scientific reports

OPEN

Neural network ensemble model for prediction of erythrocyte sedimentation rate (ESR) using partial least squares regression

Jaejin Lee[1], Hyeonji Hong[1], Jae Min Song[2,3]✉ & Eunseop Yeom[1]✉

The erythrocyte sedimentation rate (ESR) is a non-specific blood test for determining inflammatory conditions. However, the long measurement time (60 min) to obtain ESR is an obstacle for a prompt evaluation. In this study, to reduce the measurement time of ESR, deep neural networks (DNNs) were applied to the sedimentation tendency of blood samples. DNNs using multilayer perceptron (MLP), long short-term memory (LSTM), and gated recurrent unit (GRU) were assessed and compared to determine a suitable length of time for the input sequence. To avoid overfitting, a stacking ensemble learning was adopted, which combines multiple models by using a meta model. Four meta models were compared: mean, median, least absolute shrinkage and selection operator, and partial least squares regression (PLSR) schemes. From the empirical results, LSTM and GRU models have better prediction than MLP over sequence lengths of 5 to 20 min. The decrease in $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ of GRU and LSTM was attenuated after a sequence length of 15 min, so the input sequence length is determined as 15 min. In terms of the meta model, the statistical comparison suggests that GRU combined with PLSR (GRU–PLSR) is the best case. Then, the GRU–PLSR was tested for prediction of ESR data obtained from periodontitis patients to check its applicability to a specific disease. The Bland–Altman plot shows acceptable agreement between measured and predicted ESR values. Based on the results, the GRU–PLSR can predict ESR with improved performance within 15 min and has potential applicability to ESR data with inflammatory and non-inflammatory conditions.

The erythrocyte sedimentation rate (ESR) is a simple and inexpensive blood test for providing general information about an inflammatory or acute response. As the gold standard of ESR measurements, the Westergren method measures the distance from the meniscus to the top of the column of sedimented erythrocytes separated from plasma in a vertical Westergren tube (inner diameter = 2.5 mm, length = 200 mm, and whole blood volume = 5 mL) after 1 h[1,2]. The ESR value represents the sedimentation speed in units of millimeters per hour.

Under low shear conditions, erythrocytes tend to form one-dimensional coin stacks (rouleaux) or three-dimensional aggregates in a reversible process. Erythrocyte aggregation (EA) is affected by two major factors: aggregating force induced by the macromolecules, and disaggregating forces including negative surface charge and shear force[3]. Although negative charges of erythrocytes repel each other, the presence of certain plasma proteins (fibrinogen, immunoglobulins, lipoproteins, and α-2 macroglobulin) and inflammation-related proteins (cytokines and chemokines) in plasma enhance the degree of EA. Elevated EA is considered to augment blood viscosity[4], which increases flow resistance and wall shear stress. In addition, the high extent of EA is somewhat correlated with cardiovascular diseases (CVDs)[4] and diabetes mellitus[5]. Considering Stokes' law, the sedimentation speed of a spherical particle in a medium is dependent on its size[6]. Therefore, the ESR value is closely associated with the degree of EA[7].

Although ESR with minor modifications is still considered to be a useful index of nonspecific disease, the Westergren method has some disadvantages, including a time-consuming procedure (60 min) and a repetitive cleaning process. In addition, the Westergren method provides only the final sedimentation speed (the ESR value) after 1 h. It is insufficient to describe the temporal variation of the sedimentation tendency. Hong et al. proposed the micro-vibrational erythrocyte sedimentation rate (MV-ESR) and demonstrated the possibility of

[1]School of Mechanical Engineering, Pusan National University, Busan, South Korea. [2]Department of Oral and Maxillofacial Surgery, School of Dentistry, Pusan National University, Yangsan, South Korea. [3]Dental and Life Science Institute, School of Dentistry, Pusan National University, Yangsan, South Korea. ✉email: songjm@pusan.ac.kr; esyeom@pusan.ac.kr

shortening the time of ESR measurement[8]. Shteinshnaider et al. demonstrated that ESR values at 60 min can be predicted by ESR values at 30 min by using simple linear regression[9]. Nishisako et al. applied a formula for calculating a normal maximum range of ESR with only age information of patients[10]. However, this formula was only applicable to Asian patients aged more than 65 years. Lapić et al. compared automated ESR methods (TEST1, Ves-Matic Cube 200) with the Westergren method[11]. Ves-Matic Cube 200 showed larger dispersion, so they concluded that automated ESR methods should be used carefully. To gain information about the temporal sedimentation tendency, a characteristic time representing the time required for sedimentation by a specific length was proposed[12].

Although ESR elevation typically results from inflammatory factor such as fibrinogen, some noninflammatory factors such as age, sex, and pregnancy can also cause elevation of the ESR, which might make it difficult to identify the patterns of ESR[13]. For analysis of complex biological systems, machine learning algorithms have been proposed to recognize, classify, and predict the patterns[14]. Thus, deep learning methods were used in this study considering their ability to learn the various patterns in ESR.

The sedimentation tendency of erythrocytes is time series data. For forecasting a time series, several deep neural networks (DNNs) have been proposed in several fields[15–19]. Multilayer perceptron (MLP) is a type of feedforward neural network that does not have cycles in the same layer. Unlike the MLP, long short-term memory (LSTM) and gated recurrent unit (GRU) have cycles within a hidden layer and filter the information through gate structure[15,16]. Kong et al. compared various benchmarks with LSTM, including the k-Nearest Neighbors' algorithm and extreme learning machine[17]. They found that LSTM outperformed the other rival algorithms in short-term forecasting for loads in individual residential households. AlDahoul et al. investigated the usage of ElasticNet linear regression, MLP, extreme gradient boosting, and LSTM for the prediction of suspended sediment in the Johor river in Malaysia[18]. The results showed that the LSTM is better than the other models. Che et al. developed a GRU-based model (GRU-D) that captures the informative missing patterns in time series analysis[19]. It was shown that GRU-D performed better than deep learning models built on GRU as well as other machine learning methods. Yilmaz and Buyrukoğlu proposed the hybrid model combining Back Propagation-Based Artificial Neural Network (BP-ANN), Correlated Additive Model (CAM) and Auto-Regressive Integrated Moving Average (ARIMA) to forecast Coronavirus disease (Covid-19) case[20]. They showed that the proposed hybrid model provides the best results, which can be used as the prediction model for Covid-19 to obtain better prediction results and to take immediate measures. Buyrukoğlu analyzed five promising cryptocurrencies with the ensembles of LSTM and single-based LSTM networks[21]. This study revealed both ensembles of LSTM and single-based LSTM can be used separately, considering that ensemble of LSTM does not always outperform the single-based LSTM.

When a small amount of data is available, a single neural network algorithm tends to exhibit generalization error for unseen instances[22]. Generalization means the capacity of the model to predict the correct output for previously unseen data[23]. Because small datasets are more prone to overfitting than large datasets, an ensemble approach was proposed to reduce the generalization error by avoiding overfitting[24]. Specifically, ensemble learning refers to the combination of multiple algorithms with different opinions and then weighting and combining them to make a complex decision. Ensemble learning is mainly used to improve prediction performance, or to avoid selection of a poor predictions by combining multiple models. In several studies, ensemble learning could improve the forecasting performance[25–27].

There are commonly used ensemble methods like bagging, boosting, and stacking. Bagging, also known as bootstrap aggregation, constructs multiple classifiers trained on bootstrap samples and averages the outputs of the predictors to obtain the final prediction with the appropriate size[28]. Since the bootstrap samples are randomly drawn with replacement from the entire training data, it causes individual bootstrap samples to overlap significantly, with many of the same instances in the training data appearing in most samples, and some instances appearing multiple times in a given sample. Unlike bagging, boosting creates different classifiers by sequentially reweighting the instances in the training data[29]. Each instance misclassified by the previous base classifier get larger weights, and the next classifier is boosted on the reweighted training data. In this way, a sequence of training data and classifiers is obtained, which can be combined by a simple majority voting or by weight majority voting in the final decision. In stacking, which stands for stacked generalization, a set of base models are constructed from the training data, then their outputs are used to train a meta model[30]. In most cases, the performance of an ensemble is better than a single model[21,31–34]. Ensemble learning has been widely used in various fields such as: health science[31], sport science[32], agriculture[33,34], finance[21].

In this study, the MLP, LSTM, and GRU models were compared to reduce the measurement time of ESR. Specifically, the sequence length (input data) is determined by monitoring predictive performance with evaluation metrics for each model. After determining the sequence length, stacking ensemble models with different ensemble sizes were investigated to select an adequate ensemble structure. Furthermore, all models including the individual DNNs were compared statistically. From the statistical comparison, the best performing meta model was determined. The proposed model was then applied to ESR prediction for patients with a particular disease. The aim of this study can be summarized as:

- To reduce measurement time of ESR, using ensemble deep learning models.
- To find the best performing ensemble model based on the statistical scores.
- To validate the applicability of the selected ensemble model for predicting ESR in nontrained samples (Patients with a particular disease).

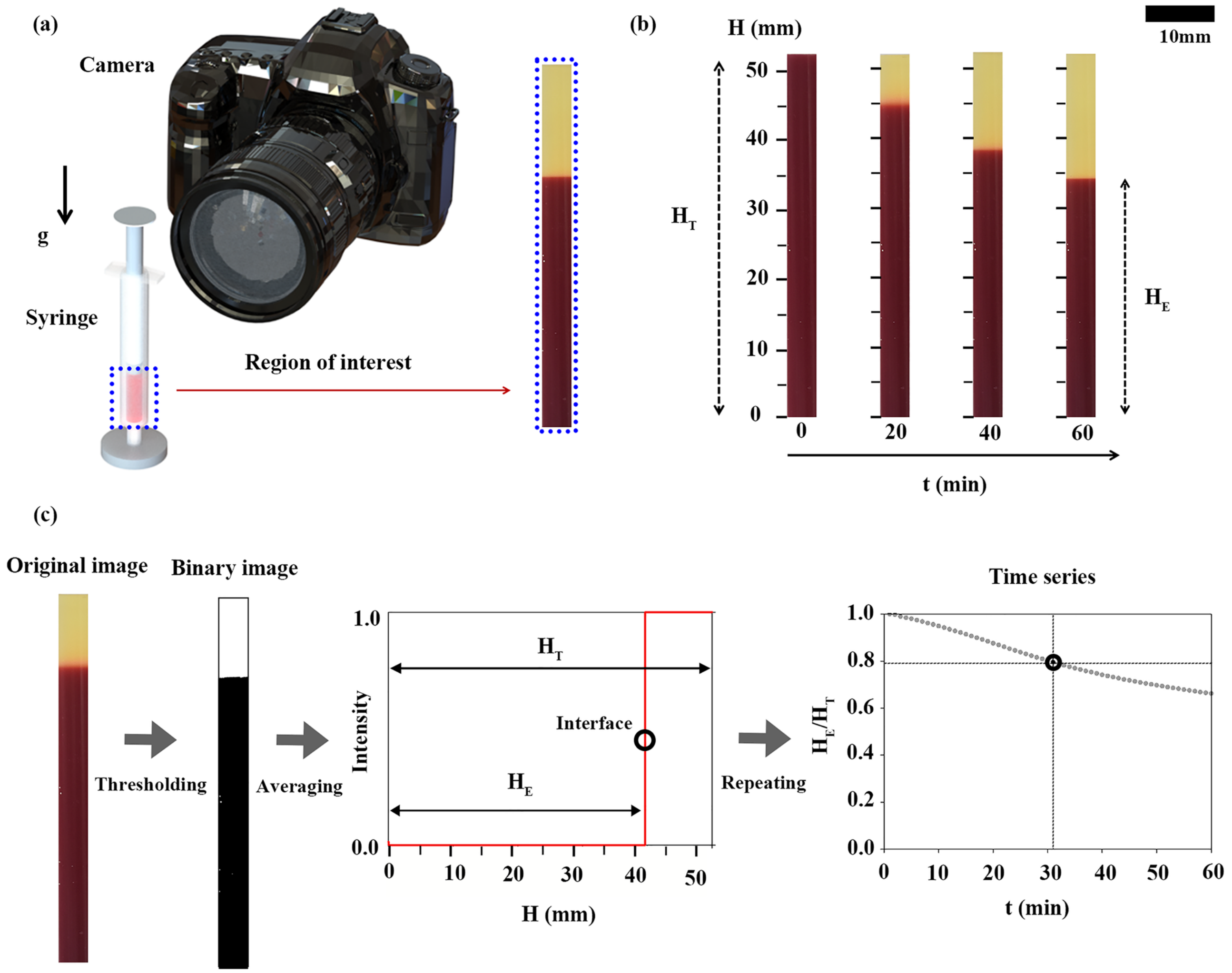

Figure 1. (**a**) Measurement system composed of a disposable syringe (1 mL) and digital camera. (**b**) Typical images of sedimented erythrocytes in a blood sample with respect to time (t). Height of the sedimented erythrocytes (H_E) is gradually decreased during 60 min. (**c**) Data processing procedure. The original image is transformed into a binary image using Otsu's method. The binary image is averaged in the horizontal direction to detect an interface (marked with a black circle) between sedimented erythrocytes and plasma. This interface, denoted by H_E/H_T, was extracted during 60 min (marked with a gray circle), and forms time series data.

Material and methods

Blood sample preparation.

A total of 310 subjects from the Pusan National University Dental Hospital (Busan, Korea) participated in this study. The study population included 304 normal subjects and 6 patients who have periodontal disease. The age of the subjects ranged from 5 to 95 years (48.369 ± 18.823). The percentages of males and females were 46% and 54%, respectively. Blood samples were collected via 3-mL BD Vacutainer® K2 EDTA tubes (Becton, Dickinson and Co., Franklin Lakes, NJ) for anticoagulation. A centrifuge (DSC-102SD, Nasco Korea) was used for separating erythrocytes from plasma. Separated erythrocytes were mixed with autologous plasma with the hematocrit (packed volume percentage of erythrocytes) fixed at 35% by mixing. Ethical permission for this study was granted by the institutional review board (IRB) of Pusan National University Hospital (H- 1911–028–085), and the study followed the rules of the Declaration of Helsinki.

Experimental setup.

Figure 1a shows the experimental setup for recording the images for erythrocyte sedimentation. A disposable syringe (1 mL syringe, BD, USA) filled with a 1 mL blood sample was mounted vertically. Consecutive images of the blood in the syringe were acquired by a digital camera (D5600, Nikon Imaging Korea Co., Ltd., Korea) with a micro-lens (AF Micro-Nikkor 105 mm f/2.8D, Nikon Co., Japan). Using software (Camera control pro 2, Nikon Co., Japan), images were taken automatically at intervals of 0.5 min for 60 min. The pixel size of the CMOS camera was 6000 × 4000 pixels. All experiments were conducted at room temperature (25 °C) with 50% humidity.

Image processing procedure. Figure 1b shows the process of erythrocyte sedimentation in a blood sample at various times (t = 0, 20, 40, and 60 min). The total height of the blood (H_T) was about 53 mm, corresponding to 1 mL. The height of the sedimented erythrocytes (H_E) decreased continuously due to gravity sedimentation with the lapse of time. Figure 1c shows the image processing procedure using commercial software (MATLAB, Mathworks, USA). The captured images were cropped to the region of interest in the syringe (120 × 2900 pixels) and then converted into gray images. The images were transformed into binary images using Otsu's method. By averaging the binary images in the horizontal direction, variation of the intensity along the gravity direction was obtained. The interface between sedimented erythrocytes and separated plasma was detected by utilizing the 'findchangepts' function in MATLAB, which detects a drastic change in signal. Based on the detected interface, H_E/H_T was determined at a given time. During ESR measurement (60 min), temporal variation of H_E/H_T for one sample was obtained by repeating the procedure, as illustrated in the right inset of Fig. 1c. For learning and training for the prediction of ESR, normalized time series data (H_E/H_T) were used as model input.

Deep learning methodology for ESR prediction. *MLP model.* The MLP is the most widely used algorithm for forecasting. It consists of multiple layers of interconnected elements called neurons with a nonlinear activation function (φ)[35]. Each neuron is only forward connected to every neuron in the subsequent layers. This structure is combined with the hidden layer activation function and allows the MLP to approximate any continuous function[36]. The mathematical model of the MLP is represented by:

$$y_i = \varphi\left(\sum_{i=1}^{H} w_{ij}x_i + b_j\right) \tag{1}$$

where y_i is the output variable, H represents the number of neurons in the hidden layer, w_{ij} is the weight connecting the jth neuron and the ith neuron, x_i is an input variable, and b_j is the bias of the jth neuron.

LSTM. A recurrent neural network (RNN) has a recurrent hidden state with an activation at each time that depends on the one from the previous time. RNN based algorithms have been developed in time series forecasting applications[37–39]. Due to infinite lookback windows, RNNs are vulnerable to long-range dependencies in the data, and the gradients tend to either vanish or explode[40]. LSTM is a variant of RNN that was developed to deal with the limitation of RNNs by using memory cells. A memory cell in each LSTM unit stores long-term information and these cells are modulated through three gates: an input gate (i_t), output gate (o_t), and forget gate (f_t). These gates adjust the hidden state (h_t) and cell states (c_t) of the LSTM as follows:

$$i_t = \sigma(W_{i1}X_t + W_{i2}h_{t-1} + b_i) \tag{2}$$

$$o_t = \sigma(W_{o1}X_t + W_{o2}h_{t-1} + b_o) \tag{3}$$

$$f_t = \sigma(W_{f1}X_t + W_{f2}h_{t-1} + b_f) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{c1}X_t + W_{c2}h_{t-1} + b_c) \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

where W and b are the weights and biases of the LSTM layer, respectively. σ is a sigmoid function, and tanh is a hyperbolic tangent function. ⊙ is the element-wise (Hadamard) product.

GRU. GRU was introduced by Cho et al.[16] and makes each recurrent unit with an adaptive selection of dependencies of different time scales. Unlike LSTM, GRU modulates the flow of information inside the unit without memory cells. GRU has two gates: an update gate (g_t) and reset gate (r_t).

$$g_t = \sigma(W_{z1}X_t + W_{z2}h_{t-1} + b_z) \tag{7}$$

$$r_t = \sigma(W_{r1}X_t + W_{r2}h_{t-1} + b_r) \tag{8}$$

$$\widetilde{h}_t = \tanh(W_{h1}X_t + W_{h2}(r_t \odot h_{t-1}) + b) \tag{9}$$

$$h_t = \left(1 - g_t\right) \odot h_{t-1} + g_t \odot \widetilde{h}_t \tag{10}$$

where $\widetilde{h}_t$ denotes candidate hidden state.

Selection of sequence length. To reduce ESR measurement time, the length of the input sequence of the prediction should be decreased. The starting time is fixed as 0 min. To select the optimal length of time, different time spans were tested for MLP, LSTM, and GRU. The input time ranged from 5 to 20 min with intervals of 2.5 min. When the input time is shorter than 5 min, the predicted results are extremely poor.

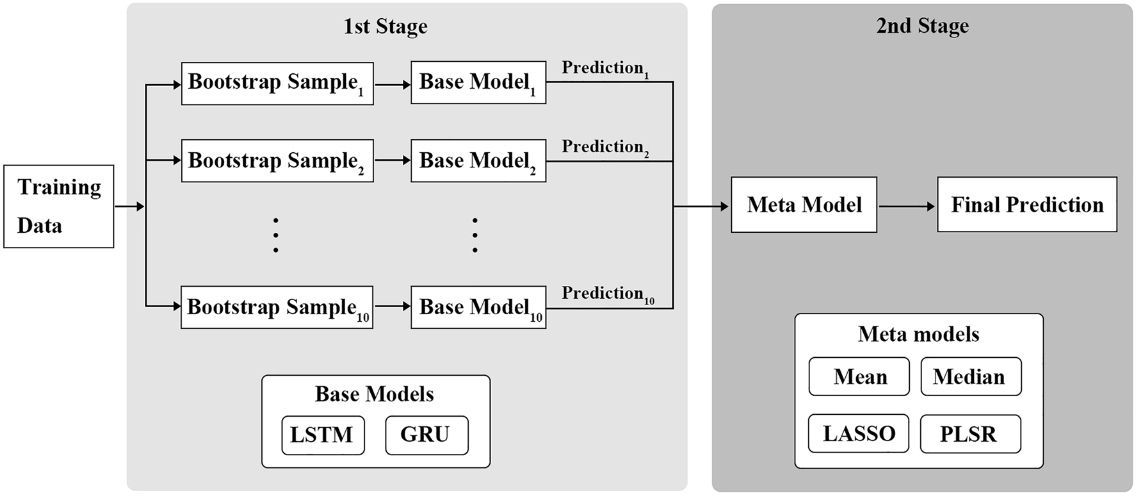

Figure 2. Structure of the ensemble model.

Ensemble model. In the present work, stacking ensemble using bootstrapped samples of the training data was proposed. As shown in Fig. 2, stacking is a two-stage approach. In the first stage, multiple base models were built on bootstrap samples, and corresponding predictions were obtained. Bootstrap samples were randomly drawn with replacement from the entire training data. In the second stage, meta model combines the predictions of the base models and produces the final prediction. The final prediction should be a weighted average of the individual predictions since some base models are more accurate than others. However, given that all base models attempt to solve the same task, their predictions are expected to be highly correlated, which can lead to performance degradation of the ensemble[24,28,41]. Feature selection technique is necessary to remove the redundant or irrelevant features or features correlated in the data. Some studies proposed deep autoencoder or deep belief network to extract the nonlinear features[42,43]. However, the predictions of the base models didn't show any complex patterns but high correlation in this study. Thus, LASSO and PLSR were used to cope with the multicollinearity due to the high degree of correlation between the individual predictions. The meta models are briefly introduced.

Mean ensemble. The mean is the simplest weight combination approach and assigns equal weights to all the individual forecasts[44]. The mean operator is intuitive, but it has a limitation that is widely known: the mean is susceptible to outliers and skewed distribution. The mean ensemble forecast (z_a) can be calculated as:

$$z_a = \frac{1}{M}\sum_{i=1}^{M} z_i \tag{11}$$

where M is the number of base models, and z_i ($i = 1, \ldots, M$) denotes the forecasted output vector of the ith base model.

Median ensemble. The median simply provides the middle value of z_i in a sorted list if M is odd or the mean of the two middle values otherwise[44]. The median is less affected by the influence of outliers. The median ensemble forecast (z_m) can be calculated as:

$$z_m = \begin{cases} z_{(M+1)/2}, & \textit{if M is odd} \\ (z_{M/2} + z_{M/2+1})/2, & \textit{if M is even} \end{cases} \tag{12}$$

Least absolute shrinkage and selection operator (LASSO). The LASSO is a regularized regression to deal with multicollinearity in data[45]. The LASSO penalty can force certain coefficients to be zero by imposing a penalty on their size, and it yields interpretable results with improved accuracy. The LASSO estimator ($\widehat{\beta}_L$) is the solution to the following optimization problem:

$$\widehat{\beta}_L = \underset{\beta}{\operatorname{argmin}} \|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 \tag{13}$$

where y represents the vector of observed ESR values, $Z = [z_1, z_2, \ldots, z_M]$ is the individual forecasts matrix, and β denotes the regression coefficient. λ is a positive regularization (penalty/tuning) parameter that controls the amount of shrinkage of β. As the value of λ depends on the data, it can be determined using data-driven methods such as cross-validation (CV). The intercept is ignored for convenience. The LASSO ensemble forecast (z_L) can be calculated as:

$$z_L = Z\widehat{\beta}_L \tag{14}$$

Partial least squares regression (PLSR). The basic idea of PLSR is to project the original variables onto a set of orthogonal latent variables. The latent variables, called components, are linear combinations of the original variables. The number of the latent components (J) is determined by CV, similar to the LASSO method. PLSR extracts these components by modeling both input and output data[46]. PLSR builds a relationship between the input matrix Z and the output vector y. The PLSR technique attempts to find a linear decomposition of Z and y by the outer relations[47]:

$$Z = TP^T + E_x = \sum_{j=1}^{J} t_j p_j^T + E_x \tag{15}$$

$$y = UQ^T + E_y = \sum_{j=1}^{J} u_j q_j^T + E_y \tag{16}$$

where T and U are score matrices, P and Q are loading matrices, and E_x and E_y are residual errors for input and output. p_j and q_j are loading vectors that can be considered as weights. The superscript "T" means transposition. Latent vectors t_j and u_j are obtained to maximize the covariance between Z and y while minimizing E_x and E_y. The inner relation between U and T is given by:

$$U = TB + E_U \tag{17}$$

where B denotes the coefficients, and E_U is an error term between U and T. If the error terms are minimized, the PLSR ensemble forecast (z_P) can be calculated as:

$$z_P = \sum_{j=1}^{J} u_j q_j^T \tag{18}$$

Deep neural network implementation. DNN models ware built using TensorFlow 2.8.0 and Keras packages in Python 3.7.12 in the Google Colaboratory environment. For training and evaluation of the model, time series data of 304 normal subjects were randomly split into training, validation, and test sets (194, 50 and 60 subjects, respectively). Models were trained for 300 epochs with an early stopping method. Early stopping was adopted that specifies an arbitrarily large number of epochs and stops training the model when there is no improvement in the validation loss for 15 epochs. Moreover, an exponential decay schedule was utilized, which decays the learning rate using an exponential function. The loss function was the mean squared error. The optimizer used in this study was the adaptive moment estimation (ADAM), which showed the highest performance among other optimizers including Adadelta, Adagrad, RMSprop, and SGD in a preliminary test.

The neural network architecture was determined empirically by monitoring the performance with different architectures. MLP consisted of two hidden layers: one with 15 neurons and one with 10 neurons. The architecture of the GRU consisted of one hidden layer with 15 neurons. The architecture of LSTM was identical to that of GRU. A dense layer with a single neuron was chosen for the output layer in all DNNs.

Performance metrics. The forecasting performance of each model was evaluated through some error metrics. The mean absolute percentage error (MAPE) and root mean squared error (RMSE) were adopted[48]. The formulas for MAPE and RMSE are as follows:

$$\text{MAPE} = \frac{1}{n_s}\sum_{i=1}^{n_s}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \times 100(\%) \tag{19}$$

$$\text{RMSE} = \sqrt{\frac{1}{n_s}\sum_{i=1}^{n_s}\left(y_i - \widehat{y}_i\right)^2} \tag{20}$$

where n_s and $\hat{y}$ denote the number of samples and the forecasted ESR, respectively.

Statistical test for all pairwise comparison. The Friedman test is a non-parametric test that detects significant differences among models[49]. It produces ranks of models' performance averaged over multiple datasets. The null hypothesis of the Friedman test states that all the models are equivalent and their mean ranks are nearly equal. If the null hypothesis is rejected with very small p-values, the Nemenyi post-hoc test can be done[50]. Although the Nemenyi test is known to be conservative, the procedure is the simplest among post-hoc tests for pairwise multiple comparisons[51]. The Nemenyi test was used to compare all models with each other. The performance between any two models is significantly different if their corresponding mean ranks differ by at least the critical distance (CD). The CD is given by:

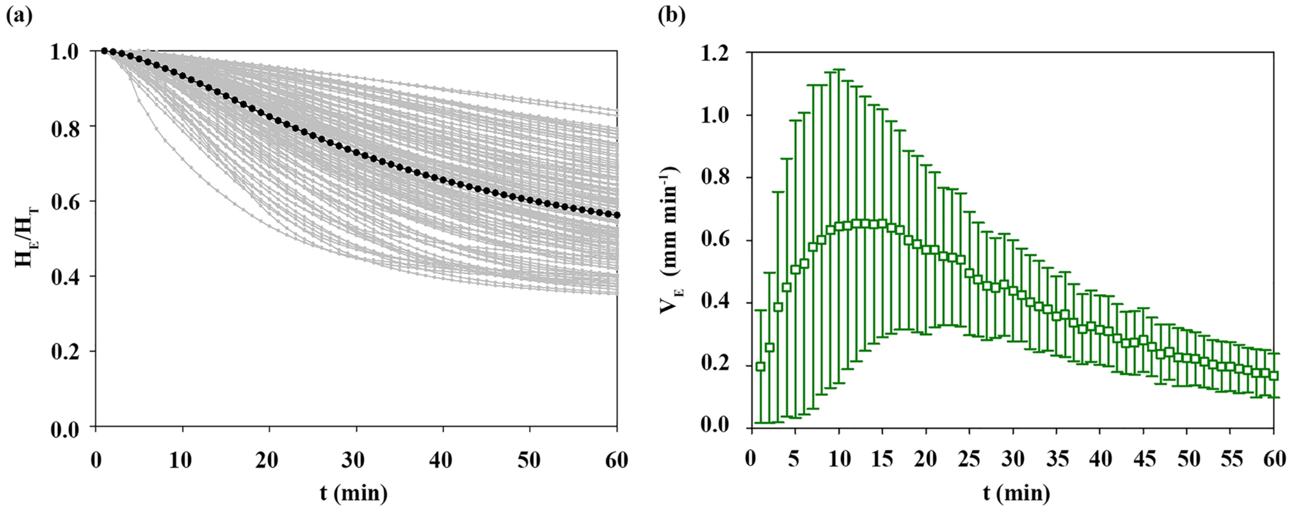

Figure 3. (**a**) Temporal variations of H_E/H_T for 304 normal subjects. Mean value of every H_E/H_T is marked by black dotted line. (**b**) Instantaneous time rate of sedimentation (V_E). Velocity data are represented as mean ± SD (n = 304).

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6D}} \tag{21}$$

where k and D are the numbers of models and datasets, respectively, and q_α is the critical value based on studentized range statistics divided by $\sqrt{2}$.

Results and discussion

Tendency of the erythrocyte sedimentation. Figure 3a shows the temporal variations of H_E/H_T for all subjects (n = 304) and mean trend of all samples (black dotted line). Since the hematocrit of blood samples was fixed at 35%, ESR values (H_E/H_T at 60 min) ranged from 0.350 to 0.859. Although some samples show similar ESR values, the variation tendency of H_E/H_T is somewhat different. To further investigate the variation tendency during the erythrocyte sedimentation process, the instantaneous time rate of sedimentation ($V_E = -dH_E/dt$) for 304 subjects is shown in Fig. 3b as the mean and standard deviation (SD). The extent of V_E can be approximately divided into two stages according to time. Within 13 min, V_E shows an acceleration stage. After that, it tends to decrease with the lapse of time. Since V_E can describe the extent of erythrocyte sedimentation for a specific time, it is closely related to the ESR sedimentation process.

The sedimentation process can be approximately described by 3 phases: aggregation, sedimentation, and packing[52,53]. During the aggregation phase, erythrocytes aggregate and form the rouleaux or aggregates. Since the sedimentation speed is proportional to the particle size[6], large rouleaux and aggregates produced in the aggregation stage promote sedimentation. As a result, erythrocytes rapidly fall in the sedimentation phase, whereas erythrocytes slowly fall and pile up at the bottom of the tube in the packing phase. It can be inferred that times with high V_E (10–20 min) may be included in the sedimentation phase (Fig. 3b). Thus, temporal variations of H_E/H_T at 0–20 min are important to predict the ESR value because they provide information on the erythrocyte sedimentation process.

Performance comparison of DNNs depending on input sequence lengths. Figure 4a,b show the variations of $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ according to the sequence length (input data at 0-t min). $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ are mean values of MAPE and RMSE obtained by repeating the prediction over 30 times. Due to the random setting of initial weights and biases, the 30 results are different. In general, prediction errors decrease with increasing sequence length for all DNNs. MLP shows relatively poor performance for all sequence lengths. This might result from the inherent structure of MLP, which cannot learn the sequential order of the time series input. As shown in Supplementary Table 1, SD values of MLP are usually higher than those of LSTM and GRU. This indicates that the prediction results of LSTM and GRU are less dependent on the initial setting than those of MLP. These results are well matched with a previous study that compared the prediction uncertainty of LSTM, GRU, and ANN[54]. Since similar accuracy was observed between LSTM and GRU, the optimal sequence length was determined based on LSTM and GRU results. $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ of both GRU and LSTM gradually decrease until the sequence length reaches 15 min and show a marginal decrease from 15 min onward.

Determination of the exact time input for the ESR is challenging since many factors also affect ESR, including sex, age, and even vertical installation angle of the specimen tubes. Previous studies reported that the ESR can be estimated by the ESR recorded at 30 min[9,55]. The inconsistent optimal time can be explained by the method to monitor erythrocyte sedimentation. In this experiment, relatively dense temporal variations of erythrocyte sedimentation (time interval of 0.5 min) were recorded. In contrast, previous reports measured the ESR values at 15, 20, 30, and 40 min. Considering the marginal decrease from 15 min with small values of $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$, the optimal sequence length was selected as 15 min.

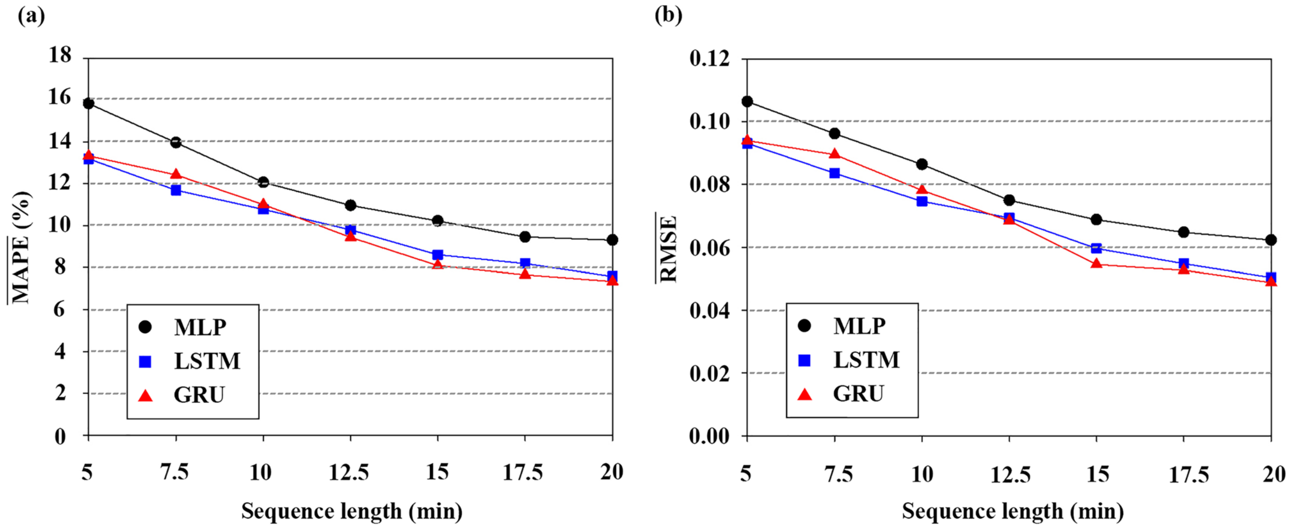

Figure 4. (**a**) $\overline{\text{MAPE}}$ (%) of the individual models according to sequence length (min). $\overline{\text{MAPE}}$ is the mean MAPE over 30 repeated experiments. (**b**) $\overline{\text{RMSE}}$ of the individual models according to sequence length (min). $\overline{\text{RMSE}}$ is the mean RMSE over 30 repeated experiments.

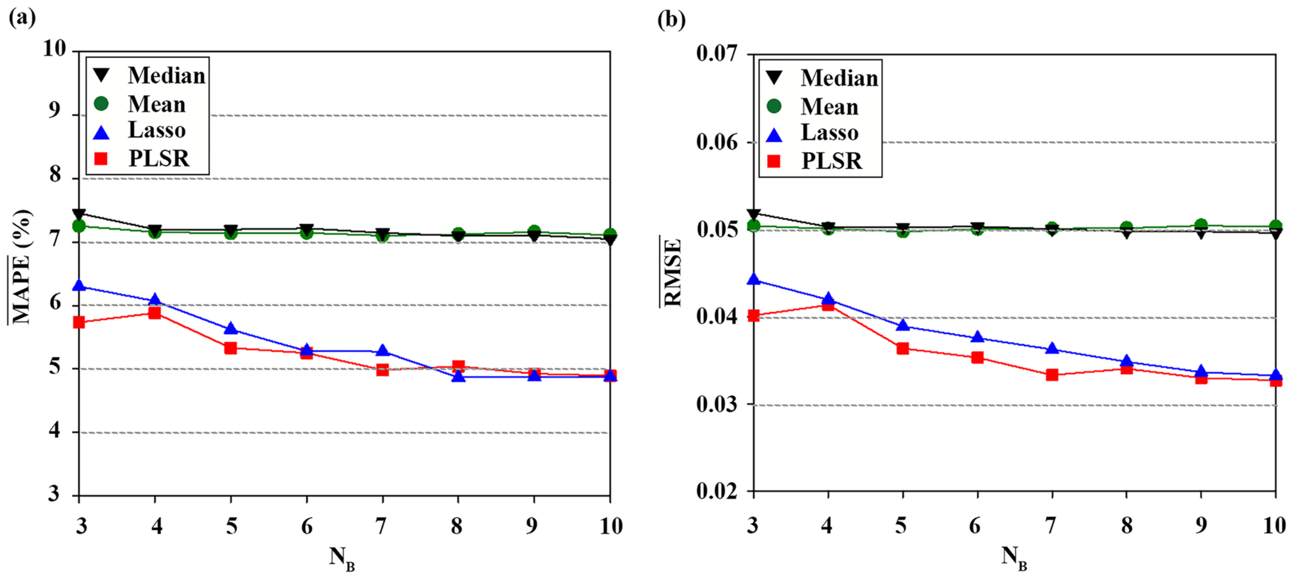

Figure 5. Forecasting accuracy of LSTM-based ensemble model according to number of base models (N_B): (**a**) $\overline{\text{MAPE}}$ (%) for different meta models, (**b**) $\overline{\text{RMSE}}$ for different meta models.

Performance of ensemble model according to ensemble size. As the number of base models (N_B) is an important parameter in ensemble performance, its impact was investigated in terms of $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$. Both LSTM and GRU were selected as base models for the ensemble. All meta models were implemented in MATLAB, where the parameters of LASSO and PLSR were determined by fivefold CV. All base models were trained on bootstrap samples and run 30 times for the fixed sequence length (15 min) of the test set. To select an adequate ensemble size, N_B was varied from 3 to 10. In a preliminary test, $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ of each ensemble model were almost saturated when N_B exceeded 10 DNNs.

Figures 5,6 show the variations of $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$ for LSTM and GRU coupled with four meta models. It can clearly be observed that the results of the mean are similar to those of the median. This marginal difference can be explained by the strong correlation between the individual forecasts. As the mean and median are measures of the central tendency, multicollinearity in the individual forecasts results in similar performance between them. However, LASSO and PLSR methods are more accurate than the mean and median over all N_B, which show superior ability to handle the multicollinearity.

Combinations of LSTM with PLSR and LASSO show similar performance, as illustrated in Fig. 5. However, compared to LASSO, PLSR shows a different trend for N_B ranging from 3 to 7. Particularly, PLSR exhibits a significantly low error when N_B increases from 4 to 5. Both methods show marginal improvement with the increase

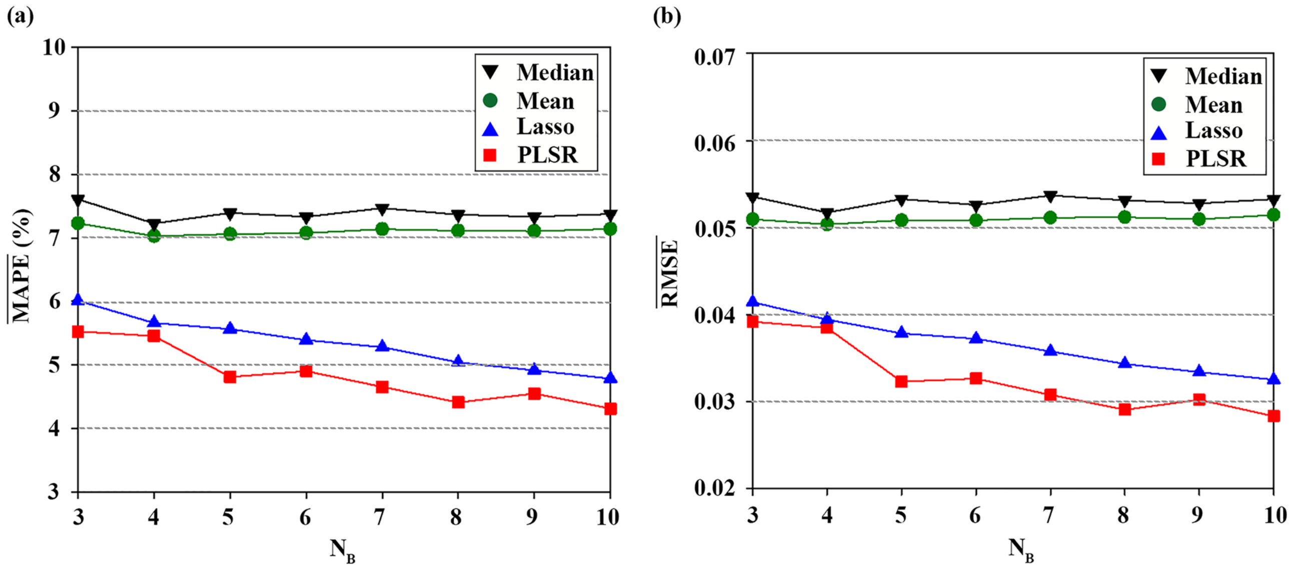

Figure 6. Forecasting accuracy of GRU-based ensemble model according to N_B: (**a**) $\overline{\text{MAPE}}$ (%) for different meta models, (**b**) $\overline{\text{RMSE}}$ for different meta models.

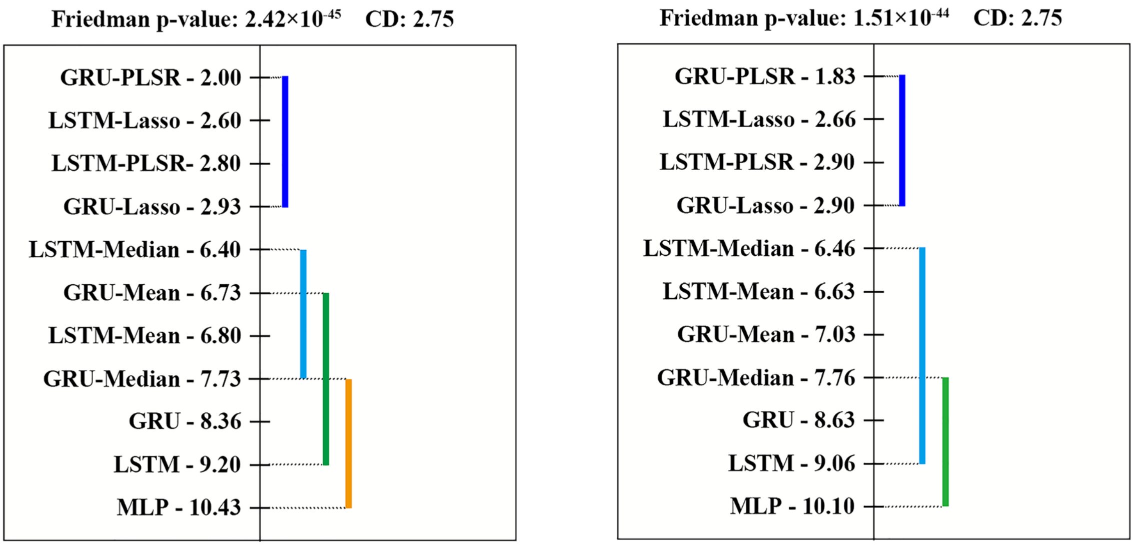

Figure 7. Nemenyi test results based on MAPE (left) and RMSE (right) for N_B of 8. The critical distance (CD) is 2.75.

in N_B ($N_B > 8$). In the case of the GRU base model, PLSR achieves better results, followed by LASSO (Fig. 6). The variations of PLSR fluctuate with increasing N_B, whereas the variations of LASSO gradually decrease. PLSR shows relatively high reduction of error when N_B varies from 4 to 5, which was similarly observed in LSTM results (Fig. 5).

Although PLSR shows the minimum error at N_B of 10, there is no significant difference at N_B ranging from 8 to 10 in terms of PLSR performance (Figs. 5,6). The performance degradation with the increase in N_B indicates that usage of all base models may not contribute to the improvement of the ensemble forecast[56,57]. The pattern learned by some base models has already been represented by other models due to serious correlation between the models when N_B is sufficiently high. Thus, the candidates for the optimal N_B are 8 and 10. In this study, the optimal N_B was determined as 8 because there was no significant difference between 8 and 10.

Statistical comparison of model performance. To examine significant differences among all models, statistical tests were carried out with the selected N_B of 8. Extremely low p-values were obtained by the Friedman test for both MAPE and RMSE at a significance level of 0.05, so the Nemenyi test was performed as illustrated in Fig. 7. P-values and CD are shown in the title, and the mean rank is marked next to each model. The ensemble models are named based on their base models and the meta models, such as LSTM-Mean or GRU-Median. As the base models coupled with meta models show better performance (low mean ranks), the models are posi-

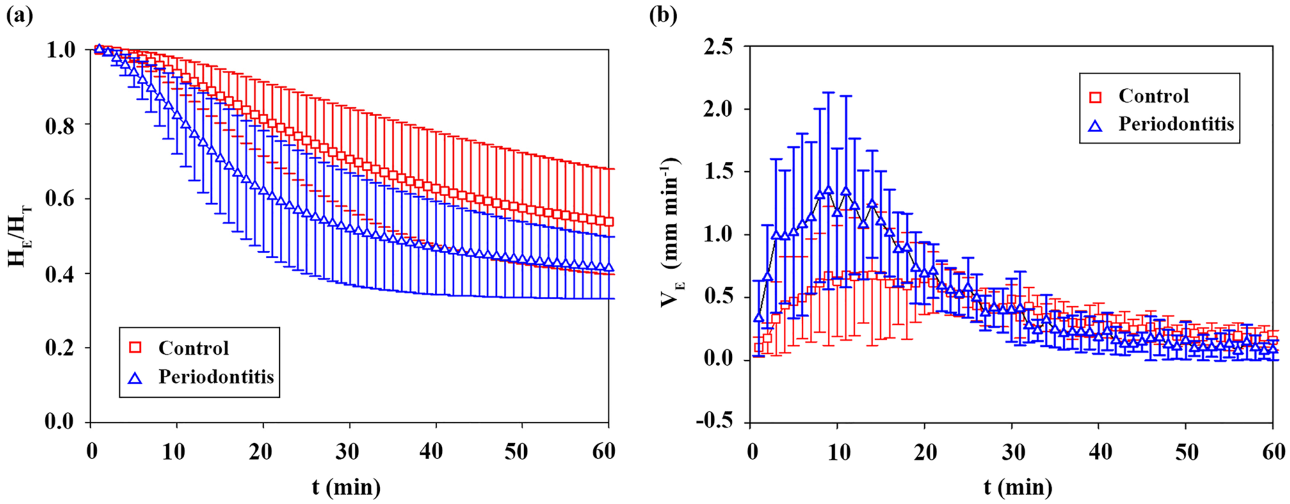

Figure 8. (**a**) Temporal variation of H_E/H_T for the control and periodontitis subjects. (**b**) Instantaneous time rate of sedimentation of V_E for the normal subjects and periodontitis patients. All data shown are values of mean ± SD (n = 6).

tioned at the top in ascending order. The methods are not significantly different if their mean ranks share the same color bar that is less than or equal to the CD.

According to the Nemenyi test rank, GRU–PLSR achieves the lowest mean rank in both of $\overline{\text{MAPE}}$ and $\overline{\text{RMSE}}$. Furthermore, the ensemble models coupled with PLSR and LASSO outperform the other meta models and the individual models with 95% confidence. The findings suggest that both PLSR and LASSO are useful meta models and that GRU–PLSR is the best-performing ensemble model for the ESR prediction.

Application of GRU–PLSR for ESR prediction of periodontitis patients.

Periodontal disease is a bacterial infection that damages the supporting tissues of the teeth. It may result in tooth loss and elevation of the body's inflammation. Some studies reported relatively high ESR values in patients with periodontitis disease[58,59]. The ESR difference between normal subjects and periodontitis patients was monitored to examine the effect of periodontitis disease. For comparison, 6 control subjects were randomly sampled from 304 normal subjects. Temporal variation of H_E/H_T for the control group and periodontitis patients is illustrated in Fig. 8a.

The ESR tendency shows differences between the two groups because ESR is increased by infectious and inflammatory diseases. The mean values of ESR for the control and periodontitis groups are 0.537 ± 0.141 and 0.414 ± 0.083, respectively. This indicates that a fast sedimentation trend was observed in periodontitis patients. The different tendency from periodontitis disease is similar to a previous result of the faster sedimentation for a patient group in a comparison of characteristic time[8]. To further examine the effect of periodontitis disease, variations of V_E for the control and periodontitis groups are illustrated in Fig. 8b. Based on V_E, the erythrocytes for periodontitis patients rapidly decreased in 15 min in comparison to the control subjects, which indicates that the periodontitis disease accelerates the aggregation and sedimentation phases.

GRU–PLSR trained using normal subjects was applied to the ESR prediction for periodontitis patients to demonstrate the practical applicability of the ensemble model. The Bland–Altman method was used to compare the similarity between methods[60]. Figure 9 displays the difference between the actual ESR values and predicted ESR values according to their mean values. The bold line and dashed lines denote the mean value of bias and 95% limits of agreement, respectively. As the presence of data in the 95% limits indicates that the two methods are well matched, ESR values estimated by the GRU–PLSR are in good agreement with the actual ESR values. However, the limited sample size of patients with periodontitis may lead to biased results, and further study with more data is needed.

Contributions of the study.

To the best of our knowledge, this is the first paper that analyzes and predicts the sedimentation tendency of erythrocytes using deep learning algorithms. The key contributions of this study are as follows:

- The use of deep learning approaches to predict ESR, in an attempt to find faster means of obtaining the ESR
- The use of ensemble models to improve the generalization ability on unseen data.
- The performance evaluation of various deep learning and ensemble models to find a best predictive model for ESR based on post-hoc analysis.

Conclusion

In this study, the optimal sequence length was determined based on LSTM and GRU to reasonably predict the ESR value before 60 min. For improvement of ESR prediction, ensemble models with four meta models were suggested and compared. Empirical and statistical results demonstrated that PLSR and LASSO were superior meta

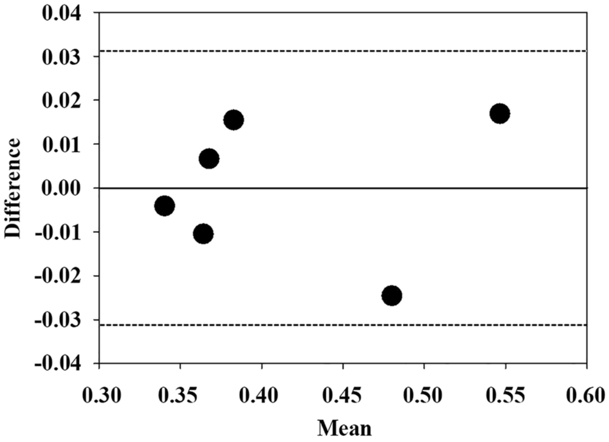

Figure 9. Bland–Altman plot demonstrating difference between the actual ESR values and predicted ESR values obtained by the GRU–PLSR model for periodontitis subjects. The horizontal axis represents the mean ESR values estimated by the two values, and the vertical axis shows the difference between them. The bold line and dashed lines denote the mean value of bias and 95% limits of agreement, respectively (n = 6).

models, and GRU–PLSR achieved the lowest mean ranks among all methods. Finally, the GRU–PLSR model was applied to the prediction of ESR values for patients with periodontitis. Based on our results, it can be concluded that GRU–PLSR is able to predict ESR values from inflammatory and non-inflammatory conditions with high accuracy in only 15 min. Therefore, GRU–PLSR could be useful for clinical applications in ESR measurement, where prompt decision-making is needed.

Data availability

The datasets used and analysed during the current study available from the corresponding authors on reasonable request.

Received: 5 April 2022; Accepted: 26 October 2022
Published online: 15 November 2022

References

1. Westergren, A. Studies of the suspension stability of the blood in pulmonary tuberculosis 1. *Acta Med. Scand.* **54**, 247–282 (1921).
2. Jou, J. *et al.* ICSH review of the measurement of the erythocyte sedimentation rate. *Int. J. Lab. Hematol.* **33**, 125–132 (2011).
3. Yeom, E., Nam, K. H., Paeng, D. G. & Lee, S. J. Effects of red blood cell aggregates dissociation on the estimation of ultrasound speckle image velocimetry. *Ultrasonics* **54**, 1480–1487 (2014).
4. Yeom, E., Jun Kang, Y. & Joon Lee, S. Hybrid system for ex vivo hemorheological and hemodynamic analysis: A feasibility study. *Sci. Rep.* **5**, 1–15 (2005).
5. Yeom, E., Byeon, H. & Lee, S. J. Effect of diabetic duration on hemorheological properties and platelet aggregation in streptozotocin-induced diabetic rats. *Sci. Rep.* **6**, 1–11 (2016).
6. Stokes, G. G. On the effect of the internal friction of fluids on the motion of pendulums. *Trans. Cambridge Philos. Soc.* **9**, 8 (1851).
7. Olshaker, J. S. & Jerrard, D. A. The erythrocyte sedimentation rate. *J. Emerg. Med.* **15**, 869–874 (1997).
8. Hong, H., Song, J. M. & Yeom, E. Micro-vibrational erythrocyte sedimentation rate (ESR) for sensitive measurement of erythrocyte aggregation. *J. Vis.* **24**, 749–760 (2021).
9. Shteinshnaider, M. *et al.* Shortened erythrocyte sedimentation rate evaluation is applicable to hospitalised patients. *Eur. J. Intern. Med.* **21**, 226–229 (2010).
10. Nishisako, H. *et al.* Investigation on prediction formulae for calculating erythrocyte sedimentation rate. *J. Gen. Fam. Med.* **18**, 146 (2017).
11. Lapić, I. *et al.* Automated measurement of the erythrocyte sedimentation rate: method validation and comparison. *Clin. Chem. Lab. Med.* **57**, 1364–1373 (2019).
12. Yeom, E. & Lee, S. J. Microfluidic-based speckle analysis for sensitive measurement of erythrocyte aggregation: A comparison of four methods for detection of elevated erythrocyte aggregation in diabetic rat blood. *Biomicrofluidics* **9**, 024110 (2015).
13. Ormseth, M. J. *et al.* Utility of a novel inflammatory marker, GlycA, for assessment of rheumatoid arthritis disease activity and coronary atherosclerosis. *Arthritis Res. Ther.* **17**, 1–8 (2015).
14. Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R. & Drăghici, S. Machine learning and its applications to biology. *PLoS Comput. Biol.* **3**, e116 (2007).
15. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
16. Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).
17. Kong, W. *et al.* Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Trans. Smart Grid* **10**, 841–851 (2017).
18. AlDahoul, N. *et al.* Suspended sediment load prediction using long short-term memory neural network. *Sci. Rep.* **11**, 1–22 (2021).
19. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **8**, 1–12 (2018).
20. Yilmaz, Y. & Buyrukoğlu, S. Hybrid machine learning model coupled with school closure for forecasting COVID-19 cases in the most affected countries. *Hittite J. Sci. Eng.* **8**, 123–131 (2021).

21. Buyrukoğlu, S. Promising cryptocurrency analysis using deep learning. In *2021 5th International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*. 372–376 (IEEE, 2021).
22. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, 1249 (2018).
23. Mc Loone, S. & Irwin, G. Improving neural network training solutions using regularisation. *Neurocomputing* **37**, 71–90 (2001).
24. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).
25. Moretti, F., Pizzuti, S., Panzieri, S. & Annunziato, M. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing* **167**, 3–7 (2015).
26. Li, C., Tang, G., Xue, X., Saeed, A. & Hu, X. Short-term wind speed interval prediction based on ensemble GRU model. *IEEE Trans. Sustain. Energy* **11**, 1370–1380 (2019).
27. Puleio, A. Recurrent neural network ensemble, a new instrument for the prediction of infectious diseases. *Eur. Phys. J. Plus* **136**, 1–15 (2021).
28. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).
29. Schapire, R. E. The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
30. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
31. Buyrukoğlu, S. Improvement of machine learning models' performances based on ensemble learning for the detection of Alzheimer disease. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*. 102–106 (IEEE, 2021).
32. Buyrukoğlu, S. & Savaş, S. Stacked-based ensemble machine learning model for positioning footballer. *Arab. J. Sci. Eng.*1–13 (2022).
33. Buyrukoğlu, S. New hybrid data mining model for prediction of Salmonella presence in agricultural waters based on ensemble feature selection and machine learning algorithms. *J. Food Saf.* **41**, e12903 (2021).
34. Buyrukoğlu, S., Yılmaz, Y. & Topalcengiz, Z. Correlation value determined to increase Salmonella prediction success of deep neural network for agricultural waters. *Environ. Monit. Assess.* **194**, 1–12 (2022).
35. Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. (Cornell Aeronautical Lab Inc Buffalo NY, 1961).
36. Hornik, K., Stinchcombe, M. & White, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* **3**, 551–560 (1990).
37. Salinas, D., Flunkert, V., Gasthaus, J. & Januschowski, T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36**, 1181–1191 (2020).
38. Rangapuram, S. S. *et al.* Deep state space models for time series forecasting. *Adv. Neural Inf. Process. Syst.* **31**, 7785–7794 (2018).
39. Geravanchizadeh, M. & Roushan, H. Dynamic selective auditory attention detection using RNN and reinforcement learning. *Sci. Rep.* **11**, 1–11 (2021).
40. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**, 157–166 (1994).
41. Kang, D. K. & Kim, M. J. Performance enhancement of SVM ensembles using genetic algorithms in bankruptcy prediction. In *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*. V2–154 (IEEE, 2010).
42. Liu, Y., Fan, Y. & Chen, J. Flame images for oxygen content prediction of combustion systems using DBN. *Energy Fuels* **31**, 8776–8783 (2017).
43. Liu, K., Zheng, M., Liu, Y., Yang, J., & Yao, Y. Deep Autoencoder Thermography for Defect Detection of Carbon Fiber Composites. *IEEE Trans. Industr. Inform*. (2022).
44. Kourentzes, N., Barrow, D. K. & Crone, S. F. Neural network ensemble operators for time series forecasting. *Expert Syst. Appl.* **41**, 4235–4244 (2014).
45. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
46. Rosipal, R. & Krämer, N. Overview and recent advances in partial least squares. in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"* (ed. Saunders, C., Grobelnik, M., Gunn, S., & Shawe-Taylor, J) 34–51 (Springer, 2005).
47. Geladi, P. & Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185**, 1–17 (1986).
48. Botchkarev, A. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006 (2018).
49. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **11**, 86–92 (1940).
50. Nemenyi, P. B. *Distribution-free multiple comparisons*. (Princeton University, 1963).
51. Garcia, S. & Herrera, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *J. Mach. Learn. Res.* **9**, 1 (2008).
52. Rovel, A., L'Huillier, J. & Vigneron, C. Mechanisms of erythrocyte sedimentation. *Biomedicine* **28**, 248–255 (1978).
53. Bull, B. Clinical and laboratory implications of present ESR methodology. *Clin. Lab. Haematol.* **3**, 283–298 (1981).
54. Gao, S. *et al.* Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. *J. Hydrol.* **589**, 125188 (2020).
55. Harrow, C., Singer, A. J. & Thode, H. C. Jr. Facilitating the use of the erythrocyte sedimentation rate in the emergency department. *Acad. Emerg. Med.* **6**, 658–660 (1999).
56. Zhou, Z. H., Wu, J. & Tang, W. Ensembling neural networks: many could be better than all. *Artif. Intell.* **137**, 239–263 (2002).
57. Ahmad, Z. & Zhang, J. Selective combination of multiple neural networks for improving model prediction in nonlinear systems modelling through forward selection and backward elimination. *Neurocomputing* **72**, 1198–1204 (2009).
58. Hutter, J. *et al.* Lower numbers of erythrocytes and lower levels of hemoglobin in periodontitis patients compared to control subjects. *J. Clin. Periodontol.* **28**, 930–936 (2001).
59. Prakash, S., Dhingra, K. & Priya, S. Similar hematological and biochemical parameters among periodontitis and control group subjects. *Eur. J. Dent.* **6**, 287–294 (2012).
60. Bland, J. M. & Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**, 307–310 (1986).

Acknowledgements

This work was supported by the National Research Foundation (NRF) of Korea grant, which is funded by the Korean government (MSIP) (NRF-2018R1A5A2023879 and 2021R1I1A3047664).

Author contributions

E.Y., and J.L. proposed the study. J.L. and H.H. performed the experiment and analyzed the experimental data. J.L., H.H., and E.Y. wrote the paper. All authors participated in completing the manuscript. All authors contributed to revising the manuscript, and approved its final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-23174-0.

Correspondence and requests for materials should be addressed to J.M.S. or E.Y.