large volumes of data that characterize this area require simple but accurate and fast methods of intellectual analysis to improve the level of medical services. Existing machine learning (ML) methods require many resources (time, memory, energy) when processing large datasets. Or they demonstrate a level of accuracy that is insufficient for solving a specific application task. In this paper, we developed a new ensemble model of increased accuracy for solving approximation problems of large biomedical data sets. The model is based on cascading of the ML methods and response surface linearization principles. In addition, we used Ito decomposition as a means of nonlinearly expanding the inputs at each level of the model. As weak learners, Support Vector Regression (SVR) with linear kernel was used due to many significant advantages demonstrated by this method among the existing ones. The training and application procedures of the developed SVR-based cascade model are described, and a flow chart of its implementation is presented. The modeling was carried out on a real-world tabular set of biomedical data of a large volume. The task of predicting the heart rate of

individuals was solved, which provides the possibility of determining the level of human stress, and is an essential indicator in various applied fields. The optimal parameters of the SVR-based cascade model operating were selected experimentally. The authors shown that the developed model provides more than 20 times higher accuracy (according to Mean Squared Error (MSE)), as well as a significant reduction in the duration of the training procedure compared to the existing method, which provided the highest accuracy of work among those considered.

Keywords: cascading; data analysis; biomedical data; Ito decomposition; ensemble model; linear Support Vector Machine; non-linear input extension; prediction tasks

1. Introduction

The modern development of the post-industrial society requires providing high-quality and timely service for potential consumers. Largely, this also applies to medical services, where the amount of information received about the patient significantly affects the accuracy of the decisions made by the doctor. However, large volumes of various types of data that need to be processed considerably affect the quality of such decisions [1]. In the Big Data era, the tasks of increasing the accuracy of the medical diagnosis, treatment, or monitoring of the patient's condition are entirely based on the intellectual analysis of biomedical data. Such data can be collected in different ways. Still, the increased computing power, the appearance of a large number of portable devices, and broadband internet access in recent years provide the possibility of collecting biomedical data precisely with the Internet of Medical Things (IoMT) [2]. This approach has many advantages in medicine, especially for remote monitoring of the patient's condition. However, it is also accompanied by significant risks, the main of which can be reduced to the problems of accurate and fast data processing. The effective solution to both problems is the key to using the IoMT in practice.

When we consider the task of intellectual analysis of large sets of numerical data, the most optimal solution in terms of their processing speed is the use of linear ML methods [3]. Effective use of such models can also occur hardware-wise, particularly for the implementation of Edge- and Fog computing due to the low computational complexity of their work. This approach will reduce the load on the server where data processing takes place and energy costs for data transfer, replacing it with the transfer of current knowledge.

Despite the high speed of operation, linear models do not always provide sufficient accuracy of intelligence analysis results for their practical use. The use of non-linear models eliminates this drawback. They significantly increase the accuracy of solving applied regression/classification tasks when processing large sets of biomedical data. However, they require considerably higher resources for the implementation of training procedures. It will significantly slow down the functioning of applied medical diagnostics or monitoring systems, which imposes several restrictions on their practical application.

Based on this, a contradiction arises between preserving the speed of system operation due to linear data analysis models and ensuring the high accuracy of intellectual analysis, which only non-linear models allow obtaining. To solve it, we proposed to use a linear model (SVR with the linear kernel) with a non-linear expansion of inputs (second-degree Ito decomposition). This approach is justified by Cover's theorem [4] which says that the transformation of data into a space of higher

dimensions increases the probability of correct classification of data by linear methods [5]. In addition, it will significantly increase prediction/classification accuracy with a slight increase in the duration of training procedures.

Even with increased accuracy, solving applied intellectual analysis tasks using single models does not always provide a sufficiently effective result. In this case, one effective way to improve prediction/classification accuracy may be to use ensembles from such models. Among the four classes of ensemble methods: boosting, bagging, stacking, and cascading, it is the last one that provides the highest accuracy of work. However, the composition of cascading methods can be different, as well as the principles based on them [6]. Depending on this, the general model can demonstrate high/low accuracy and be very slow in operation.

This paper aims to develop an effective cascading model based on the linear ML algorithm with the non-linear expansion of inputs. This model is based on the principle of response surface linearization by considering the outputs from the previous cascade level to the following one. Therefore, it will provide high accuracy during the analysis of large biomedical tabular datasets with a slight increase or even decrease in the duration of its training procedure compared with analogs.

The main contribution of this paper can be summarized as follows:

- We have created a new ensemble model based on Ito decomposition and linear Support Vector Machine (SVM) for improving the prediction accuracy of biomedical data processing. It is based on the idea of cascading and the principles of the response surface linearization;
- We have studied and chosen optimal parameters of the proposed non-linear SVR-based cascade model that provide the best prediction accuracy using different performance indicators;
- We have compared our model with other ML-based methods from different classes and showed its effectiveness when solving biomedical data analysis tasks in case of large tabular data processing.

The structure of the paper is as follows. The second section provides an overview and analysis of existing approaches to solving the stated problem. Section 3 contains a detailed description of the proposed model and the method for its implementation. Section 4 presents simulation results based on a real-world dataset. A comparison with existing approaches and a discussion on the effectiveness of using the proposed model is given in Section 5. Conclusions are given in Section 6.

2. State-of-the-arts

The applied problem that was solved in this paper consists of predicting heart rate based on a large volume biomedical data set. Its primary purpose is to determine a person's stress level. That is why the review of existing works focused on linear, non-linear and ensemble ML methods for solving the stated task.

The paper [7] considered the problem of heart rate prediction based on multiple regression. The dataset contained a set of physiological and demographic indicators. The authors obtained a high prediction accuracy based on various accuracy indicators. However, the shortcoming of this study is the small dataset on which the simulation took place.

The review [8] analyzes linear and non-linear ML methods for solving the stated task. The authors selected studies that use various tools to solve the heart rate prediction task by solving classification or regression tasks using datasets of different volumes. The results of the analysis show that linear methods provide high prediction accuracy only in the case of processing small datasets. Therefore it is advisable to use non-linear methods to analyze other volumes of data.

Research [9] is dedicated to solving the classification task for heart rate prediction. The authors suggested using Bayesian Inference Federated Learning for its solution. The modeling of the proposed method took place using a real-world dataset collected by the IoMT device. The results regarding the accuracy of the work based on the author's errors look satisfactory. However, this approach requires setting a large number of parameters. That is why it is computationally inefficient to solve the stated problem.

The authors of [10] carried out a study on the evaluation of the effectiveness of using linear and ensemble ML methods to solve the stated task. In this case, the authors used many time series collected for different intervals. The results demonstrate the high accuracy of all linear methods, among which SVR stands out. In addition, ensemble methods and long-short-term memory did not provide sufficient accuracy and an unstable result, especially when using different sliding windows.

In [11], the task of early diagnosis of heart diseases is considered. The authors used a large medical dataset to solve the classification task. Synthetic Minority Over-sampling Technique (SMOTE) algorithm was used to balance the dataset. The effectiveness of single models, homogeneous and heterogeneous ensembles, was investigated. The basis of each of the considered approaches is the use of AdaBoost. The simulation results showed that the highest accuracy was obtained using a heterogeneous ensemble. However, the disadvantage of this approach is the need to adjust all optimal parameters of each heterogeneous ensemble member. It requires a lot of resources.

Research [12] is devoted to applying a boosting strategy to increase the accuracy of solving the heart rate prediction task. Such a strategy provides sufficient prediction accuracy with satisfactory time performance of the model. To increase the prediction accuracy, the authors suggested using an approach to feature selection called recursive feature elimination. The modeling results confirmed the higher accuracy of using the hybrid approach compared to the basic Gradient Boosting Regressor.

The authors of [13] proposed a two-step model for solving the stated task. It is based on the use of both numerical datasets and images. Patient's data was collected by authors of [13] using IoMT device. In the first stage, numerical data is classified using hybrid linear discriminant analysis with the modified ant lion optimization. If the results do not meet the specified criteria, a second stage is added-processing the echocardiogram by deep learning tools based on an extensive known dataset. This approach demonstrated good prediction accuracy but required many resources for its practical implementation.

The two-step multistage model, which is essentially a cascade, was developed in [14]. The authors used a General Regression Neural Network (GRNN) as the fundamental element of the cascade. At the first level of the cascade, the desired value is predicted. The second level of the cascade is designed to predict the errors of the first one. The results of using this approach have demonstrated their effectiveness but are limited to the use of not only small datasets. This is explained by the peculiarities of GRNN operation, which becomes very slow and significant in analyzing large datasets.

A multistage model was also developed in [15]. However, unlike in the previous study, the user chooses an arbitrary number of levels of the cascade model. The basis of the model is using the principle of response surface linearization by taking into account the outputs of the previous cascade in the next one. SVR with the radial basic functions (RBF) kernel was used as weak regressors. It eliminates the need to apply additional procedures for the non-linear expansion of inputs. However, the simulation results did not demonstrate a significant increase in the accuracy of the model in comparison with the classic SVR with the RBF kernel. In addition, due to the use of non-linear SVR, the developed model requires a lot of time to implement the training procedures. Overcoming these limitations is the basis of the cascade developed in this paper, which demonstrates a significant increase

in accuracy with a small duration of the training procedure.

3. A nonlinear SVR-based cascade model

3.1. SVR

The SVM is one of the most effective and flexible ML methods that was developed in the 1960s and 1970s. The basis of the method is the need to build an optimal hyper-surface, which draws the border between data clusters with the most significant distance to them. Even though this technique was developed for solving classification tasks, it has been successfully adapted and is widely used for solving regression tasks. SVR is the same algorithm but it is used for solving regression task.

The learning process of this method boils down to solving a quadratic programming problem with a unique solution. This method remains a somewhat effective linear model even when processing large datasets. In addition, SVR ensures efficient use of RAM as it uses only a subset of support points in the objective function. Detailed mathematical descriptions of the implementation of linear SVR are given in [16].

Generalization of the method for the case of non-linear response surfaces analysis is based on several kernel functions. Their main essence consists in mapping the primary space of input data in the space of a higher dimension in order to build a better separating hypersurface. Even in this case, the SVR learning algorithm does not change significantly, which ensures the flexibility of this method when solving various classification or regression tasks.

3.2. Ito decomposition

The Wiener polynomial or Kolmogorov-Gabor polynomial or Ito decomposition [17] is a discrete analog of the Voltaire series. It was developed in parallel by several scientists from different countries to solve various approximation tasks, particularly for synthesizing a non-linear prediction filter [18].

Ito decomposition of the second degree can be written as follows:

$$Y(x_1,...,x_n) = a_i + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=i}^{n} a_{i,j} x_i x_j \,, \tag{1}$$

where $x_1,...,x_n$ are independent variables; Y is a function from independent variables; a_i are the coefficient of this decomposition, $i = 1,n$.

The peculiarity of this non-linear decomposition is its high approximation properties, which ensure its use when solving several applied problems [19].

In this paper, we will use second-order Ito decomposition for non-linear expansion of the input data space. It will ensure an increased approximation accuracy by linear models (justified by Cover's theorem). Higher degrees of this decomposition should not be considered when processing large volumes of data. It is due to a significant increase in the number of attributes of the dataset for processing, which will increase the complexity of calculations and the training time of the selected ML model. In addition, a higher degree of Ito decomposition can cause overfitting of the chosen ML model, especially when processing large datasets.

3.3. SVR-based cascade model

In this paper, we developed a new SVR-based cascade model. It is based on one of the approaches to the ensemble of ML methods, namely cascading. Cascading provides the most accurate prediction or classification results when solving various applied tasks among the four ensemble strategies.

The training sample division accompanies the developed model's implementation into several subsamples with the same or almost the same number of vectors. The number of subsamples will determine the levels numbers of the developed SVR-based cascade model.

At the input of each cascade level, the dataset is normalized and subjected to a non-linear transformation using the Ito decomposition. This step maps the original input data space (each subsample) into a higher dimensional space to obtain a higher prediction accuracy. In addition, the predicted value from the previous level of the cascade model is used by the next level as an additional attribute for prediction. The principle of the response surface linearization underlying this approach also increases the accuracy of the approximation of a current dataset.

A SVR with a linear kernel was chosen as the basic regressor of the developed model. Its use as a weak learner at each cascade level has several significant advantages. First, according to Cover's theorem, using SVR with a linear kernel during non-linear data analysis increases the classification/prediction accuracy [4]. In addition, this regressor will search for the optimal global solution with a limited number of support vectors, which results in low computational complexity [20].

Let us consider the main steps of the training procedure, which are visualized in Figure 1 in more detail:

1) We divide the dataset into parts with the same or almost the same number of observations;

2) At the first level of the cascade model, we perform a non-linear transformation of each vector of the first subsample and learn the first SVR with a linear kernel;

3) We expand the second subsample with the Ito decomposition and apply it to the pre-trained SVR_1 at the first level. We get the predicted values, which we add to the same sample as an additional attribute and send the modified dataset to the cascade model's second level. At the second level, we perform a non-linear transformation of the already extended by one attribute of the second subsample and train the second SVR_2 with a linear kernel;

4) The third subsample is expanded by the Ito decomposition and applied to the pre-trained SVR_1 of the first cascade level. We get the predicted output, which we add as an additional feature to this subsample and pass to the second level of the cascade model. At the second level, we perform a non-linear transformation of the third subsample extended by one attribute and apply it to SVR_2. We get the predicted output, which we add as an additional attribute to the third subsample and pass to the third level of the cascade. At the third level, we perform a non-linear transformation of the third subsample extended by one attribute and train SVR_3.

5) We use the same logic to train all subsequent cascade levels using the following subsamples that remained after the first step of the training algorithm.

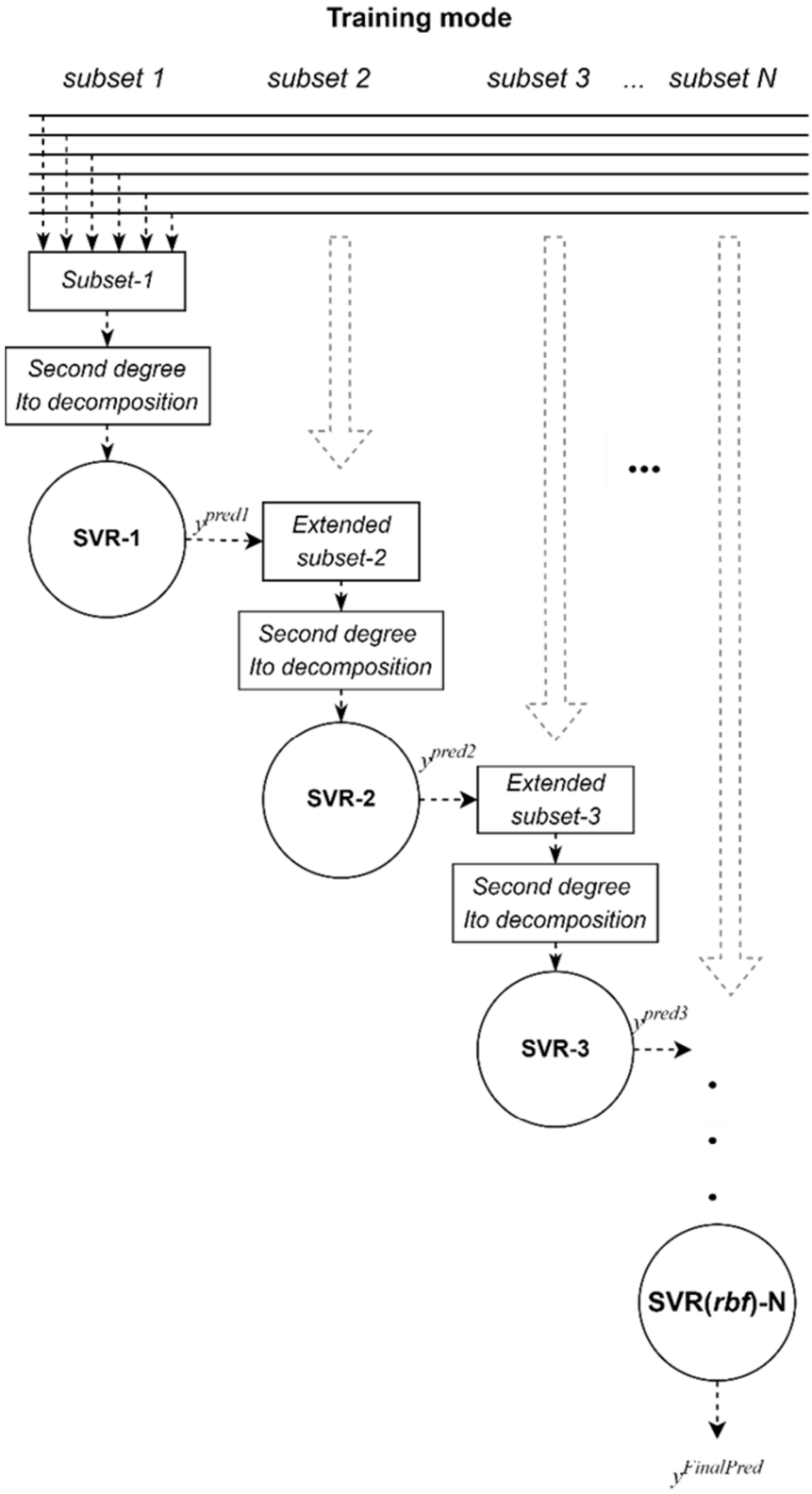

Figure 1. Flow chart for the training process of the proposed non-linear SVR-based cascade model.

In the application mode, one test dataset (or one data vector) is specified, and the number of pre-trained levels of the cascade model is determined. We apply the test dataset to the SVR-based cascade

model by performing all the procedures for the training algorithm (for the last level of the cascade).

For example, for the SVR-based cascade model, which will contain three cascade levels, the application procedures will be as follows: we expand the test subsample of the data with the Ito decomposition and apply it to the pre-trained SVR_1 of the first level of the cascade. We get the predicted output, which we add as an additional feature to this sample and pass to the second level of the cascade model. At the second level, we perform a non-linear transformation of the test sample extended by one attribute and apply it to SVR_2. We get the predicted output, which we add as an additional attribute and pass to the third level of the cascade. At the third level, we perform a non-linear transformation of the extended by one feature of the third subsample and apply it to SVR_3. We get the predicted value of the sought value.

4. Modeling and results

This section describes the modeling process on real-world data and obtained results. We used such performance indicators for evaluation the accuracy of the proposed SVR-based cascade scheme:

$$\text{Mean absolute error: } MAE = \frac{1}{N}\sum_{i}^{N}\left|y_i^{true} - y_i^{pred}\right|, \tag{2}$$

$$\text{Mean square error: } MSE = \frac{1}{N}\sum_{i}^{N}\left(y_i^{true} - y_i^{pred}\right)^2, \tag{3}$$

$$\text{Maximum residual error: } ME = \max\left(\left|y_i^{true} - y_i^{pred}\right|\right). \tag{4}$$

4.1. Dataset description

This paper investigated the heart rate prediction task based on the dataset of large volumes. The biomedical dataset for solving the stated task was taken from an open repository [21]. It contains 18 attributes derived from the signals that were measured from the ECG (electrocardiography). These, selected by the author of the dataset 18 features, have a significant but different effect on each individual's heart rate at a certain moment (Table 1). The dataset consists of 369,289 unique observations.

The dataset is pre-cleaned and ready for use, so it does not contain missing or abnormal values. It ensures the avoidance of many preliminary processing procedures and the possibility of directly performing the experimental part of the research on building a prediction model of increased accuracy by ML tools. The researched dataset was divided into training and test samples in a ratio of 70% to 30% to implement modeling procedures. Thus, the training sample contained 258,503 observations, and the test sample included 110,787 observations.

The practical value of solving this task can provide the possibility of determining the level of human stress [22]. That is why constructing a highly accurate heart rate prediction model will allow highly precise identification of a person's stress level, which is critical in various application areas [23–25].

Table 1. Training and test errors for different levels of the SVR-based cascade model.

Attribute Title	Min Value	Std	Mean Value	Max Value
Mean of RR intervals (MEAN_RR)	547.595	124.485	845.914	1322.01
Median of RR intervals (MEDIANR_R)	517.51	132.003	841.156	1653.12
Standard deviation of RR intervals (SDRR)	27.2406	76.8158	109.26	563.48
Root mean square of successive RR interval differences (RMSSD)	5.53346	4.12688	14.9808	26.6232
Standard deviation of successive RR interval differences (SDRR)	5.53336	4.12688	14.9801	26.623
Ratio of SDRR/RMSSD	2.66038	5.12581	7.38995	54.3399
Percentage of successive RR intervals that differ by more than 25 ms (pNN25)	0	8.20845	9.84384	39.4
Percentage of successive RR intervals that differ by more than 50 ms (pNN50)	0	0.9921	0.86997	5.4
Kurtosis of distribution of successive RR intervals (KURT)	−1.8947	1.78593	0.52599	62.6724
Skew of distribution of successive RR intervals (SKEW)	−2.1363	0.69987	0.044	6.56471
Mean of relative RR intervals (MEAN_REL_RR)	−0.0012	0.00016	−0.001	0.00123
Median of relative RR intervals (MEDIAN_REL_RR)	−0.0044	0.00087	−0.0005	0.0021
Standard deviation of relative RR intervals (SDRR_REL_RR)	0.00899	0.00547	0.01859	0.03654
Root mean square of successive relative RR interval differences (RMSSD_REL_RR)	0.00322	0.00392	0.00972	0.02695
Standard deviation of successive relative RR interval differences (SDSD_REL_RR)	0.00322	0.00392	0.00972	0.02695
Ratio of SDRR/RMSSD for relative RR interval differences (SDRR_RMSSD_REL_RR)	1.18126	0.37551	2.005	3.70231
Kurtosis of distribution of relative RR intervals (KURT_REL_RR)	−1.8947	1.78593	0.52599	62.6724
Skew of distribution of relative RR intervals (SKEW_REL_RR)	−2.1363	0.69987	0.044	6.56471
Heart rate of the patient at the time of data recorded (HR)	48.7372	10.3811	74.0103	113.727

4.2. Scalers

As studied in [26], data normalization significantly affects the accuracy and speed of ML methods when processing large volumes of medical and biomedical data [27]. Accordingly, the selection of the correct normalization method will allow increasing the efficiency of the work of the SVR-based

cascade model developed in this paper. When applying SVR to the entire dataset, the authors investigated three well-known normalization methods (MAX-ABS scaler, MIN-MAX scaler, and Standard scaler).

The obtained results for the training and application modes are shown in Figure 2.

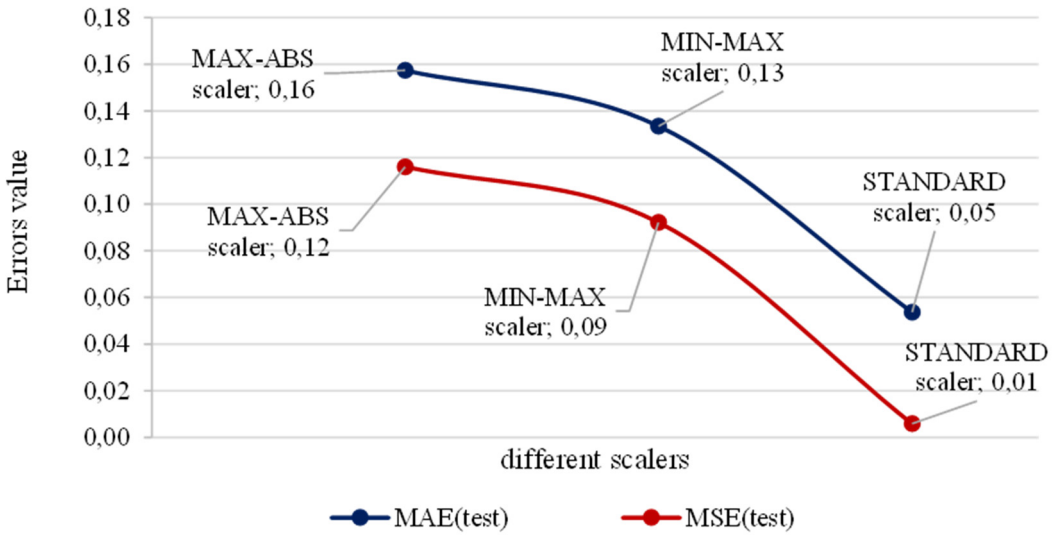

Figure 2. MAE and MSE values for training SVR on whole data using different scalers.

As shown in Figure 2, the MAX-ABS scaler demonstrated the most significant errors (MAE and MSE) in the SVR operation. Somewhat better results were obtained when using the MIN-MAX scaler. The best results in prediction accuracy were obtained when using the STANDARD scaler (data normalization by Standard scaler means subtracting from each vector components the mean value of current feature and dividing the result using standard deviation). This normalization method reduced both accuracy errors of the SVR application mode. In particular, the scaler used by MSE will provide a 12 times more minor error in the application of the method compared to the MAX-ABS scaler. For MAE, the error is more than three times smaller. That is why, during the practical implementation of the developed SVR-based cascade model, we used the STANDARD scaler to normalize data at each cascade level.

4.3. Cascade levels

The developed SVR-based cascade model is based on the ensemble of ML methods approach, namely cascading [28,29]. In addition, the principle of response surface linearization is used here [30]. In theory, all this will increase accuracy with each new level of the cascade [17]. However, the process of growing accuracy will occur until a certain point. The possible accumulation of errors from the previous levels of the cascade will affect the accuracy of the prediction at its following levels, and it will decrease. Therefore, an essential parameter for the practical application of the developed model is the number of its cascades. Determining the optimal value of the number of cascades of the SVR-based model will not only ensure obtaining the sought-after highest accuracy of solving the stated problem but also the possibility of using a model with lower computational complexity and with less training

time (in particular, by using a smaller number of cascades)

That is why, experimental studies were conducted to select the optimal value of this parameter. We built the SVR-based cascade model with the number of cascades from 1 to 6. The SVR parameters for all cascades were the same (kernel = 'linear', gamma = 'scale', coef0 = 0.0, tol = 0.001, C = 1.0, epsilon = 0.1, max_iter = −1). The results of this experiment for both training and application modes using MSE and MAE are shown in Table 2.

Table 2. Training and test errors for different levels of the SVR-based cascade model.

SVR-based cascade level	MAE*	MSE*	MaxE*
Training mode			
2	0,03668	0,00195	0,13377
3	**0,03413**	**0,00181**	**0,16356**
4	0,03522	0,00193	0,74910
5	0,03836	0,00218	0,41678
6	0,03853	0,00227	0,51560
Test mode			
2	0,0368	0,00202	1,06643
3	**0,03449**	**0,00187**	**0,78698**
4	0,03543	0,00195	0,88296
5	0,03828	0,00225	1,94576
6	0,03851	0,00238	0,82189

*MAE = Mean Absolute Error, MSE = Mean Squared Error, MaxE = Maximum Residual Error

Table 2 clearly shows that the both errors (training and test) SVR-based cascade model with two cascade levels significantly decreased in comparison with processing the whole dataset. It is explained by the principle of response linearization and the high approximation properties of the Ito decomposition. The best results were obtained when building a model with three cascade levels. In particular, the MSE error has decreased by more than 30 percent compared to the first level of the studied model. It should be noted that when using a four-level cascade, the model does not demonstrate overfitting. Testing errors are higher than training errors. However, this difference is insignificant, showing the developed model's high generalization properties. In addition, constructing a three-level cascade is significantly more efficient regarding computational complexity and training time than a model with a larger number of cascades.

It is also clear from Table 1 that the studied model, when using four or more cascade levels shows an increase in error in both modes. Moreover, starting from level 5, overfitting of the model is observed. Two reasons explain this. First, the rise in input attribute number is due to considering the output signals of all previous levels and the Ito decomposition. Secondly, the accumulation of errors at the earlier cascade levels and their impact on the subsequent levels of the developed model. That is why the optimal number of cascades of the developed model when solving the stated task is 3.

5. Comparison and discussion

To evaluate the effectiveness of the developed SVR-based cascade model, we compared its work with several linear, non-linear, and ensemble methods used to solve the stated problem.

Linear models are chosen because they are pretty fast when processing large datasets, non-linear because they provide increased accuracy compared to linear ones, and boosting ensemble methods are very similar to the developed cascade model. Performance indicators (MSE and MAE) and the duration of the training procedure were selected as efficiency criteria.

Therefore, the comparison was made with the following methods:

1) Linear Regression [7];
2) Ridge Regression [8];
3) Automatic Relevance Determination (ARD) Regression [8];
4) Bayesian Ridge Regression [9];
5) SVR with different kernels [10]
6) Ada Boost Regressor [11];
7) Gradient Boosting Regressor [12].

The results of applying all the methods studied in this paper are summarized in Table 3. It should be noted that Table 2 shows the MSE and MAE errors of the methods in the application mode.

When analyzing Table 3, it can be seen that all linear models provide almost the same level of prediction accuracy based on both performance indicators. Interesting results were obtained for SVR with non-linear kernels. Non-linear SVR with RBF kernel shows 16 times higher accuracy than SVR with linear kernel. Despite this, non-linear SVR with a polynomial kernel shows the lowest prediction accuracy of all the studied methods. It can be explained by the significant increase in the number of attributes of a given large dataset due to the non-linear expansion of the inputs. As a result, this model has significantly lower generalization and approximation properties.

Table 3. Comparison with different ML-based methods.

Method	MAE*	MSE*
SVR (linear)	1,065	3,007
SVR (rbf)	0,063	0,041
SVR (poly)	1,652	5,928
Proposed SVR-based cascade model	0,035	0,002
GradientBoostingRegressor	0,343	0,264
AdaBoostRegressor	1,585	3,359
ARDRegression	1,121	2,677
BayesianRidgeRegression	1,121	2,677
LinearRegression	1,121	2,676
RidgeRegression	1,122	2,678

*MAE = Mean Absolute Error, MSE = Mean Squared Error.

If we consider boosting ensembles, the results here are also twofold. Despite selecting optimal operating parameters, AdaBoost demonstrates significantly lower accuracy than linear models. At the same time, Gradient boosting makes it possible to obtain an acceptable predicted result, particularly

regarding the accuracy of the approximation of the current dataset.

The highest prediction accuracy was obtained using the developed cascade model (that are baseon Ito decomposition and SVR with linear kernel). If we analyze the MSE errors from Table 2, the developed SVR-based cascade model provides more than 3000 times minor errors than SVR with the polynomial kernel; more than 1500 times minor errors than SVR with the linear kernel; and more than 20 times minor error in comparison with SRV with the RBF kernel. It is explained both by the high approximation properties of Ito decomposition, which is used at all cascade levels, and by the principles of response surface linearization, which are the basis of the developed model.

The main disadvantage of methods based on cascading is the lack of parallel data processing, as, for example, when using the stacking strategy of assembling ML methods. That is why, in this paper, we also compared the duration of the training procedure of all studied techniques. The experiments was conducted on the PC with following characteristics: processor Intel Core i5-8250U CPU @ 1.60 GHz, RAM: 16 GB DDR4 2400 MHz. The results are summarized in Figure 3.

As expected (Figure 3), linear models are the fastest ones. Boosting methods demonstrate a significant increase in accuracy compared to linear algorithms. But some of the latter, in particular Gradient boosting, shows a substantial increase in prediction accuracy compared to linear models. The developed SVR-based cascade model requires five times more time than Gradient boosting (however, it provides 135 times higher MSE-based accuracy than this method). It considers the non-linear expansion of inputs due to Ito decomposition, which significantly increases the duration of the developed model's training procedure and is not used during modeling by the Gradient boosting algorithm.

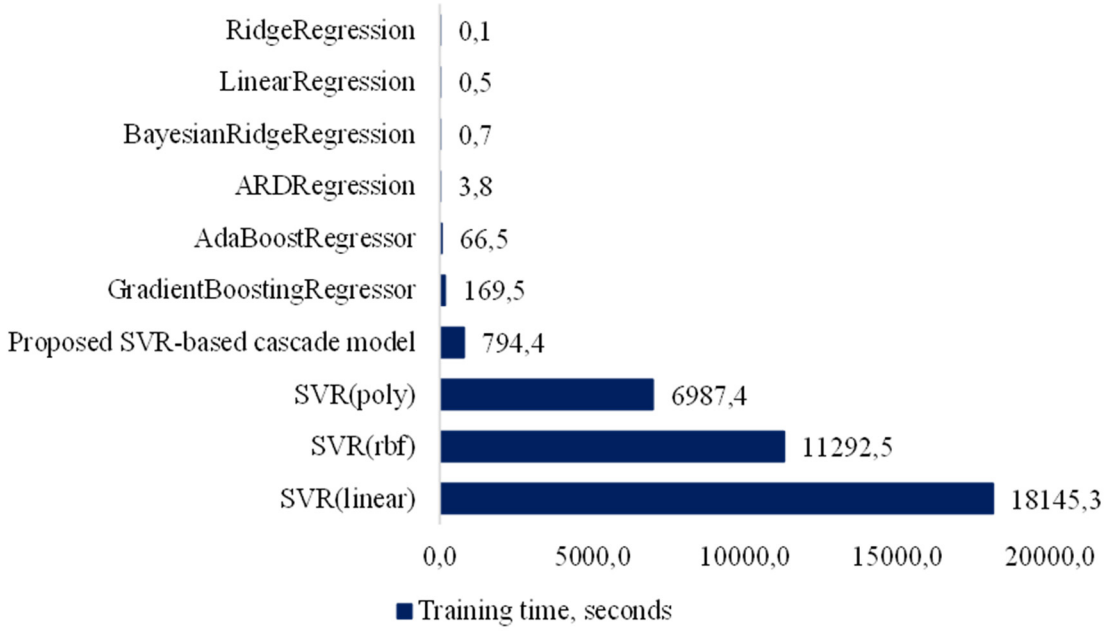

Figure 3. The training time for all investigated methods.

SVRs with different kernels show the highest training time. It is explained by the fact that these methods processed the entire large dataset, while the developed model worked with its parts separately. Despite this fact, in the perspective of further research, one should focus on reducing the duration of

the developed model's training procedure while maintaining its operation's accuracy. In particular, it is worth using a neural network variant of Principal Component Analysis (PCA) at each level of the cascade model to reduce the dimensionality of the input data space and, as a result, to reduce the duration of the training procedure [31]. Among other options, one should consider using non-iterative artificial neural networks [32] instead of SVR as weak elements for each level of the developed cascade model.

In addition, the proposed method can be used to analyse large volumes of data in other fields [33–35]. Moreover, it can be adapted to solve the classification task.

6. Conclusions

This paper is devoted to solving the problem of increasing the accuracy of approximation of biomedical datasets of large volumes. The authors developed a new SVR-based cascade model. It is based on the principles of cascading ML methods. Each subsequent level of the cascade of the developed model takes into account (as an additional feature) the output of the previous one. In addition, at each level of the cascade, non-linear expansion of the inputs due to Ito decomposition is implemented. This provides increased prediction accuracy due to the principle of response surface linearization. The authors used SVR with the linear kernel as the weak regressors.

The paper describes the training and application algorithms of the developed model. A flow chart of its implementation is also provided.

The modeling was based on a real-world biomedical dataset. The authors solved the task of heart rate prediction of individuals based on more than 350,000 observations. We found the high efficiency of proposed cascade scheme through a comparison with existing methods based on different performance indicators.

The shortcoming of the proposed approach is the impossibility of its parallelization due to the use of cascading principles. It determines the need for considerable time and resources to implement the training procedure. That is why, in the prospect of further research, it is planned to use non-iterative artificial neural networks as fundamental elements for each level of the developed cascade model. It will also be considered the possibility of using the neural network variant of PCA to reduce the space of input data of each level of the developed cascade model while maintaining the high accuracy of its operation.

Acknowledgments

The authors thank the reviewers for the relevant comments that helped to present the paper better. The National Research Foundation of Ukraine funds this study from the state budget of Ukraine within the project No. 30/0103. Michal Kovac was supported by the APVV grant No. 21-0448 and received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie COFUND SASPro2 grant agreement No. 2207/02/01.

Conflict of interest

Ivan Izonin is an editorial board member for Mathematical Biosciences and Engineering and was not involved in the editorial review or the decision to publish this article. All authors declare that there

are no competing interests.

References

1. N. Melnykova, N. Shakhovska, M. G. ml, V. Melnykov, Using big data for formalization the patient's personalized data, *Proc. Comput. Sci.*, **155** (2019), 624–629. https://doi.org/10.1016/j.procs.2019.08.088
2. K. Kakhi, R. Alizadehsani, H. M. D. Kabir, A. Khosravi, S. Nahavandi, U. R. Acharya, The internet of medical things and artificial intelligence: trends, challenges, and opportunities, *Biocybern. Biomed. Eng*., **42** (2022), 749–771. https://doi.org/10.1016/j.bbe.2022.05.008
3. I. H. Sarker, Machine learning: Algorithms, real-world applications and research directions, *SN Comput. Sci.*, **2** (2021). https://doi.org/10.1007/s42979-021-00592-x
4. I. Izonin, A. Trostianchyn, Z. Duriagina, R. Tkachenko, T. Tepla, N. Lotoshynska, The combined use of the wiener polynomial and SVM for material classification task in medical implants production, *Int. J. Intell. Syst. Appl*., **10** (2018), 40–47. https://doi.org/10.5815/ijisa.2018.09.05
5. I. Krak, O. Barmak, E. Manziuk, A. Kulias, Data classification based on the features reduction and piecewise linear separation, in *International Conference on Intelligent Computing & Optimization*, (2020), 282–289. https://doi.org/10.1007/978-3-030-33585-4_28
6. G. Heitz, S. Gould, A. Saxena, D. Koller, Cascaded Classification Models: Combining Models for Holistic Scene Understanding, 2008. Available from: https://proceedings.neurips.cc/paper/2008/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html
7. S. Kim, H. Park, W. Jung, K. Lim, Predicting heart rate variability parameters in healthy korean adults: A preliminary study, *Inquiry*, **58** (2021). https://doi.org/10.1177/00469580211056201
8. E. E. Tripoliti, T. G. Papadopoulos, G. S. Karanasiou, K. K. Naka, D. I. Fotiadis, Heart failure: Diagnosis, severity estimation and prediction of adverse events through machine learning techniques, *Comput. Struct. Biotechnol. J.*, **15** (2017), 26–47. https://doi.org/10.1016/j.csbj.2016.11.001
9. L. Fang, X. Liu, X. Su, J. Ye, S. Dobson, P. Hui, et al., Bayesian inference federated learning for heart rate prediction, in *International Conference on Wireless Mobile Communication and Healthcare*, **362** (2021), 116–130. https://doi.org/10.1007/978-3-030-70569-5_8
10. M. Oyeleye, T. Chen, S. Titarenko, G. Antoniou, A predictive analysis of heart rates using machine learning techniques, *Int. J. Environ. Res. Public Health*, **19** (2022), 2417. https://doi.org/10.3390/ijerph19042417
11. T. R. Mahesh, V. D. Kumar, V. V. Kumar, J. Asghar, O. Geman, G. Arulkumaran, et al., Adaboost ensemble methods using k-fold cross validation for survivability with the early detection of heart disease, *Comput. Intell. Neurosci*., **2022** (2022), 9005278. https://doi.org/10.1155/2022/9005278
12. P. Theerthagiri, Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique, *Intell. Syst. Appl.*, **16** (2022), 200121. https://doi.org/10.1016/j.iswa.2022.200121
13. S. Manimurugan, S. Almutairi, M. M. Aborokbah, C. Narmatha, S. Ganesan, N. Chilamkurti, et al., Two-stage classification model for the prediction of heart disease using iomt and artificial intelligence, *Sensors,* **22** (2022), 476. https://doi.org/10.3390/s22020476

14. R. Tkachenko, I. Izonin, I. Dronyuk, M. Logoyda, P. Tkachenko, Recovery of missing sensor data with grnn-based cascade scheme, *Int. J. Sens. Wireless Commun. Control*, **11** (2021), 531–541. https://doi.org/10.2174/2210327910999200813151904
15. I. Izonin, R. Tkachenko, R. Holoven, M. Shavarskyi, S. Bukin, I. Shevchuk, Multistage SVR-RBF-based model for heart rate prediction of individuals, in *International Conference of Artificial Intelligence, Medical Engineering, Education*, **159** (2023), 211–220. https://doi.org/10.1007/978-3-031-24468-1_19
16. J. Hsia, C. Lin, Parameter selection for linear Support Vector Regression, *IEEE Trans. Neural Networks Learn. Syst*., **31** (2020), 5639–5644. https://doi.org/10.1109/TNNLS.2020.2967637
17. I. Izonin, R. Tkachenko, An approach towards the response surface linearization via ANN-based cascade scheme for regression modeling in Healthcare, *Proc. Comput. Sci.*, **198** (2022), 724–729. https://doi.org/10.1016/j.procs.2021.12.313
18. A. G. Ivakhnenko, Polynomial theory of complex systems, *IEEE Trans. Syst. Man Cybern.*, **4** (1971), 364–378. https://doi.org/10.1109/TSMC.1971.4308320
19. V. Kotsovsky, A. Batyuk, On-line relaxation versus off-line spectral algorithm in the learning of polynomial neural units, in *International Conference on Data Stream Mining and Processing*, (2020), 3–21. https://doi.org/10.1007/978-3-030-61656-4_1
20. Y. B. Youssef, M. Afif, R. Ksantini, S. Tabbane, A novel QoE model based on boosting Support Vector Regression, in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, (2018), 1–6. https://doi.org/10.1109/WCNC.2018.8377092
21. V. Shanawad, Heart Rate Prediction to Monitor Stress Level, 2023. Available from: https://www.kaggle.com/datasets/vinayakshanawad/heart-rate-prediction-to-monitor-stress-level
22. L. Mochurad, Y. Hladun, Modeling of psychomotor reactions of a person based on modification of the tapping test, *Int. J. Comput.*, **20** (2021), 1–10. https://doi.org/10.47839/ijc.20.2.2166
23. G. Shanmugasundaram, S. Yazhini, E. Hemapratha, S. Nithya, A comprehensive review on stress detection techniques, in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, (2019), 1–6. https://doi.org/10.1109/ICSCAN.2019.8878795
24. Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, C. Ersoy, Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches, *IEEE Access*, **8** (2020), 38146–38163. https://doi.org/10.1109/ACCESS.2020.2975351
25. A. Hasanbasic, M. Spahic, D. Bosnjic, H. H. adzic, V. Mesic, O. Jahic, Recognition of stress levels among students with wearable sensors, in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, (2019), 1–4. https://doi.org/10.1109/INFOTEH.2019.8717754
26. I. Izonin, B. Ilchyshyn, R. Tkachenko, M. Greguš, N. Shakhovska, C. Strauss, Towards data normalization task for the efficient mining of medical data, in *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, (2022), 1–5. https://doi.org/10.1109/ACIT54803.2022.9913112
27. V. Shymanskyi, Y. Sokolovskyy, Finite element calculation of the linear elasticity problem for biomaterials with fractal structure, *Open Bioinf. J.*, **14** (2021), 114–122. https://doi.org/10.2174/1875036202114010011

28. N. García-Pedrajas, D. Ortiz-Boyer, R. del Castillo-Gomariz, C. Hervás-Martínez, Cascade ensembles, in *International Work-Conference on Artificial Neural Networks*, (2005), 598–603. https://doi.org/10.1007/11494669_73
29. Y. V. Bodyanskiy, O. K. Tyshchenko, A hybrid cascade neural network with ensembles of extended neo-fuzzy neurons and its deep learning, in *Conference on Information Technology, Systems Research and Computational Physics*, **945** (2018), 164–174. https://doi.org/10.1007/978-3-030-18058-4_13
30. A. G. Ivakhnenko, Development of models of optimal complexity using self-organization theory, *Int. J. Comput. Inf. Sci.*, **8** (1979), 111–127. https://doi.org/10.1007/BF00989666
31. J. Zhou, Y. Ye, J. Jiang, Kernel principal components based cascade forest towards disease identification with human microbiota, *BMC Med. Inform. Decis. Mak.*, **21** (2021), 360. https://doi.org/10.1186/s12911-021-01705-5
32. I. Tsmots, O. Skorokhoda, Methods and VLSI-structures for neural element implementation, in *2010 Proceedings of VIth International Conference on Perspective Technologies and Methods in MEMS Design*, (2010), 135–135.
33. I. G. Kryvonos, I.V. Krak, O. V. Barmak, A. S. Ternov, V. O. Kuznetsov, Information technology for the analysis of mimic expressions of human emotional states, *Cybern. Syst. Anal.*, **51** (2015), 25–33. https://doi.org/10.1007/s10559-015-9693-1
34. V. Babenko, A. Panchyshyn, L. Zomchak, M. Nehrey, Z. Artym-Drohomyretska, T. Lahotskyi, Classical machine learning methods in economics research: Macro and micro level examples, *WSEAS Trans. Bus. Econ.*, **18** (2021), 209–217. https://doi.org/10.37394/23207.2021.18.22
35. D. Chumachenko, T. Chumachenko, I. Meniailov, P. Pyrohov, I. Kuzin, R. Rodyna, On-line data processing, simulation and forecasting of the coronavirus disease (COVID-19) propagation in ukraine based on machine learning approach, in *International Conference on Data Stream Mining and Processing*, **1158** (2020), 372–382. https://doi.org/10.1007/978-3-030-61656-4_25

AIMS Press