Intraoperative Hypotension Prediction Based on Features Automatically Generated Within an Interpretable Deep Learning Model

Eugene Hwang, Yong-Seok Park, Jin-Young Kim, Sung-Hyuk Park, Junetae Kim, and Sung-Hoon Kim

[1] ***Abstract*—Monitoring arterial blood pressure (ABP) in anesthetized patients is crucial for preventing hypotension, which can lead to adverse clinical outcomes. Thus, several efforts have been made to develop an artificial intelligence-based hypotension prediction index. Nevertheless, the use of these indices is limited because they may not provide a convincing interpretation of the association between predictors and hypotension. Herein, we developed an interpretable deep learning model that forecasts hypotension occurrences 10 min before a given 90 s ABP record. Internal and external validations of model performance reported the area under the receiver operating characteristic curve (AUC) as 0.9145 and 0.9035, respectively. Furthermore, the hypotension prediction mechanism can be physiologically interpreted by using predictors representing ABP trends that are automatically generated in the proposed model. Ultimately, we demonstrate high-applicability of a deep learning model that has a high accuracy performance and provides an interpretation of the association between ABP trends and hypotension in clinical practices.**

***Index Terms*—Forecast, Hypotension, Interpretable deep learning, Intraoperative monitoring, Signal processing.**

I. Introduction

INTRAOPERATIVE hypotension (IOH), a frequent adverse event that occurs in anesthetized patients, is widely known to be associated with negative outcomes, including postoperative mortality, acute kidney injury, and myocardial injury [1], [2]. Accordingly, monitoring arterial blood pressure (ABP) in anesthetized patients is a critical task for anesthesiologists to minimize the risk of IOH occurrences [1]. Nevertheless, it is not always possible to take appropriate preemptive measures prior to the onset of IOH because of the high workload of anesthesiologists, who manage various monitoring parameters and respond to unexpected events during surgeries of patients in real time [3]. In such a busy environment, artificial intelligence (AI) could lessen the burden on anesthesiologists by predicting the occurrence of IOH [4].

Various attempts have been made to develop AI-based IOH prediction models [5]–[9], where some of which (i.e. hypotension prediction index (HPI) [6]) have been commercially available. Although these models provide successful predictive results, there are several limitations to their practical use. First, existing commercialized models may not provide convincing explanation of IOH predictive mechanism [10]. The lack of interpretability may lead clinicians to ignore or passively respond to the warnings provided by the model [10], [11]. Given that experts are more likely to take action based on their knowledge and experience than concepts that are not yet well-proven or unfamiliar [11], anesthesiologists may be hesitant to adopt a monitoring index that only provides predicted results without a sufficient basis for their predictions [12]. Second, although a few studies have attempted to provide model interpretability, clinical verification of the existing methods has not yet been sufficiently evaluated [5]. In fact, providing clinically valid interpretation is crucial for encouraging anesthesiologists to actively intervene based on the model [4], [11]. Despite its importance, there has been a lack of in-depth discussion on this aspect. Lastly, the input features utilized in the existing models may be limited in addressing the ABP trend, which is one of the key factors of IOH [5], [6]. Because IOH is often accompanied by alterations in blood volume [13], predictors that specify ABP trends involving intravascular volume status over time may be appropriate for both prediction and interpretation. However, the use of raw ABP records and the microscopic featurization of the ABP waveform employed as predictors in previous models may not accurately reflect the ABP trend [5]–[7].

To address these limitations, we propose a deep learning model that forecasts IOH 10 min prior to its onset with a 90 s ABP record sampled at a rate of 100 Hz as the input. One of the main attributes of this work is developing a framework that makes deep learning-based interpretations compatible with statistical hypothesis testing. Accordingly, the deep learning model in the framework was designed to be decomposed into two broad parts (Fig. 1). First, the generation part of the model

*Corresponding authors: Junetae Kim (*lyjune0070@gmail.com*) and Sung-Hoon Kim* (shkimans@amc.seoul.kr). Eugene Hwang and Yong-Seok Park contributed equally.

Eugene Hwang and Sung-Hyuk Park are with the School of Management Engineering, Korea Advanced Institute of Science and Technology, Seoul, Republic of Korea.

The code for this work is available at https://github.com/JunetaeKim/DWT-HPI.

Yong-Seok Park, Jin-Young Kim, and Sung-Hoon Kim are with the Biosignal Analysis and Perioperative Outcome Research Laboratory, Department of Anesthesiology and Pain Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. Jin-Young Kim is also with the Department of Medical Engineering, University of Ulsan College of Medicine, Seoul, Republic of Korea.

Junetae Kim is with Graduate School of Cancer Science and Policy; and Healthcare AI Team, Healthcare Platform Center, National Cancer Center, Goyang-si, Gyeonggi-do, Republic of Korea.

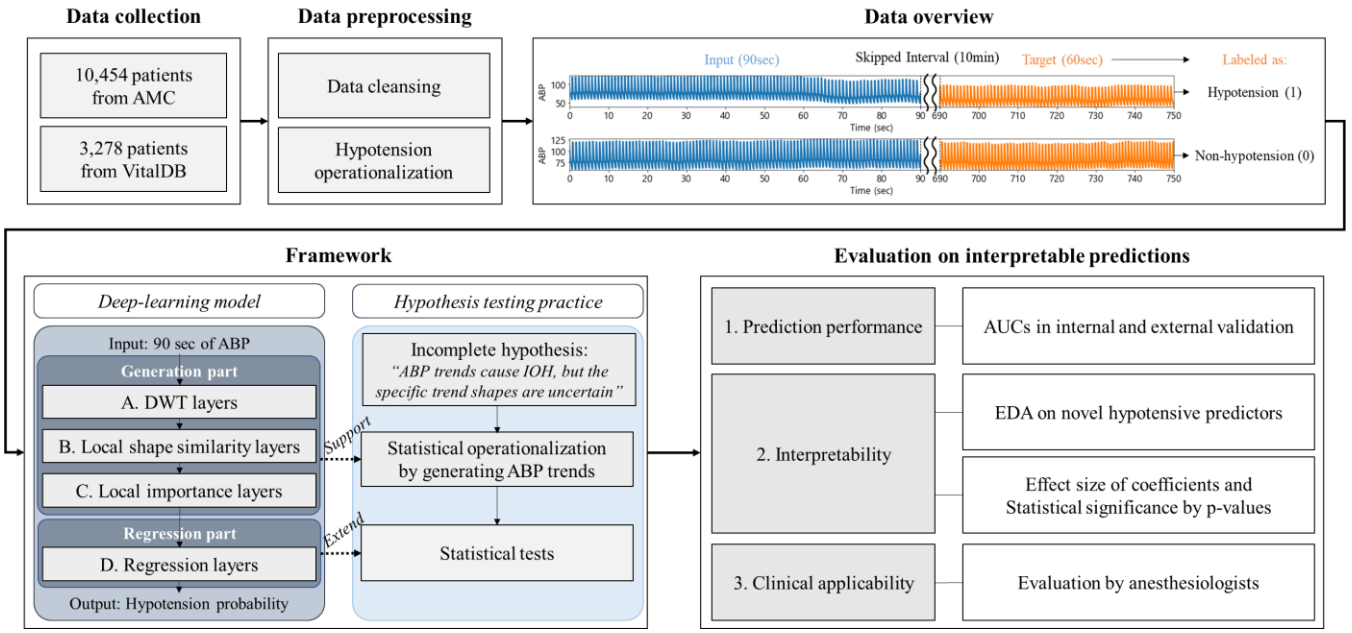


Fig. 1. Research framework.

supports statistical operationalization by generating ABP trend shapes, which are IOH predictors. Second, the regression part extends the generated predictors to be used in post-hoc statistical analysis for significance testing. Another attribute of this study is generating AI-based predictors that can address the physiological basis of hypotension, to enhance its clinical fidelity. Because anesthesiologists are likely to be receptive to a predictor that provides familiar meaning [11], clinically valid predictors may facilitate its adoption into clinical practice.

We summarize our three main contributions as follows:

- *Generation of a well-predictive and interpretable predictor*: A predictor that intuitively addresses the IOH prediction mechanism and has great predictive power was generated through the proposed model.
- *Development of a framework that extends AI-based generated features to be used for statistical analysis*: A framework was proposed to support statistical operationalization and enable the operationalized features to be used in both deep learning-based predictions and statistical tests.
- *In-depth evaluation of the interpretable method:* Our method was evaluated multi-dimensionally through a benchmark test with an existing representative method, feedbacks from anesthesiologists, and theoretical discussions.

The remainder of this paper is organized as follows (Fig. 1). In Section 2, the related works on the existing IOH prediction models are listed. In Section 3, the model's architecture and its mathematical detail are described. In Section 4, the experimental settings are presented. In Section 5, the results are presented in terms of predictive performance, interpretability, and applicability. In Section 6, all the results are discussed, along with several implications and limitations. Finally, Section 7 concludes the paper.

II. Related Works

A. Existing Hypotension Prediction Models

IOH prediction models have been developed with various machine learning and deep learning algorithms based on various predictors. [5]–[8] are representative studies that developed IOH prediction models based on only ABP-driven predictors. Specifically, [6] developed HPI, which indicates IOH-related risk based on a range of 1 to 100. Herein, 3,022 microscopic features extracted from 20 s of ABP waveform were employed in a logistic regression model to predict IOH 15 min in advance. [7] fed multiple ABP-based features extracted by statistical analyses to a random forest model that predicts IOH 5 min in advance. [8] developed an ensemble average deep learning model from convolutional neural network (CNN) and recurrent neural network (RNN) layers to predict IOH 5 min in advance. Herein, 20 s of ABP waveform was employed as a time-series input without hand-crafted feature extraction. [5] also processed 30 s of ABP waveform to train a 1D-CNN based deep learning model for predicting IOH 5, 10, and 15 min in advance. Additionally, a multichannel model with additional predictors (ABP, electrocardiogram, photoplethysmography, and capnography) was developed to compare the prediction performance with the single-channel model using ABP.

In addition to ABP-based features, various other clinical features have been used in the machine learning algorithms predicting IOH [9], [14], [15]. For instance, [9] employed features extracted from physiological signals, time-evolving treatment characteristics, and baseline characteristics to predict IOH 10 min in advance [14] and [15] predicted post-induction hypotension (IOH occurrence during the first 20 min after anesthesia) using features extracted from preoperative medications, medical comorbidities, induction medications, and intraoperative vital signs.

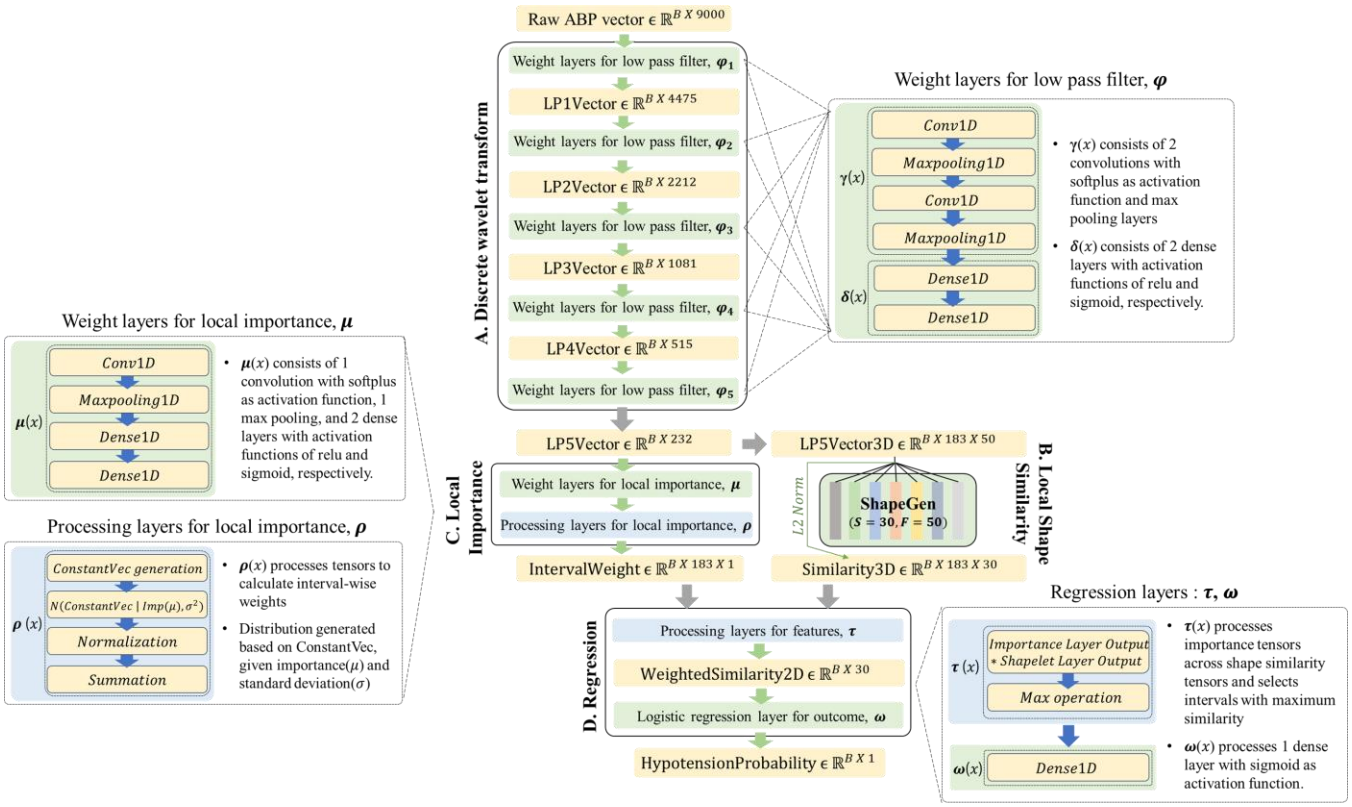


Fig. 2. Architecture of the proposed hypotension prediction model

B. Explainable AI and Hypotension Prediction Model Interpretability

Recently, explainable AI (XAI), which discloses black box characteristics, has been receiving great attention in the medical field [16]. XAI can be grouped into two broad categories: model-specific and model-agnostic methods [17]–[19]. Model-specific methods focus on constructing a transparent mechanism that allows intrinsic interpretation of the model itself. Examples include variable importance computed from boosting or bagging machine learning algorithms and feature maps extracted from certain layers or weights in neural networks [17], [18]. In contrast, model-agnostic methods are applied independently from the model by approximating the relationship between input and output data. Shapley additive explanation (SHAP) and local interpretable model-agnostic explanations (LIME) are representative examples [17], [18].

Existing hypotension prediction models are mainly focused on its prediction performance, presenting either no interpretation [6], [8], [9] or a global interpretation of how each predictor used as input data was relatively important in predicting hypotension. [7], [14] and [15] presented model-specific variable importance from random forest models and stochastic gradient boosting model, respectively. [20] and [21], which trained light gradient boosting algorithms to predict postoperative hypotension, presented model-agnostic variable importance using SHAP.

In terms of the local interpretation based on deep learning models, [5] applied a model-specific method of Grad-CAM to the 1D-CNN model to present the temporal importance within a given input in predicting IOH. Deep learning based agnostic methods have rarely been applied to predict hypotension, to the best of our knowledge. However, an attempt similar to our task was presented in [17] and [22], where electrocardiogram signal was employed as time-series input to predict arrhythmia. Among the three model-agnostic methods of SHAP, LIME, and Anchor applied in [17], only SHAP was evaluated to have an adequate interpretability in signal analysis. Likewise, [23] evaluated SHAP as superior to LIME in terms of explanation invariance (i.e., identity, stability, and separability) in breast cancer prediction and chest X-ray diagnosis. Therefore, the interpretable method proposed through this study was benchmarked only against SHAP, which is the most representative of the agnostic methods.

III. Design for Hypotension Prediction Model

The architecture of the hypotension prediction model, with four broad sections of neural network layers, is displayed in Fig. 2. In the first section (A in Fig. 2), input records are compressed into the overall trend of the ABP waveform by performing five levels of discrete wavelet transform (DWT). Subsequently, vectors of certain shapes are generated to characterize local intervals within the ABP trend vector, followed by a calculation of the similarity between all generated trend shapes and local intervals (B in Fig. 2). Furthermore, parameters are trained in the third section to weight the local intervals (C in Fig. 2). After multiplying the similarity values by the weights, hypotension probability is computed as the model output (D in Fig. 2). The single logit layer (ω) in Fig. 2 facilitates the interpretation of the linear relationship between the local trend shape of the ABP data and the hypotension occurrence.

A. Discrete Wavelet Transform Layers

1) Trend Extraction with Discrete Wavelet Transform

DWT was applied to extract the overall trend by decomposing the ABP waveform into two coefficients (i.e., approximation and detail) with convolutional filters [24], [25]. The DWT of a discrete signal record of x with filter h is defined as follows:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k] * h[n-k]. \quad (1)$$

While the approximation coefficient extracts low-frequency components that represent the overall trend of the signal, the detail coefficient extracts high-frequency components that indicate regional randomness among the signals [24], [25]. Therefore, with recursive applications of such a transformation solely on the approximation coefficient, the overall trend of ABP, which functions as a hypotension predictor, was obtained.

2) Implementation of Trend Extraction within the Model

The role of DWT layers (A in Fig. 2) was to extract the overall trends from the ABP data. The convolutional filter, which decomposed the ABP signal into low-frequency and high-frequency sub-data, was defined as the multiplication of the sinc function and Blackman window.

The sinc function is a time-domain representation of a low pass filter, which encloses information of a certain cut-off point in the frequency, indicated as f_c [26], [27]:

$$sinc(x) = \frac{sin\,(\pi x)}{\pi x}, S[n] = sinc\left(2f_c\left(n - \frac{N-1}{2}\right)\right). \quad (2)$$

Herein, *sinc(x)* and $S[n]$ in Eq. (2) denotes the sinc function and sinc filter, respectively. Because the sinc function is derived from the sine function, the gradient essential for backpropagation can be well defined. Because of this property, f_c can be trained within the model to reduce hypotension prediction error. However, a limitation of the sinc function is the ripples that occur at both ends of the function, which may cause a deviation from the ideal frequency cut-off point in transition. Thus, the Blackman window, defined as

$$w[n] = 0.42 - 0.5\,cos\left(\frac{2\pi n}{N-1}\right) + 0.08\left(\frac{4\pi n}{N-1}\right), \quad (3)$$

was applied to the sinc filter to smoothen the ripples toward zero [28]. The length of the filter (N) was determined as a hyperparameter based on previous studies that utilized the Blackman window for signal processing [28], [29]. An odd number of 51 was selected for N to ensure perfect symmetry in all discrete values within the filter, which were normalized as follows:

$$h[n] = S[n] * w[n], \;\; h_{normalized}[n] = \frac{h[n]}{\sum_{i=0}^{N-1} h[i]}\;. \quad (4)$$

As the initial model input, 9,000 raw ABP records were used (batch, 9,000), and convolution between the filter and the record vector was operated in the second-dimension direction. f_c was learned through weight layers for the low pass filter (φ), as described in Fig. 2. Then, this vector was downsized by averaging all elements in pairs. This trend extraction procedure was performed in five steps, which resulted in compressed ABP records with lengths of 4,475 (LP1Vector), 2,212 (LP2Vector), 1,081 (LP3Vector), 515 (LP4Vector), and 232 (LP5Vector).

B. Shape Similarity Layers

Generalization is important for interpreting associations among variables and can be guaranteed when the parameters for interpretation (i.e., coefficients) are constant after model training. In other words, these parameters must not be endogenous; this condition is satisfied when the parameters cannot be computed from the tensors fed toward the model output [30]. Therefore, multiple vectors representing local shapes of the ABP trends were trained to be independent constants, unaffected by other tensors (ShapeGen in the shape similarity layers in Fig. 2) [31].

Element sizes in the first (S) and the second (F) dimension of the ShapeGen matrix were set as hyperparameters, and their numerical values were 30 and 50, respectively. Using the same length as the local shape of the trend, LP5Vector was reshaped into 183 local intervals (L) with a frame size of 50 (LP5Vector3D (B, 183, 50)). Each local interval in LP5Vector3D, which consisted of 50 data points, was formed by sliding one unit horizontally across the LP5Vector. Similarity (Similarity3D (B, 183, 30)) was calculated between all 30 local shapes and all 183 local intervals of LP5Vector3D as

$$\begin{aligned} &Distance3D_{(b,l.s)} \in \mathbb{R}^{B\times 183\times 30} \\ &= \sqrt{\sum_{f=1}^{F=50}\left(LP5Vector3D_{(b,l,f)}\text{-}ShapeGen_{(s,f)}\right)^2}, \\ &Similarity3D_{(b,l.s)} \in \mathbb{R}^{B\times 183\times 30} = exp(\text{-}\,Distance3D), \end{aligned} \quad (5)$$

where B, l, s, and f refer to batch size, elements of local interval indices, local shape indices, and frame indices, respectively, and $\forall b \in batch, l \in L, s \in S$.

C. Local Importance Layers

Weighting certain elements in tensors may accelerate the convergence of model training and enhance model interpretability by highlighting what needs to be considered during the interpretation [32]. To benefit from these strengths, the model was designed to weight the temporal points at which the ABP shape was critical in predicting hypotension.

Since values within adjacent local intervals of the LP5Vector3D were mostly identical because of the one-unit shift between the intervals, weights were estimated using a Gaussian distribution to give the greatest weight to the most significant intervals and symmetrically weak weights to the remaining intervals. Furthermore, by training the local importance as the summation of probabilities from five Gaussian distributions, the importance was learned with a high flexibility [33]. Specifically, if the mean values that determine the location of Gaussian distribution were scattered, the importance was accordingly spread over 183 local intervals. However, if the mean values were close to each other, the importance was concentrated around a particular local interval because of the summation of the weights.

The procedure of weighting local intervals within the model was implemented through the local importance layers of the overall architecture (Fig. 2). Weight layers for local importance (μ) were trained to obtain the five mean values for Gaussian distributions (M). Given each mean value as a trainable parameter and the standard deviation value (σ) as a

hyperparameter of 0.075 in a set, the processing layers for local importance (ρ) computed probabilities according to the discrete random variable (x) from 0 to 183 as follows:

$$GausDist_{(b,m,l)} \in \mathbb{R}^{B\times5\times183} = N(X|\mu,\sigma^2) \equiv \frac{exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}. \quad (6)$$

All probability values in GausDist were scaled using

$$NormGD_{(b,m,l)} \in \mathbb{R}^{B\times5\times183} = \frac{(GausDist\text{-}MinGD)}{(MaxGD\text{-}MinGD)}, \quad (7)$$

where $MaxGD$ indicates $max_{l\in L}GausDist_{b,m,l}$ and $MinGD$ indicates $min_{l\in L}GausDist_{b,m,l}$ with $\forall b \in batch, m \in M$. Then, weights of the individual distribution were summed across the axis of five probability values, as follows:

$$IntervalWeight_{(b,l)} \in \mathbb{R}^{B\times183} = \sum_{m=1}^{M=5} NormGD_{(b,m,l)}, \quad (8)$$

where $\forall b \in batch$ and $l \in L$. As a result, each discrete probability in the summed distribution was multiplied by the corresponding local interval to obtain the weights.

D. Regression Layers

For models whose outcome is greatly affected by latent variables produced through multiple nonlinear operations, an interpretation of the association between each input variable and output variable may not be possible [34]. However, although multiple layers are implemented in this model structure, the association can still be interpretable as the interpretably extracted ABP trend is fed into the final regression layers (D in Fig. 2). Because the output of the model was a binary variable indicating hypotension occurrence, a logistic regression layer was specified.

Each feature used as the independent variable of the logistic regression layer was based on the weighted similarity between the generated local shapes and the trend extracted in each batch. The weighted similarity values were obtained by multiplying the relative importance of local intervals from the IntervalWeight by the similarity values, element wise, in the second dimension of Similarity3D. Among all weighted values, only the maximum value within 183 local intervals was selected, as follows:

$$IntervalWeight \in \mathbb{R}^{B\times183} \rightarrow IntervalWeight \in \mathbb{R}^{B\times183\times1},$$

$$WeightedSimilarity3D_{(b,l,s)} \in \mathbb{R}^{B\times183\times30} = IntervalWeight_{(b,l,1)} * Similarity3D_{(b,l,s)}, \quad (9)$$

$$WeightedSimilarity_{(b,s)} \in \mathbb{R}^{B\times30} = max_{l\in L} WeightedSimilarity3D_{(b,l,s)},$$

where $\forall b \in batch, s \in S$, and $*$ indicates multiplication operation on each element between two tensors. Thus, second dimension of WeightedSimilarity3D was reduced.

Using 30 weighted similarity values, which represent the most similar local intervals of each shape, the logistic regression was specified as

$$P(y_{(b)} = 1) = \sigma\left(\alpha + \sum_{s=1}^{S=30} \beta_{(b,s)} WeightedSimilarity_{(b,s)}\right), \quad (10)$$

where σ indicates the sigmoid function and y value of 0 and 1 indicate non-hypotension and hypotension, respectively, with $\forall b \in batch$. Under this model, the linear association between the local ABP shapes and hypotension occurrence can be interpreted based on the coefficient values (β_s).

E. Objective Function

The objective function was defined as a summation of the binary cross-entropy and the shape loss. While binary cross-entropy minimized the prediction error (i.e., hypotension), the shape loss generated ABP trend shapes by reducing the distance between the generated features and the local ABP intervals. Because the similarity values used as the input variables in the logistic regression layer were weighted, the Euclidean distance values were also weighted for computing shape loss.

In the local importance layers (C in Fig. 2), the normalized probability values from the five Gaussian distributions at each local interval were summed into a single weight value. Hence, the importance may be concentrated in a certain local interval if the five mean values were learned by focusing on that local interval. Although this strategy is efficient for training parameters for the task of hypotension prediction, it may impede learning various local shapes. To alleviate the restriction, all the normalized probability values from the five Gaussian distributions were applied as weights independently when generating local shapes. Hence, each importance from the five distributions contributed equally when training parameters for the local shapes, which facilitated learning various patterns of local shapes.

The implementation of the shape loss computation within the model started from the Distance3D tensor in Eq. (5). Distance3D was expanded into Distance4D by repeating the newly added vector five times in the second dimension. NormGD was transformed into MaxNormGD by replacing all non-one values in NormGD with zeros. MaxNormGD was expanded to a 4D tensor to perform an element-wise multiplication with Distance4D. Then, the summation of all multiplicative products between Distance4D and MaxNormGD along the interval axis returned ShapeLoss3D, as follows:

$$Distance3D \in \mathbb{R}^{B\times183\times30} \rightarrow Distance4D \in \mathbb{R}^{B\times5\times183\times30},$$

$$MaxNormGD_{(b,m,l)} = \begin{cases} 1, & NormGD_{(b,m,l)} = 1 \\ 0, & \text{otherwise} \end{cases},$$

$$MaxNormGD \in \mathbb{R}^{B\times5\times183} \rightarrow MaxNormGD \in \mathbb{R}^{B\times5\times183\times1}, \quad (11)$$

$$ShapeLoss3D_{(b,m,s)} \in \mathbb{R}^{B\times5\times30} = \sum_{l=1}^{L=183} Distance4D_{(b,m,l,s)} * MaxNormGD_{(b,m,l,1)},$$

where $\forall b \in batch, m \in M$, and $s \in S$. Next, the minimum distance among the 30 local shapes in ShapeLoss3D was selected, which returned ShapeLoss2D:

$$ShapeLoss2D_{(b,m)} \in \mathbb{R}^{B\times5} = min_{s\in S} ShapeLoss3D_{(b,m,s)}, \quad (12)$$

where $\forall b \in batch$ and $m \in M$. The second dimension in ShapeLoss2D represents the distances computed at five particular local intervals between local shapes that were the most similar. The average of all elements in ShapeLoss2D was

used as the final shape loss:

$$ShapeLoss = \frac{1}{(B*5)}\sum_{b=1}^{B}\sum_{m=1}^{M=5} ShapeLoss2D_{(b,m)}. \quad (13)$$

To minimize the loss, the final objective function of the model was

$$Loss = -\frac{1}{B}\sum_{b=1}^{B}\{p_b log(p_b) + (1-p_b) log(1-p_b)\} + \frac{1}{(B*5)}\sum_{b=1}^{B}\sum_{m=1}^{M=5} ShapeLoss2D_{(b,m)}, \quad (14)$$

$$\hat{\omega} = Argmin_{\omega} Loss\left(y_{(1=hypo,0=Non)}, \hat{y} \mid \omega\right),$$

where w is the matrix of all network weights.

IV. Experimental Setting

A. Data Processing

1) Data Acquisition

ABP data records were obtained from two independent medical institutions and processed according to model settings. A total of 10,454 patients from Asan Medical Center, stratified with surgical durations of less than 3 hours, 3–5 hours, and greater than 5 hours, were processed into 1,548,927 data samples. The number of patients in each group was 3,181, 3,244, and 4,029, respectively. The entire set of processed data was then randomly sampled into a training dataset of 600,000 samples and an internal validation dataset of 60,000 samples. The study of these data was approved by the Institutional Review Board of Asan Medical Center (No. 2021-1000), and the requirement for written informed consent was waived because of minimal risk to the study participants. Additionally, ABP records from 3,278 patients in the VitalDB database, an open data repository of intraoperative vital signs from Seoul National University Hospital, were obtained and processed into an external validation dataset of 60,000 samples [35]. Patient information collected from the two institutions are listed in Table I.

2) Data Cleansing

Raw ABP data was cleansed by following two procedures. The first procedure excluded the highly deviating raw ABP records with values less than 25 or greater than 200. After this exclusion, only subsets with a series of at least 17 min without data discontinuity were selected and were smoothed by moving average with a window size of 3.

The second filtering procedure removed the subsets in which at least one of the 20 adjacent ABP cycles differed from the centroid value (i.e., the mean) of the 20 adjacent cycles. Specifically, systolic (i.e., peak) and diastolic (i.e., valley) pressures at every cycle of the ABP waveform were identified. Then, ABP values within a cycle were resampled to have the same vector length (i.e., 100 data points) using a fast Fourier transform [36]. Any cycle with ABP records deviating by at least 15% from the mean value of 20 adjacent cycles were considered as noise candidates. Among these, mean values were calculated from a set of 20 peaks and valleys, without considering the remaining data points. Next, cycles with a peak or valley value deviating by at least 15% from the mean values of the consecutive cycles were excluded. Finally, subsets with a series of at least 15 min without data discontinuity were included in the training dataset.

TABLE I
Patient Characteristics for Internal and External Validation

	VitalDB (n=3,278)	AMC (n=10,454)
Age (years)	59.4 (14.3)	58.2 (14.3)
Sex (male/female)	1,831 / 1,447	5,646 / 4,808
Weight (kg)	61.5 (11.4)	63.4 (12.2)
Anesthesia duration (min)	213.2 (107.3)	234.6 (171.4)

Average and standard deviation values were reported for age, weight, and anesthesia duration, whereas sex was reported as counts.

3) Hypotension Operationalization

The model outcomes (1= hypotension or 0= non-hypotension) were operationalized based on mean arterial pressure (MAP = $\frac{\text{SystolicPressure}+2*\text{DiastolicPressure}}{3}$) [37]. Based on previous research, periods where MAP was maintained at less than 65 mmHg for at least 1 min were defined as hypotension [5], [6]. In contrast, periods where MAP was maintained at more than 75 mmHg for at least 1 min were operationalized as non-hypotension [6]. Remaining periods with MAP values between 65 and 75 mmHg were considered as a "gray zone" [6]. Only definite hypotension (i.e., MAP < 65 mmHg) and non-hypotension (i.e., MAP > 75 mmHg) events were employed to train the prediction model. However, because it is worth validating the sensitivity of model accuracy according to the gray zone during external validation, three additional tests with various MAP thresholds applied to non-hypotension samples were performed. Accordingly, in terms of the external validation set, non-hypotension was redefined as 1) samples with at least one incidence of the MAP being greater than 65 mmHg within 1 min, or samples where all MAP values were greater than 2) 65 mmHg or 3) 70 mmHg for at least 1 min.

As hypotension cases are less common than non-hypotension cases, label distribution was adjusted using random sampling, with a maximum limit of 30 non-hypotension cases per patient. This final procedure mitigated any imbalance in the distribution, which may adversely affect convergence in model training [38].

B. Evaluation on Interpretability

1) Statistical Significance of Association Between Generated ABP Shapes and Hypotension

In the proposed deep learning model, the logistic regression coefficients (β) from the D. Regression layers in Fig. 2 were estimated using the gradient descent based optimizer [39]. Although the association between the local ABP shapes and hypotension occurrence can be interpreted by these coefficient values, the deep learning model without p-values cannot provide statistical significance for this relationship [40]. Therefore, additional logistic regression analysis, based on the Newton-Raphson optimizer, was conducted to estimate the p-values of these coefficients.

The Newton-Raphson method approximates the solution of an equation of the form f(x) = 0 by identifying a tangent from the current value and by updating the value to the point where the

tangent meets the x-axis [41], [42]. The Newton-Raphson method was applied to the maximum likelihood estimation in the logit model. Using this method, parameters that lead the first-derivative form of the likelihood function to 0 were iteratively estimated until convergence [42]. Finally, the statistical significance of coefficients (θ) can be determined by the hypothesis test in regression analysis where the null hypothesis states that each coefficient equals 0. In this model, the similarity values, which were predicted from the WeightedSimilarity layer after training the proposed model, were employed as the independent variables.

In regression analysis, high correlations between independent variables may lead to biased coefficient estimates [43]. Hence, data preprocessing was performed to deal with high correlations between all pairs of 30 weighted similarity values. Specifically, only one trend shape was selected from the groups with correlations greater than 0.8, allowing only distinct shapes that were not highly correlated with others to be further analyzed. Subsequently, only the weighted similarity values corresponding to the representative shapes were included as independent variables in the logistic regression model, as in Eq. (10).

After fitting the regression model, only statistically significant ($p<0.01$) shapes were selected, and the sign and effect size of their coefficients were further analyzed. Additionally, the consistency between the selected coefficients (θ) and those estimated in the deep learning model (β) were analyzed.

2) *Exploratory Data Analysis based on Descriptive Statistics*

Exploratory data analysis (EDA) was additionally performed for multidimensional review of the proposed model and predictor. First, prediction reliability of the maximum similarity shape was assessed based on both the true positive rate (i.e., accurate prediction rates for hypotension) and the true negative rate (i.e., accurate prediction rates for non-hypotension) [43]. Herein, a confusion matrix was computed based on the number of samples that had the maximum similarity to each shape.

Additionally, the overall magnitude of the values of the generated ABP trend shapes was analyzed to demonstrate the consistency of the association between hypotension and these shapes with clinical knowledge. Accordingly, statistically significant ABP trend shapes were examined by extracting the shape vectors from ShapeGen layer.

Furthermore, a distribution of local importance locations was investigated to identify temporal positions within the compressed ABP trend that were significantly influential in predicting hypotension. Because there were five Gaussian distributions (C in Fig. 2), the number of mean positions over 183 intervals was counted.

C. Evaluation on Applicability of Proposed Interpretable Method

1) *Interpretability Comparison with SHAP*

The interpretability of the proposed model was further evaluated by benchmarking the SHAP, which has been widely used to interpret the prediction mechanism of machine learning or deep learning models. Given that SHAP explains model predictions by assigning weights to certain features [44], the absolute value of SHAP represents how much a feature has contributed to the prediction. Because SHAP values in this study were computed by approximating the expected gradients [45] with respect to 9,000 temporal ABPs, they indicated the influence of temporal position within the input ABPs on hypotension prediction. The SHAP values were compared with those of the proposed method to explore in which aspects our interpretable approach has more interpretable power.

2) *Assessment by Survey for Anesthesiologists*

The proposed interpretation method was evaluated in three aspects by conducting a survey on anesthesia experts [46]. The first aspect concerns whether the given information addressing the basis of the prediction is clinically relevant to hypotension predictors (i.e., clinical fidelity). The second is whether providing interpretative predictors along with the predicted probability of hypotension is useful in clinical practice (i.e., clinical usefulness). The last has to do with the clinician's willingness to intervene, given the information on which the prediction is based (i.e., willingness to intervene).

A group of 17 board-certified anesthesiologists from Asan Medical Center with 5–21 years of experience participated in this survey. Model interpretations were grouped into three categories:(1) SHAP values, (2) trend shape similarity, and (3) odds ratio of hypotension occurrence and history of predicted probabilities in addition to trend shape similarity. Accordingly, a scenario-based clinical guideline was presented as a set of visual summaries of 10 randomly selected samples for each category. Based on the given visual summaries, participants were asked three questions measuring (1) clinical fidelity, (2) clinical usefulness, and (3) willingness to intervene, on a 5-point Likert scale. In summary, each participant was given 30 visual summaries (10 samples * 3 categories) and nine questions (3 questions * 3 categories). A list of nine questions and a sample visual summary for each category are illustrated in Figs. 7, 8, and 9 in supplementary document.

D. Ablation Study

Two ablation experiments were conducted to test how the removal of specific layers in the certain parts of the proposed model (A to D in Fig. 2) affected both the interpretability and predictability. In the first ablation experiment, local importance layers were removed (C in Fig. 2) to disregard the weights applied to the temporal position of the extracted ABP trend. The second ablation experiment removed both local importance layers and DWT layers (A and C in Fig. 2). Instead, Conv1D and Maxpooling1D layers replaced the DTW layers to compress the raw ABP. Structural details of the models for the ablation study are provided in Fig. 6 in supplementary document. The results of both models were compared with our proposed method in terms of predictive performance and model interpretability.

V. Result

A. Prediction Performance

The hypotension prediction performance of the proposed model was evaluated with a training set, an internal validation set, and an external validation set. The areas under the receiver

operating characteristic curves (AUCs) of the three sets with non-hypotension, defined as all MAP values being above 75 for at least 1 min, are reported in Table II-a as 0.9146, 0.9145, and 0.9035, respectively. The AUCs of the external validation sets with non-hypotension defined as 1) at least one MAP above 65, 2) all MAP above 65, and 3) 70 for more than 1 min, are reported in Table II-b as 0.8337, 0.8588, and 0.8831, respectively. Although the prediction performance slightly decreased as more indefinite samples were included as non-hypotension samples in the dataset, the overall classification performance was excellent.

TABLE II-a
OVERALL PREDICTION PERFORMANCE

	Proposed Model	Ablation Model 1	Ablation Model 2
Training set	0.9152	0.9090	0.9192
Internal validation set	0.9145	0.9087	0.9165
External validation set	0.9035	0.8923	0.9057

AUC score is reported. Non-hypotension samples were defined as all MAP above 75 for at least 1 min.

TABLE II-b
PREDICTION PERFORMANCE WITH GRAY-ZONE SAMPLES

At least one MAP ≥ 65	All MAPs ≥ 65	All MAPs ≥ 70	All MAPs ≥ 75
0.8337	0.8588	0.8831	0.9035

Definition of non-hypotension samples was set accordingly. Results are based on the proposed model with external validation set as AUC score.

B. Evaluation on Interpretability

1) Statistical Significance of Association Between Generated ABP Shapes and Hypotension

Among the 30 shapes generated, four were found to have low correlations in weighted similarity. Thus, the coefficient values corresponding to these four shapes within the regression layer (ω) were further analyzed. Among the four coefficients estimated by the Newton-Raphson estimator, three were statistically significant but one was not within the 1% significance level. Details of the coefficients for each estimation method, including p-values and effect sizes, are given in the proposed model column in Table III. Shapes A and B, which had negative coefficients, were associated with non-hypotension, whereas shape C, which had a positive coefficient, was associated with hypotension. Moreover, the coefficient of A was smaller than that of B, which suggests that shape A is more closely related to non-hypotension than shape B. These results are consistent in both the gradient descent and the Newton-Raphson optimizer, which may ensure a robust interpretation of the association between the shapes and hypotension in the deep learning model.

2) Exploratory Data Analysis Results

Table IV presents the confusion matrix summarizing how well the samples most similar to the chosen shapes were classified as hypotension. The column of proposed model in the table reports that the true negative rate is outstanding for shapes A and B, whereas the true positive rate is superlative for shape C. Because the true positive rate or true negative rate for each shape is high in Table IV, the generated ABP shapes may have

TABLE III
COEFFICIENT VALUES FROM THE LOGIT LAYER

	Proposed Model		Ablation Model 1		Ablation Model 2	
Trend shape ID	**Gradient Descent Method**	**Newton-Raphson Method (p-value)**	**Gradient Descent Method**	**Newton-Raphson Method (p-value)**	**Gradient Descent Method**	**Newton-Raphson Method (p-value)**
A	-2.9222	-17.7662 (<0.001)	1.3649	-4.0752 (<0.001)	-1.5114	-24.1641 (<0.001)
B	-2.8927	-5.2918 (<0.001)	10.6953	10.0754 (<0.001)	1.3149	15.1472 (<0.001)
C	0.4064	1.5521 (<0.001)	–	–	–	–
D	0.7578	0.0289 (0.6644)	–	–	–	–

TABLE IV
CONFUSION MATRIX OF MAXIMUM SIMILARITY WITH THE REPRESENTATIVE SHAPES

	Proposed Model		Ablation Model 1		Ablation Model 2	
A	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension
Actual non-hypotension	77 (0.939)	0 (0)	28,355 (0.749)	1,950 (0.051)	27,621 (0.461)	5,190 (0.086)
Actual hypotension	4 (0.049)	1 (0.012)	4,320 (0.114)	3,249 (0.086)	3,682 (0.061)	23,507 (0.392)
B	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension
Actual non-hypotension	21,092 (0.906)	43 (0.002)	4 (0)	2,502 (0.113)	0 (0)	0 (0)
Actual hypotension	2,126 (0.091)	34 (0.001)	10 (0.001)	19,610 (0.886)	0 (0)	0 (0)
C	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension	Predicted as non-hypotension	Predicted as hypotension
Actual non-hypotension	7,038 (0.192)	4,561 (0.125)	–	–	–	–
Actual hypotension	2,133 (0.058)	22,891 (0.625)	–	–	–	–

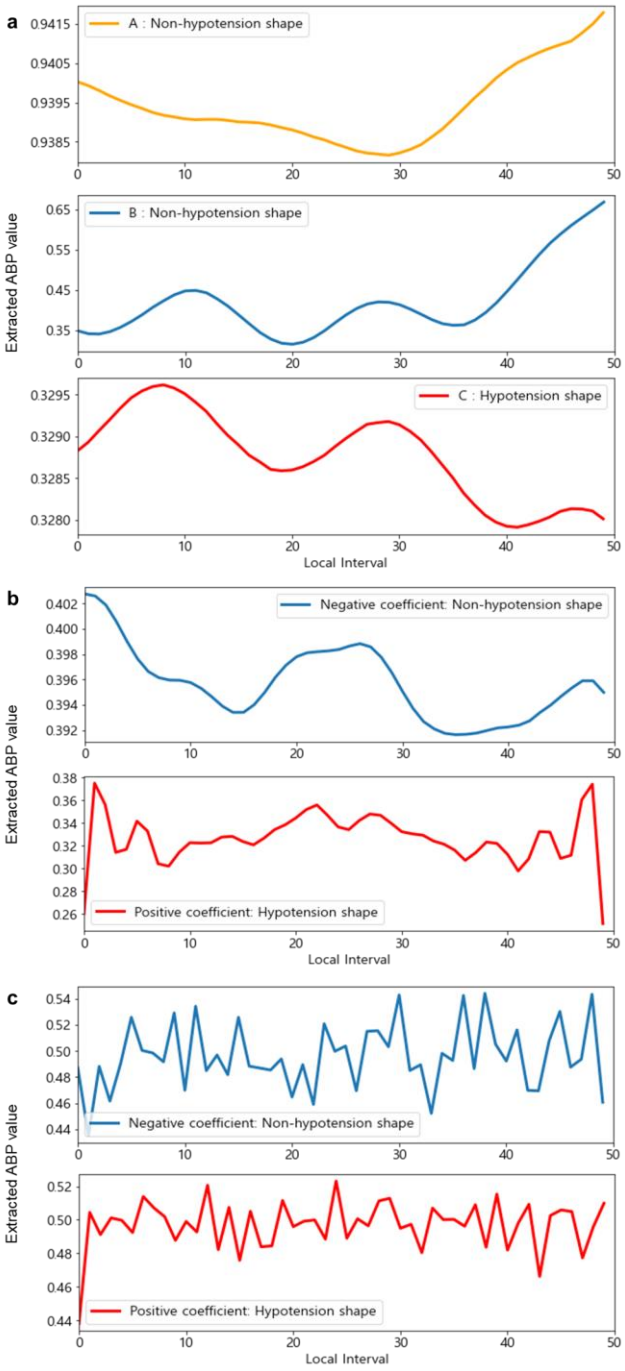


Fig. 3. Representative shapes of local ABP trends: a) Proposed model; b) Ablation model 1; and c) Ablation model 2.

a high reliability in predicting hypotension.

Next, Fig. 3a presents the overall magnitude of the value of the three statistically significant local ABP shapes extracted from ShapeGen layer. In terms of the overall trend in values, shapes A and B gradually increase over time, whereas shape C gradually decreases over time. Going by the definition of hypotension, in which low MAP values persist for more than 1 min, these results suggest that the shape C is associated with a precursor to developing hypotension. Additionally, the exceptionally large values of shape A (i.e., 0.94) were greater than those of shape B (i.e., 0.40), which may suggest that shape A has a stronger relation to non-hypotension. Thus, these results may demonstrate a reasonable association between hypotensive development and the morphological characteristic of each ABP trend shape.

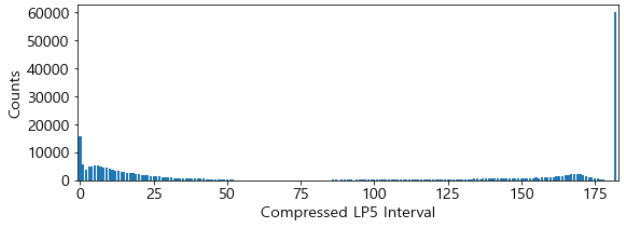


Fig. 4. Counts for each local interval that is influential in the prediction of hypotension.

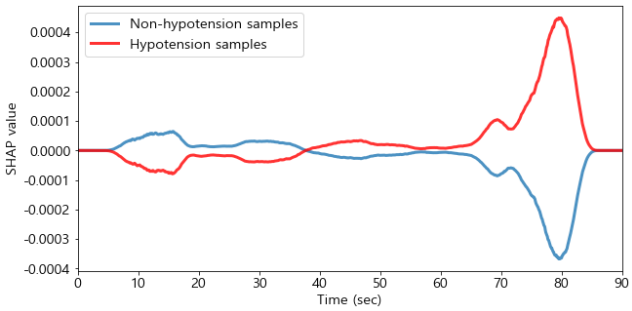


Fig. 5. Mean SHAP values of hypotension and non-hypotension samples.

Finally, Fig. 4 indicates the distribution of weights that have an influence on hypotension prediction according to each of the 183 local intervals. Because weights were assigned by the five parameters for every sample (C in Fig. 2), the distribution was formed by 300,000 values from 60,000 samples of the validation set. As demonstrated in Fig. 4, 20% of the weights are located in the last interval, with only a small number of counts in the forepart. Thus, local trends at the rear end of a given ABP trend may be more critical when predicting hypotension.

C. Evaluation of Applicability of the Proposed Interpretable Method

1) Model Interpretability by SHAP

Fig. 5 shows the mean of SHAP values for hypotensive and non-hypotension samples in the external validation set. Positive or negative SHAP values indicate the temporal-location contribution of the input ABP to the prediction of hypotension or non-hypotension, respectively. For both classes of samples, the averages of absolute SHAP values were high around the end of the input (i.e., 70–85 seconds), which means that ABP data close to the prediction time were more important for inferring the probability. This result is in accordance with the distribution of weights in Fig. 4, where most of the local weights were given to the ends of the extracted ABP trends during hypotension prediction.

2) Survey Results for Anesthesiologists

Table V presents the average and standard deviation values of the survey questionnaires in the three aspects according to the three categories. For all assessments, the mean score increased sequentially from Category 1 to Category 3. Specifically, the average score of all questionnaires in Categories 2 and 3, which evaluates the interpretable components of the proposed model, was 3 or higher. Compared to Category 1, Category 2 was rated

13.2%, 10.2%, and 15.5% higher, and Category 3 was rated 24.4%, 41.0%, and 22.6% higher in terms of clinical fidelity, clinical usefulness, and willingness to intervene, respectively.

D. Ablation Study

Ablation model 1 showed AUCs of 0.9090, 0.9087, and 0.8923 for the training, internal validation, and external validation sets, respectively, which were approximately 0.006 to 0.01 lower than the proposed model (Table II-a). Although two statistically significant shapes were identified (Table III), the coefficient signs for non-hypotension estimated by gradient descent and Newton-Raphson methods were inconsistent with each other. Moreover, trend shapes illustrated in Fig. 3b may not address the association with hypotensive development. In particular, the gradual decline in ABP trends in the shape statistically associated with non-hypotension and the irregular pattern of ABP trend in the hypotension-related shape are not clinically rational. These results may suggest they have little explanatory power as a predictor of hypotension.

Ablation model 2 scored AUCs of 0.9192, 0.9165, and 0.9057 for the training, internal validation, and external validation sets, respectively (Table II-a). Through the model, two shapes that were statistically significant at the significance level of 0.001 were generated (Table III). However, because all samples had the maximum similarity only in shape A in Table IV, there is a lack of representativeness of the shapes that account for hypotension prediction. Moreover, irregular values that fluctuate over time are almost random walks and are unlikely to provide plausible interpretation for the hypotension prediction (Fig. 3c). Thus, ablation model 2, whose AUC was as high as the proposed model, had a poor explanatory power.

VI. Discussion

A. Performance Verification of New Hypotensive Predictor

Through this study, an attempt was made to develop a model with excellent performance in predicting hypotension based on a new predictor, which is different from previous models. In the existing hypotension prediction models, raw ABP data or its features (i.e., time, amplitude, area, and slope) characterizing the ABP waveform were often utilized as input variables [5], [6]. Regarding performance, the model in which raw ABP records were used to forecast hypotension 10 min prior reported an AUC of 0.882 in internal validation [5]. Additionally, in the model where multiple features extracted from the ABP waveform were utilized for the forecast, AUC was reported as 0.92 in external validation [6].

Likewise, our model using the local ABP trends as a new predictor showed excellent performance compared to the existing model. Specifically, in the internal and the external validation, the AUCs of the model were 0.9145 and 0.9035, respectively, showing similar scores to the previous models (Table II-a). These results may suggest that the local ABP trends generated via deep learning layers can be an effective predictor of hypotension.

Moreover, an additional test was conducted for a rigorous performance evaluation. Specifically, unlike previous studies, in which the performance was assessed based only on the well-behaved samples [5], [6], this study evaluated model performance with the inclusion of ill-behaved samples, which accounted for a gray zone of MAP values when labeling non-hypotension. In the external validation of these samples, AUC was reported from 0.8337 to 0.9035, as the MAP value threshold for defining hypotension varied from 65 to 75 (Table II-b). Although it tends to be less accurate in a rigorous setting, the AUC of 0.8337 still indicates a good performance. Thus, these results suggest that the compressed local ABP trends are a robust predictor of hypotension.

TABLE V
Clinical Survey Result

	Category 1	**Category 2**	**Category 3**
Clinical fidelity	3.12 (0.99)	3.53 (0.87)	3.88 (0.78)
Clinical usefulness	2.88 (1.27)	3.18 (1.01)	4.06 (1.14)
Willingness to intervene	3.41 (1.28)	3.94 (0.66)	4.18 (0.73)

Categories 1, 2, and 3 indicate interpretation with SHAP values, trend shape similarity, and odds ratio of hypotension occurrence and history of predicted probabilities in addition to trend shape similarity, respectively. Average and standard deviation values are reported for each category.

B. Clinical Applicability of the Proposed Interpretable Method

In addition to the performance verification, the clinical applicability of the proposed interpretable method was verified in three aspects using the results of the survey. First, the predictors generated by the proposed model may have high clinical fidelity. In terms of theoretical perspective, it is a well-established in physiology that changes in cardiac preload are an important cause of intraoperative hypotension [47]. Specifically, cardiac preload can be indicated by the respiratory variability of the ABP waveform [47], which is an important piece of information that anesthesiologists can obtain by monitoring ABP waveforms. Because the low-frequency component of ABP is mainly caused by respiratory variation [13], generally associating the ABP trends with low-frequency components can play an important role as precursors for hypotension [48]. Accordingly, trends extracted from the ABP waveform components reflecting respiratory variability may have a high degree of clinical fidelity. This theoretical interpretation can be empirically supported by the evaluations of anesthesiologists, which reaffirmed the high association between the proposed predictors and the practice knowledge of clinicians (Table V). Specifically, the highest rating in Category 3 in terms of clinical fidelity may indicate that the interpretable method provided by the proposed model is clinically more informative for understanding hypotension prediction than the conventional method.

Second, the higher ratings of clinical usefulness in Categories 2 and 3 compared with Category 1 (Table V) may indicate that the proposed interpretable method provides more useful information than SHAP. A common feature between SHAP and the proposed interpretable method is that both provide the importance of certain local points for predicting hypotension in a given 90 s of ABP data. However, the difference is that the proposed method provides additional information on the level of similarity between the current ABP trend and representative morphological characteristics of the ABP trend. Given that the identified ABP trend shapes are precursors significantly associated with hypotension, the similarity value may be practically useful by enabling anesthesiologists to acquire

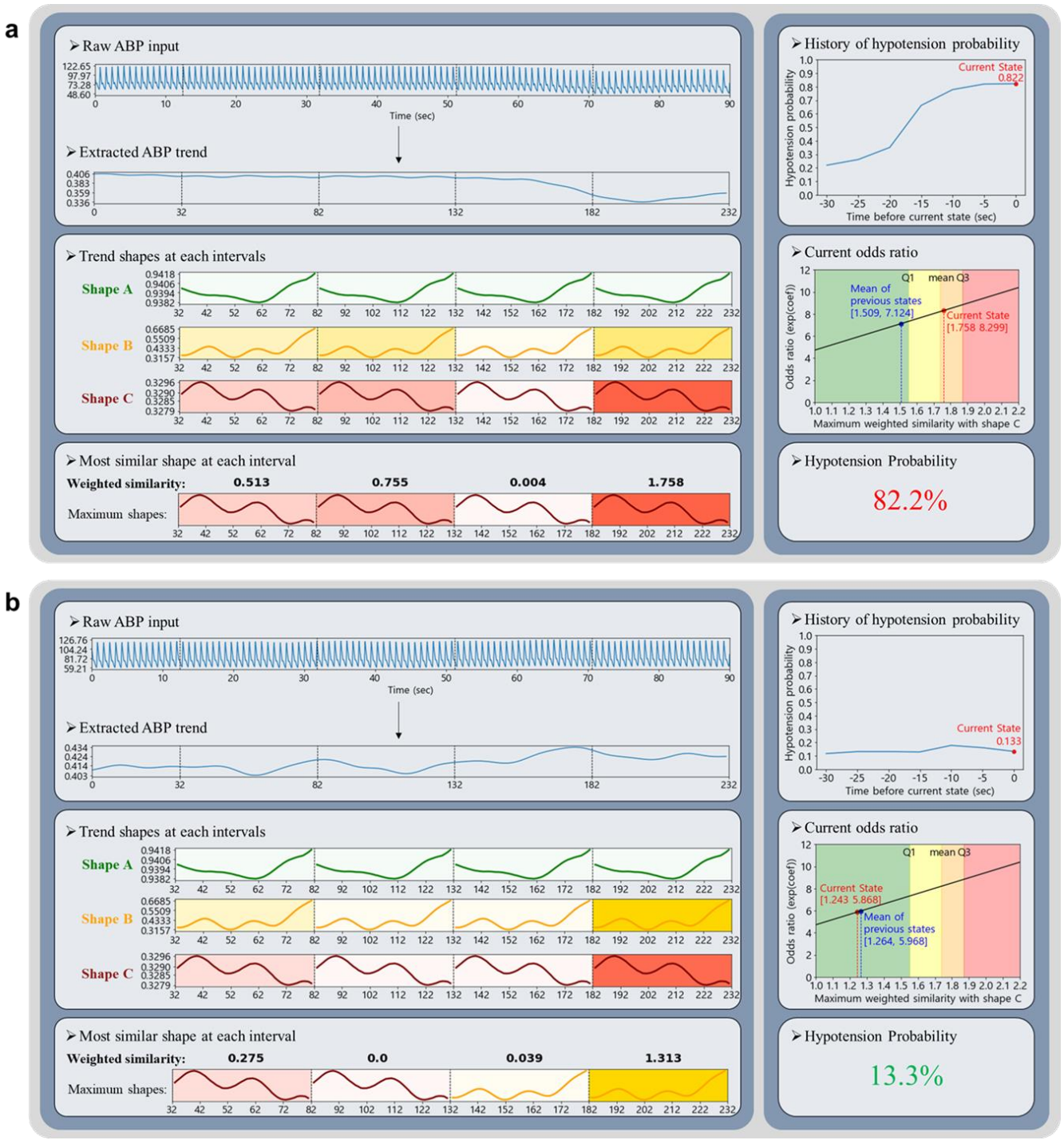


Fig. 6. Visual summary of a scenario-based clinical guideline for interpretable hypotension prediction: a) hypotension sample case and b) non-hypotension

information intuitively.

Lastly, the highest scores in Category 3 in terms of willingness to intervene (Table V) may suggest that the history of predicted hypotension probability and odds ratio provide convincing information about the prognostic signs of hypotension. Especially, because the interpretation based on the odds ratio of logistic regression is one of the most widely accepted methods [43], it may increase the confidence of anesthesiologists in accepting information that warns of possible hypotension occurrence. Consequently, anesthesiologists may be more willing to intervene based on the information that is already familiar to them.

C. A Scenario-based Clinical Guideline Using Visual Summary

To encourage the practical adoption of the proposed interpretable method, we visualize an end-to-end process in Fig. 6, which starts from raw ABP input data and displays the final probability computation of hypotension.

In the first section, raw ABP and its trend, extracted using DWT, are shown. Specifically, the extracted ABP trends are more readable than the raw ABP in terms of overall changes. Because the information on respiratory variation is captured by

the ABP trend, observing trend dynamics in blood pressure may help anesthesiologists monitor a precursor of hypotension.

Second, weighted similarities of the three significant ABP local trends (Fig. 3a) with four consecutive local intervals from the tail of the trend are presented and differentiated by color. This is a local interpretation of the extracted ABP trend, and may yield critical information by foreshadowing future blood pressure status. In particular, local morphological changes of the hypotensive precursor allow anesthesiologists to identify a transition from high to low or from low to high risk of hypotension over time (Fig. 6b). In a surgical environment that requires treatment based on real-time information reading, flagging this notable transition may help anesthesiologists decide when to intervene appropriately.

Third, the association between hypotensive development and each generated ABP shape can be interpreted independently, because a regression analysis is performed with the assumption that independent variables are not highly correlated with each other [49], [50]. This means a relationship between hypotension and the only ABP shape of interest can be interpreted while holding all other ABP shapes constant ("all else equal"). In Fig. 6, the odds of hypotension in terms of the gradient-based coefficient of shape C are plotted to illustrate the shift in the odds ratio from the mean of previous states to the current state, given the maximum similarity value with shape C among the four intervals (Table III). This independent interpretability may improve the efficiency of information acquisition by allowing anesthesiologists to focus on the desired shapes according to their primary interests (i.e., early detection of hypotension development or transition to normal blood pressure).

Finally, the upper right section indicates the change in hypotension probability every 5 s from 30 s prior to the current state, and the hypotension probability at the current state is reported as a single number. In conclusion, the distinct difference in the results of the two sample cases under hypotension and non-hypotension illustrates the clinical intuitiveness of the proposed interpretable method.

D. Theoretical discussion of contributions

Our work contributes to incorporating clinical practices into the development of XAI, providing insights that can bridge the gap between AI-based interpretations and medical practices. The contributions are discussed from a perspective of technology adoption theories, starting with identifying some limitations of existing interpretable methods in medicine.

Traditionally, new knowledge has not been well-integrated into medical practice [51]. Among various reasons, changes proposed without considering clinical needs can be a major obstacle to adopting AI-based emerging concepts [51]. Recently, various attempts have been made to apply XAI to medical prediction problems, to improve interpretability [16]. However, when limiting these applications to the realm of signal data, there is no evidence that existing methods can address clinical needs. Specifically, interpretable methods such as SHAP reveal important temporal points in a signal for predicting an event, as shown in Fig. 5 [22], [52]–[54]. Previous studies have neglected in-depth discussions on whether the presentation of the temporal importance satisfies the needs. Unfortunately, our survey results (Table V), where SHAP-based interpretations were rated as the lowest by anesthesiologists, call into question how seriously previous studies have considered the needs of clinical practice.

Given that change often entails additional burdens [51], low-value innovations further increase barriers to the adoption of new concepts. Thus, efforts should be channeled toward developing interpretable methods acceptable to practitioners in the field. Previous research theorized that perceived usefulness and compatibility with current practice are essential factors for medical technology adoption [55]. Another well-established theoretical study conceptualized that job relevance, defined as capability for supporting one's task, positively affects perceived usefulness [56]. Accordingly, we have attempted to incorporate some important practices relevant to hypotension monitoring into the development of an AI-based interpretation framework.

First, in our proposed method, we have addressed one rationale for hypotension detection in the interpretable mechanism. As discussed previously, the relation between ABP trends and IOH can be well-explained in terms of physiology [13], [47], [48]. Because physiological evidence is a key factor in clinical decision-making [57], the relation can be a fundamental premise for anesthesiologists to detect hypotension. Thus, the interpretation of ABP trends provided via the proposed method is basically compatible with hypotension monitoring practice, which has a positive effect on perceived usefulness. Moreover, the proposed method only highlights trends in ABP data, thereby facilitating anesthesiologists' understanding of physiological changes. Evidently, it is more efficient for anesthesiologists to understand the presented trends than to have to cognitively isolate the trends from raw ABP data and then try to understand them. Increased work efficiency has been theorized to have a positive effect on perceived usefulness [55], thus suggesting that our method is highly useful for anesthesiologists.

Second, we have proposed a framework in which the AI-based interpretation is compatible with hypothesis testing practices involving two main stages: hypothesis building and significance testing. Intriguingly, the structure of our deep learning model can be decomposed to correspond to these two stages. The generation part of the framework in Fig. 1 may assist anesthesiologists in hypothesis building. Specifically, even if anesthesiologists have hypothetical ABP trend shapes causing IOH, they are likely to be uncertain of the specific shapes for a statistical operationalization. Given this incomplete hypothesis, this part completes the operationalization by generating interpretable ABP trend shapes. This completion is attributed to the DWT layers implemented within the generation part. Generally, the typical nonlinear learning structure of deep learning models may generate features that do not account for domain specificity [58]. However, our DWT layers preserve the original nature of ABP waveform when generating trend shapes. Thus, the property of the part, which supports hypothesis building, suggests its high association with job relevance, which positively affects perceived usefulness. Next, the regression part of the framework shown in Fig. 1 has a scalability for statistical significance testing. Statistical models may have advantages over common interpretable methods such as SHAP because they provide generalizable interpretation. Specifically, the SHAP values change according to the distribution of the given input data [18], making it

difficult to generalize the interpretation. However, since p-values and odd ratios are constant regardless of the given input data, logistic regression can provide a generalizable interpretation [43]. Generalizability has been very important in clinical practice for pursuing evidence-based decision-making [16]. Nevertheless, existing studies on XAI have tended not to take this important practice seriously. By contrast, the proposed framework, which integrates AI-based and statistical interpretations, enhances generalizability, suggesting its high compatibility with a major clinical practice.

Third, we have presented a guideline for our method and received evaluations on the method by clinicians, which was often neglected in previous studies. The insufficient guidance provision in previous methods may be due to the paucity of clinically relevant information because domain-specificity was not well incorporated into these methods [12], [58], [59]. By contrast, our interpretable method can provide new types of clinically relevant information by addressing physiological mechanisms. However, even advanced new concepts in medicine may not propagate without proper guidance [51]. Accordingly, we have provided a scenario-based guideline as a visual summary that outlines how the information emerging through our method can be well integrated into current practices. Furthermore, as the ABP trend shapes generated by our model are new, an evaluation of their relevance to the physiological basis is essential. Because the physiological basis is close to the qualitative domain, it was investigated by a group of experts in this study [57]. Although the evaluation was performed by a small number of anesthesiologists, we believe that the limited number of evaluators may not be a major concern when considering the technology adoption lifecycle model [51], [60]. Specifically, the majority adopt new medical concepts through interaction with a minority who have already embraced the concepts [51]. This suggests that evaluation of a new concept by a small number of early adopters can have a positive impact on its future dissemination. Auspiciously, the high evaluation of relevance of the ABP trend shapes to the physiological basis in this work may indicate that anesthesiologists have become aware of its potential usefulness during evaluation. Although this discussion is less empirical, it has implications in that an important aspect in the development of medical XAI was discussed in conjunction with the technology acceptance theory, providing insight into the direction of future research.

E. Limitations and Future Works

Despite the novelty of the model architecture and interpretable method proposed through this study, there are some limitations that need to be addressed in future research. The first limitation relates to the nature of the retrospective study setting. In particular, medical intervention bias may exist in ABP data, which may lead to predictive bias [5]. Hence, further studies in a prospective setting are encouraged to adequately address interventional biases induced when anesthesiologists respond to symptoms associated with the development of hypotension. In addition, the results of clinical investigations conducted in a retrospective setting showed the anesthesiologist's intention rather than the actual use of the proposed interpretable method. Therefore, it would be worthwhile to evaluate the practical use of the proposed method through follow-up studies in a prospective setting.

The second limitation concerns the computational inefficiency of learning ABP trend shapes using the Euclidean distance between the generated vectors and local intervals. Because the Euclidean distance assumes that the i^{th} position in one sequence is aligned with the exact same position in another sequence, the similarity between two sequence data is measured inflexibly [61] and may produce unnecessary shapes highly correlated with each other. This limitation may be alleviated by employing dynamic time warping (DTW), which allows a nonlinear alignment between two sequence datasets to account for the similarity of local shapes that are adjacent but different in locations [61], [62]. With the application of DTW in deep learning algorithms still in its infancy, further research on methods that can efficiently measure the similarity of sequence data with well-defined gradients [62], should be encouraged.

The last limitation is the lack of flexibility of the post-hoc interpretation, which is caused by the dependence of model-specific interpretation on the model structure [18]. Specifically, because the ABP trend shapes were generated based on 30 s of ABP in this study, the association between the hypotension development and ABP trends at different time lengths such as 40 s is not interpretable. As such, models must be redesigned to address the association between factors and the outcome based on different interests. Given the recent complexity of model structures in deep learning and the need for large amounts of data for training, training all models according to the interest of each interpretation is computationally inefficient. Therefore, a hasty generalization that model-specific interpretations are superior to model-agnostic interpretations should be avoided. Rather, studies on how the two methods can be combined and used together for a better interpretation of hypotension prediction should be encouraged.

VII. Conclusion

Given the clinical importance of monitoring hypotension during surgery, the recent emergence of XAI is expected to provide ground-breaking support in the medical field. The hypotension prediction model proposed in this study aims to focus on the local trends of ABP to interpret how certain shapes of local trends are associated with hypotension, along with a good predictive power. To facilitate ABP monitoring in clinical practice, we expect that more empirical validations of hypotension prediction models will be actively conducted in a prospective environment.

References

[1] T. G. Monk *et al.*, "Association between Intraoperative Hypotension and Hypertension and 30-day Postoperative Mortality in Noncardiac Surgery," *Anesthesiology*, vol. 123, no. 2, pp. 307–319, Aug. 2015, doi: 10.1097/ALN.0000000000000756.

[2] A. Gregory *et al.*, "Intraoperative Hypotension Is Associated With Adverse Clinical Outcomes After Noncardiac Surgery," *Anesthesia & Analgesia*, vol. 132, no. 6, pp. 1654–1665, Jun. 2021, doi: 10.1213/ANE.0000000000005250.

[3] G. M. Gurman, M. Klein, and N. Weksler, "Professional stress in anesthesiology: a review," *J Clin Monit Comput*, vol. 26, no. 4, pp. 329–335, Aug. 2012, doi: 10.1007/s10877-011-9328-7.

[4] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang, "The practical implementation of artificial intelligence technologies in medicine," *Nat Med*, vol. 25, no. 1, pp. 30–36, Jan. 2019, doi: 10.1038/s41591-018-0307-0.

[5] S. Lee *et al.*, "Deep learning models for the prediction of intraoperative hypotension," *British Journal of Anaesthesia*, vol. 126, no. 4, pp. 808–817, Apr. 2021, doi: 10.1016/j.bja.2020.12.035.

[6] F. Hatib *et al.*, "Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis," *Anesthesiology*, vol. 129, no. 4, pp. 663–674, Oct. 2018, doi: 10.1097/ALN.0000000000002300.

[7] S. Lee, M. Lee, S.-H. Kim, and J. Woo, "Intraoperative Hypotension Prediction Model Based on Systematic Feature Engineering and Machine Learning," *Sensors*, vol. 22, no. 9, Art. no. 9, Jan. 2022, doi: 10.3390/s22093108.

[8] S. Choe *et al.*, "Short-Term Event Prediction in the Operating Room (STEP-OP) of Five-Minute Intraoperative Hypotension Using Hybrid Deep Learning: Retrospective Observational Study and Model Development," *JMIR Medical Informatics*, vol. 9, no. 9, p. e31311, Sep. 2021, doi: 10.2196/31311.

[9] M. Cherifa, A. Blet, A. Chambaz, E. Gayat, M. Resche-Rigon, and R. Pirracchio, "Prediction of an Acute Hypotensive Episode During an ICU Hospitalization With a Super Learner Machine-Learning Algorithm:," *Anesthesia & Analgesia*, vol. 130, no. 5, pp. 1157–1166, May 2020, doi: 10.1213/ANE.0000000000004539.

[10] W. H. van der Ven, D. P. Veelo, M. Wijnberge, B. J. P. van der Ster, A. P. J. Vlaar, and B. F. Geerts, "One of the first validations of an artificial intelligence algorithm for clinical use: The impact on intraoperative hypotension prediction and clinical decision-making," *Surgery*, vol. 169, no. 6, pp. 1300–1303, Jun. 2021, doi: 10.1016/j.surg.2020.09.041.

[11] K. Maheshwari *et al.*, "Hypotension Prediction Index for Prevention of Hypotension during Moderate- to High-risk Noncardiac Surgery," *Anesthesiology*, vol. 133, no. 6, pp. 1214–1222, Dec. 2020, doi: 10.1097/ALN.0000000000003557.

[12] S. Kundu, "AI in medicine must be explainable," *Nat Med*, vol. 27, no. 8, pp. 1328–1328, Aug. 2021, doi: 10.1038/s41591-021-01461-z.

[13] B. Lamia, D. Chemla, C. Richard, and J.-L. Teboul, "Clinical review: Interpretation of arterial pressure wave in shock states," *Crit Care*, vol. 9, no. 6, p. 601, 2005, doi: 10.1186/cc3891.

[14] A. R. Kang *et al.*, "Development of a prediction model for hypotension after induction of anesthesia using machine learning," *PLoS ONE*, vol. 15, no. 4, p. e0231172, Apr. 2020, doi: 10.1371/journal.pone.0231172.

[15] S. Kendale, P. Kulkarni, A. D. Rosenberg, and J. Wang, "Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension," *Anesthesiology*, vol. 129, no. 4, pp. 675–688, Oct. 2018, doi: 10.1097/ALN.0000000000002374.

[16] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv:1712.09923 [cs, stat]*, Dec. 2017, Accessed: Sep. 23, 2021. [Online]. Available: http://arxiv.org/abs/1712.09923

[17] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava, "How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 4211–4222. Accessed: Jul. 05, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/2c29d89cc56cdb191c60db2f0bae796b-Abstract.html

[18] A. Carrillo, L. F. Cantú, and A. Noriega, "Individual Explanations in Machine Learning Models: A Survey for Practitioners." arXiv, Apr. 11, 2021. doi: 10.48550/arXiv.2104.04144.

[19] Q. Ai and L. Narayanan.R, "Model-agnostic vs. Model-intrinsic Interpretability for Explainable Product Search," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, New York, NY, USA: Association for Computing Machinery, 2021, pp. 5–15. Accessed: Jul. 04, 2022. [Online]. Available: https://doi.org/10.1145/3459637.3482276

[20] N. Prasad and K. Palla, "The Role of Context in the Prediction of Acute Hypotension in Critical Care," p. 6.

[21] M. W. Kang *et al.*, "Machine learning model to predict hypotension after starting continuous renal replacement therapy," *Sci Rep*, vol. 11, no. 1, p. 17169, Dec. 2021, doi: 10.1038/s41598-021-96727-4.

[22] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, "Explainable AI decision model for ECG data of cardiac disorders," *Biomedical Signal Processing and Control*, vol. 75, p. 103584, May 2022, doi: 10.1016/j.bspc.2022.103584.

[23] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "An Empirical Evaluation of AI Deep Explainable Tools," in *2020 IEEE Globecom Workshops (GC Wkshps*, Dec. 2020, pp. 1–6. doi: 10.1109/GCWkshps50303.2020.9367541.

[24] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen, "Discrete wavelet transform-based time series analysis and mining," *ACM Comput. Surv.*, vol. 43, no. 2, pp. 1–37, Jan. 2011, doi: 10.1145/1883612.1883613.

[25] R. Madan, S. K. Singh, and N. Jain, "Signal Filtering Using Discrete Wavelet Transform," vol. 2, no. 3, p. 4, 2009.

[26] Xiao-Ping Zhang, Li-Sheng Tian, and Ying-Ning Peng, "From the wavelet series to the discrete wavelet transform-the initialization," *IEEE Trans. Signal Process.*, vol. 44, no. 1, pp. 129–133, Jan. 1996, doi: 10.1109/78.482020.

[27] S. P. Nanavati and P. K. Panigrahi, "Wavelet transform: A new mathematical microscope," *Reson*, vol. 9, no. 3, pp. 50–64, Mar. 2004, doi: 10.1007/BF02834988.

[28] S. M. S. Alam and T. Hasan, "Performance Analysis of FIR Filter Design by Using Optimal, Blackman Window and Frequency Sampling Methods," vol. 10, no. 01, p. 6.

[29] A. A. Eleti and A. R. Zerek, "FIR digital filter design by using windows method with MATLAB," in *14th International Conference on Sciences and Techniques of Automatic Control & Computer Engineering - STA'2013*, Sousse, Dec. 2013, pp. 282–287. doi: 10.1109/STA.2013.6783144.

[30] M. Akrout, C. Wilson, P. C. Humphreys, T. Lillicrap, and D. Tweed, "Deep Learning without Weight Transport," *arXiv:1904.05391 [cs, stat]*, Jan. 2020, Accessed: Jul. 05, 2021. [Online]. Available: http://arxiv.org/abs/1904.05391

[31] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, Paris, France, 2009, p. 947. doi: 10.1145/1557019.1557122.

[32] Z. Niu, "A review on the attention mechanism of deep learning," p. 15, 2021.

[33] P. Michel, O. Levy, and G. Neubig, "Are Sixteen Heads Really Better than One?," p. 11.

[34] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, Feb. 2018, doi: 10.1016/j.dsp.2017.10.011.

[35] H.-C. Lee and C.-W. Jung, "Vital Recorder—a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices," *Sci Rep*, vol. 8, no. 1, p. 1527, Dec. 2018, doi: 10.1038/s41598-018-20062-4.

[36] D. Zhu and L. Lu, "Resampling method of computed order tracking based on time-frequency scaling property of fourier transform," in *2015 International Conference on Estimation, Detection and Information Fusion (ICEDIF)*, Jan. 2015, pp. 248–253. doi: 10.1109/ICEDIF.2015.7280200.

[37] H. Lee *et al.*, "Deep Learning Model for Real-Time Prediction of Intradialytic Hypotension," *CJASN*, vol. 16, no. 3, pp. 396–406, Mar. 2021, doi: 10.2215/CJN.09280620.

[38] M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *IJMLC*, pp. 224–228, 2013, doi: 10.7763/IJMLC.2013.V3.307.

[39] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," in *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Oct. 2019, pp. 135–139. doi: 10.1109/ICCSNT47585.2019.8962457.

[40] V. N. L. Duy, S. Iwazaki, and I. Takeuchi, "Quantifying Statistical Significance of Neural Network Representation-Driven Hypotheses by Selective Inference," *arXiv:2010.01823 [cs, stat]*, Oct. 2020, Accessed: Jul. 02, 2021. [Online]. Available: http://arxiv.org/abs/2010.01823

[41] A. Ben-Israel, "A Newton-Raphson method for the solution of systems of equations," *Journal of Mathematical Analysis and Applications*, vol. 15, no. 2, pp. 243–252, Aug. 1966, doi: 10.1016/0022-247X(66)90115-6.

[42] S. A. Czepiel, "Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation."

[43] G. A. Morgan, J. J. Vaske, J. A. Gliner, and R. J. Harmon, "Logistic Regression and Discriminant Analysis: Use and Interpretation," *Journal of the American Academy of Child & Adolescent Psychiatry*,

vol. 42, no. 8, pp. 994–997, Aug. 2003, doi: 10.1097/01.CHI.0000046896.27264.0F.

[44] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Jun. 07, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[45] S. Lundberg, “slundberg/shap.” Jun. 07, 2022. Accessed: Jun. 07, 2022. [Online]. Available: https://github.com/slundberg/shap

[46] E. Kilsdonk, L. W. Peute, and M. W. M. Jaspers, “Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis,” *International Journal of Medical Informatics*, vol. 98, pp. 56–64, Feb. 2017, doi: 10.1016/j.ijmedinf.2016.12.001.

[47] Y.-S. Jeong *et al.*, “Prediction of Blood Pressure after Induction of Anesthesia Using Deep Learning: A Feasibility Study,” *Applied Sciences*, vol. 9, no. 23, p. 5135, Nov. 2019, doi: 10.3390/app9235135.

[48] F. Michard *et al.*, “Relation between Respiratory Changes in Arterial Pulse Pressure and Fluid Responsiveness in Septic Patients with Acute Circulatory Failure,” *Am J Respir Crit Care Med*, vol. 162, no. 1, pp. 134–138, Jul. 2000, doi: 10.1164/ajrccm.162.1.9903035.

[49] L. Tanzi, P. Piazzolla, and E. Vezzetti, “Intraoperative surgery room management: A deep learning perspective,” *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 5, p. e2136, 2020, doi: 10.1002/rcs.2136.

[50] M. A. Poole and P. N. O’Farrell, “The Assumptions of the Linear Regression Model,” *Transactions of the Institute of British Geographers*, no. 52, pp. 145–158, 1971, doi: 10.2307/621706.

[51] D. M. Berwick, “Disseminating Innovations in Health Care,” *JAMA*, vol. 289, no. 15, p. 1969, Apr. 2003, doi: 10.1001/jama.289.15.1969.

[52] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, “Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey.” arXiv, Apr. 02, 2021. Accessed: Nov. 24, 2022. [Online]. Available: http://arxiv.org/abs/2104.00950

[53] R. Assaf, I. Giurgiu, F. Bagehorn, and A. Schumann, “MTEX-CNN: Multivariate Time Series EXplanations for Predictions with Convolutional Neural Networks,” in *2019 IEEE International Conference on Data Mining (ICDM)*, Jan. 2019, pp. 952–957. doi: 10.1109/ICDM.2019.00106.

[54] K. S. Choi, S. H. Choi, and B. Jeong, “Prediction of IDH genotype in gliomas with dynamic susceptibility contrast perfusion MR imaging using an explainable recurrent neural network,” *Neuro-Oncology*, vol. 21, no. 9, pp. 1197–1209, Sep. 2019, doi: 10.1093/neuonc/noz095.

[55] F. Tung, S. Chang, and C. Chou, “An extension of trust and TAM model with IDT in the adoption of the electronic logistics information system in HIS in the medical industry,” *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 324–335, May 2008, doi: 10.1016/j.ijmedinf.2007.06.006.

[56] V. Venkatesh and F. D. Davis, “A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies,” *Management Science*, vol. 46, no. 2, pp. 186–204, Feb. 2000, doi: 10.1287/mnsc.46.2.186.11926.

[57] M. R. Tonelli, “Integrating evidence into clinical practice: an alternative to evidence-based approaches,” *Journal of Evaluation in Clinical Practice*, vol. 12, no. 3, pp. 248–256, 2006, doi: 10.1111/j.1365-2753.2004.00551.x.

[58] A. Rai, “Explainable AI: from black box to glass box,” *J. of the Acad. Mark. Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.

[59] M. Langer *et al.*, “What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research,” *Artificial Intelligence*, vol. 296, p. 103473, Jul. 2021, doi: 10.1016/j.artint.2021.103473.

[60] E. M. Rogers, in *Diffusion of innovations*, 4th ed., New York, 1995.

[61] Z. Yu, Z. Niu, W. Tang, and Q. Wu, “Deep Learning for Daily Peak Load Forecasting–A Novel Gated Recurrent Neural Network Combining Dynamic Time Warping,” *IEEE Access*, vol. 7, pp. 17184–17194, 2019, doi: 10.1109/ACCESS.2019.2895604.

[62] X. Cai, T. Xu, J. Yi, J. Huang, and S. Rajasekaran, “DTWNet: a Dynamic Time Warping Network,” p. 11.

Supplementary Document for Intraoperative Hypotension Prediction Based on Features Automatically Generated Within an Interpretable Deep Learning Model

Eugene Hwang, Yong-Seok Park, Jin-Young Kim, Sung-Hyuk Park, Junetae Kim, and Sung-Hoon Kim

I. Specifications of the Proposed Model

A. Discrete Wavelet Transform (DWT) Layers

To extract the low-frequency components of arterial blood pressure (ABP), a convolutional filter with a length of 51 values is operated as a multiplicative form of the *sinc* function and the Blackman window, as presented in Fig. 1. The frequency cutoff value, which determines the level of frequency to be filtered, was trained within the deep learning model.

With the utilization of the identified filter, Fig. 2 demonstrates how the input data are processed within the first section of the layers in the proposed model. Since the ABP records utilized in this model were processed at 100 Hz, 90 s of ABP records consisted of 9,000 values. Starting with a raw vector with 9,000 values, five levels of the discrete wavelet were performed to obtain a compressed vector with 232 values (LP5Vector). The morphological change at each of the five levels is shown in Fig. 2. Next, the LP5Vector is expanded in dimension by 208 local intervals with a length of 50, which is illustrated as LP5Vector3D.

B. Shape Similarity Layers

With the generation of 30 local shape vectors as a matrix, the similarity between each generated feature and each local interval of the compressed ABP was calculated, as illustrated in Fig. 3. Distance3D and Simiarlity3D illustrate the distance and similarity, respectively, between certain intervals and generated features.

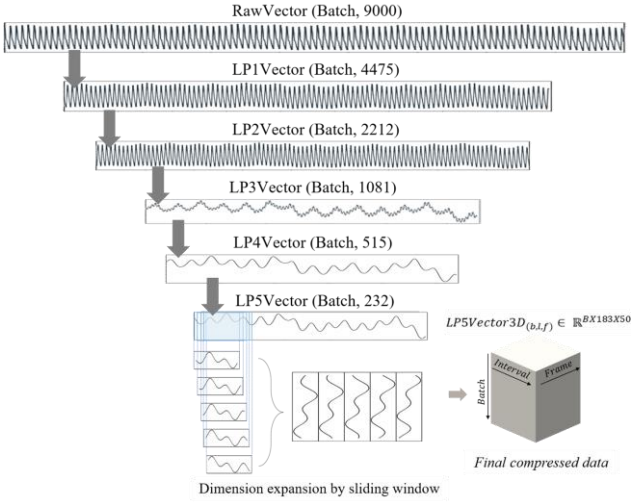


Fig. 2. Example of tensor flows via discrete wavelet transform layers.

C. Local Importance Layers

Fig. 4 shows how importance is weighted within the compressed ABP shape. First, five parameters were trained to weight certain local intervals that were important in hypotension prediction. A Gaussian distribution, which weights local intervals, was formed based on the five parameters as μ. Next, the probability values of five Gaussian distributions over the compressed vector were summed as a single set of weight values. Subsequently, multiplication of the weight and similarity values was performed along the shape axis. At this point, the maximum value along the shape axis was only considered during the prediction.

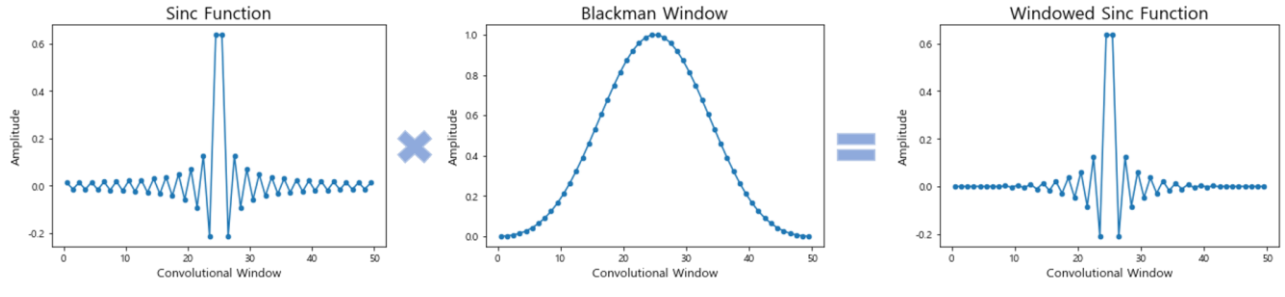


Fig. 1. Convolutional filter employed for the discrete wavelet transform.

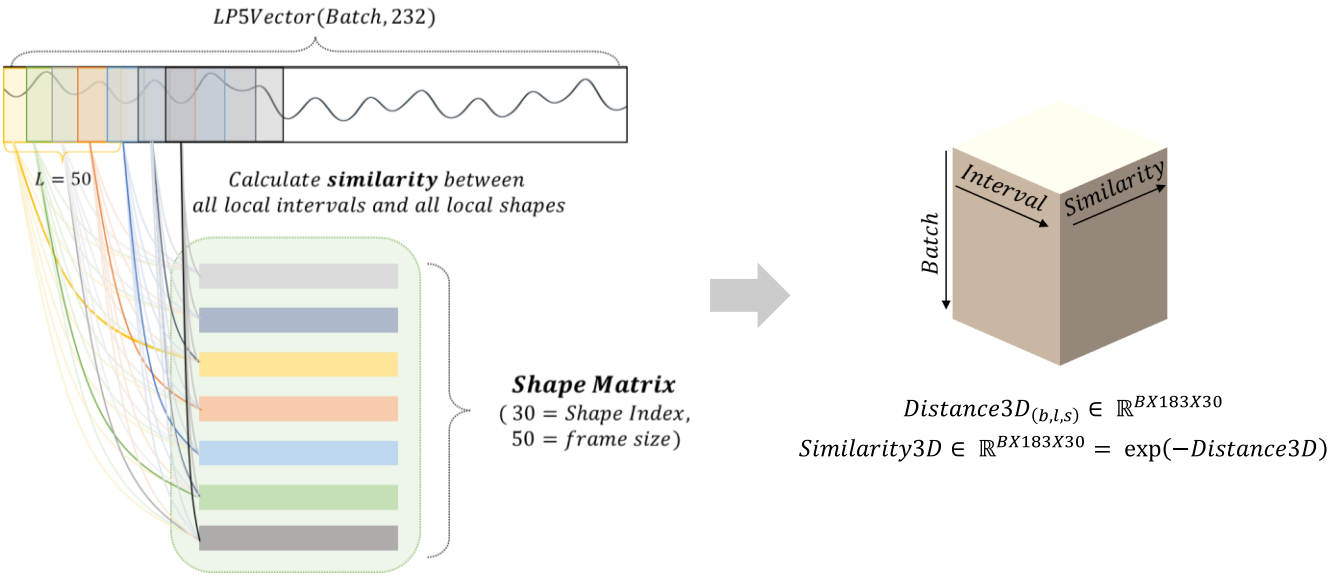


Fig. 3. Local shape similarity layers.

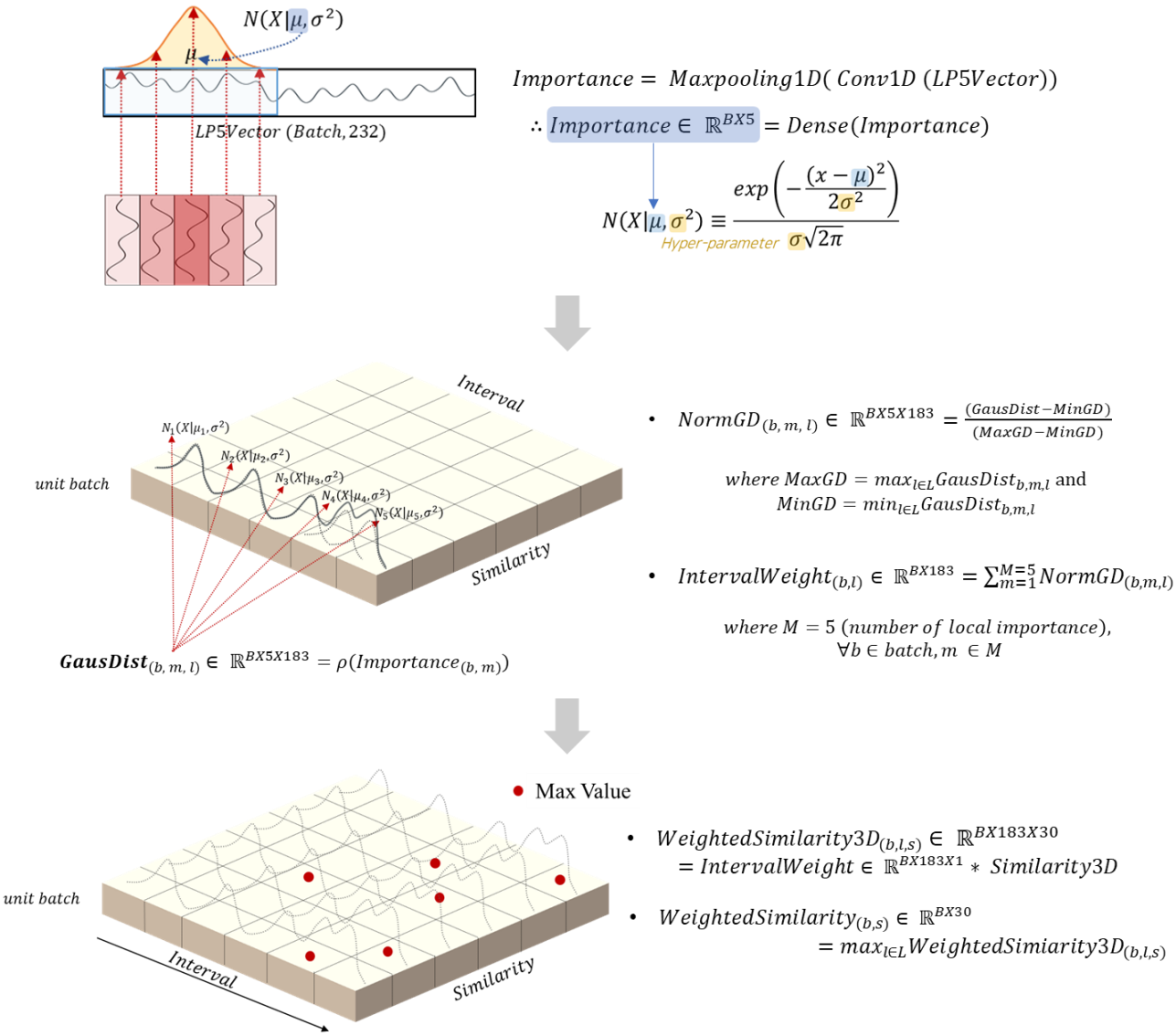


Fig. 4. Local importance layers.

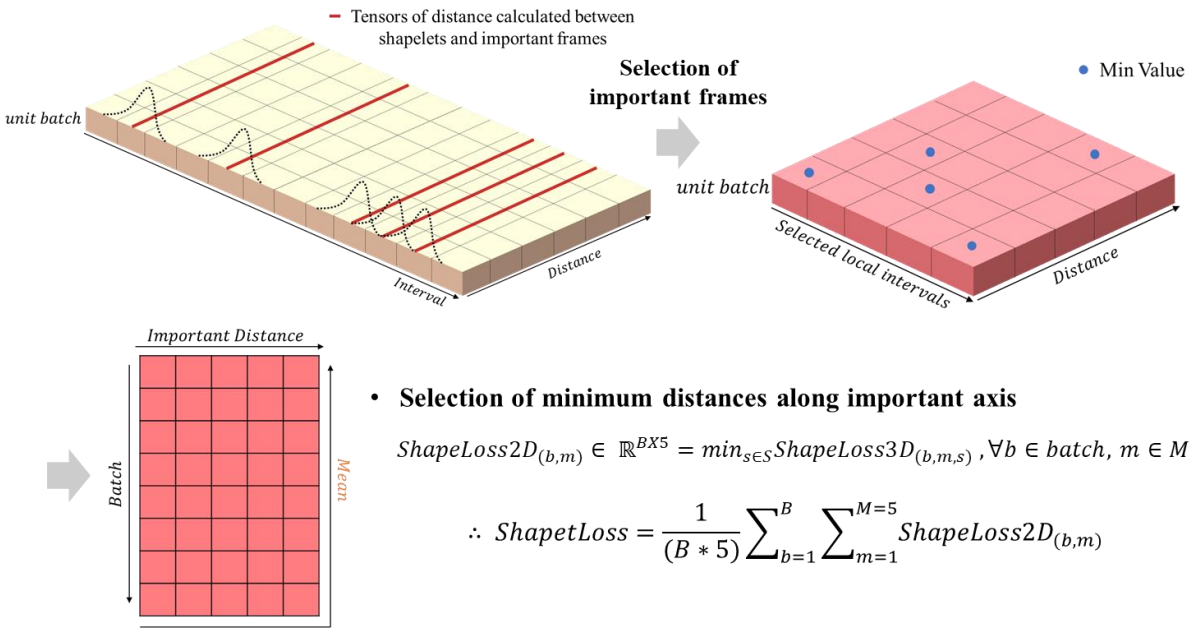


Fig. 5. Shape loss function.

D. Objective Function

Fig. 5 presents the formation of the shape loss function for model training. First, the distance values in only five specific local intervals influential in hypotension prediction were considered. Next, only the minimum value along the distance axis in a matrix was considered during the back-propagation update. Finally, all five values at batch-level were averaged into a single loss, which was minimized during model training.

II. Specifications of Ablation Models

Fig. 6 provides the model architectures of two ablation experiments. Layers that were removed or modified from the proposed model are colored in gray.

In Ablation model 1 (Fig. 6a), the local importance layers were removed. Accordingly, tensors from LP5Vector were propagated directly to LP5Vector3D without going through any layers for computing IntervalWeight (μ, ρ). As a result, only

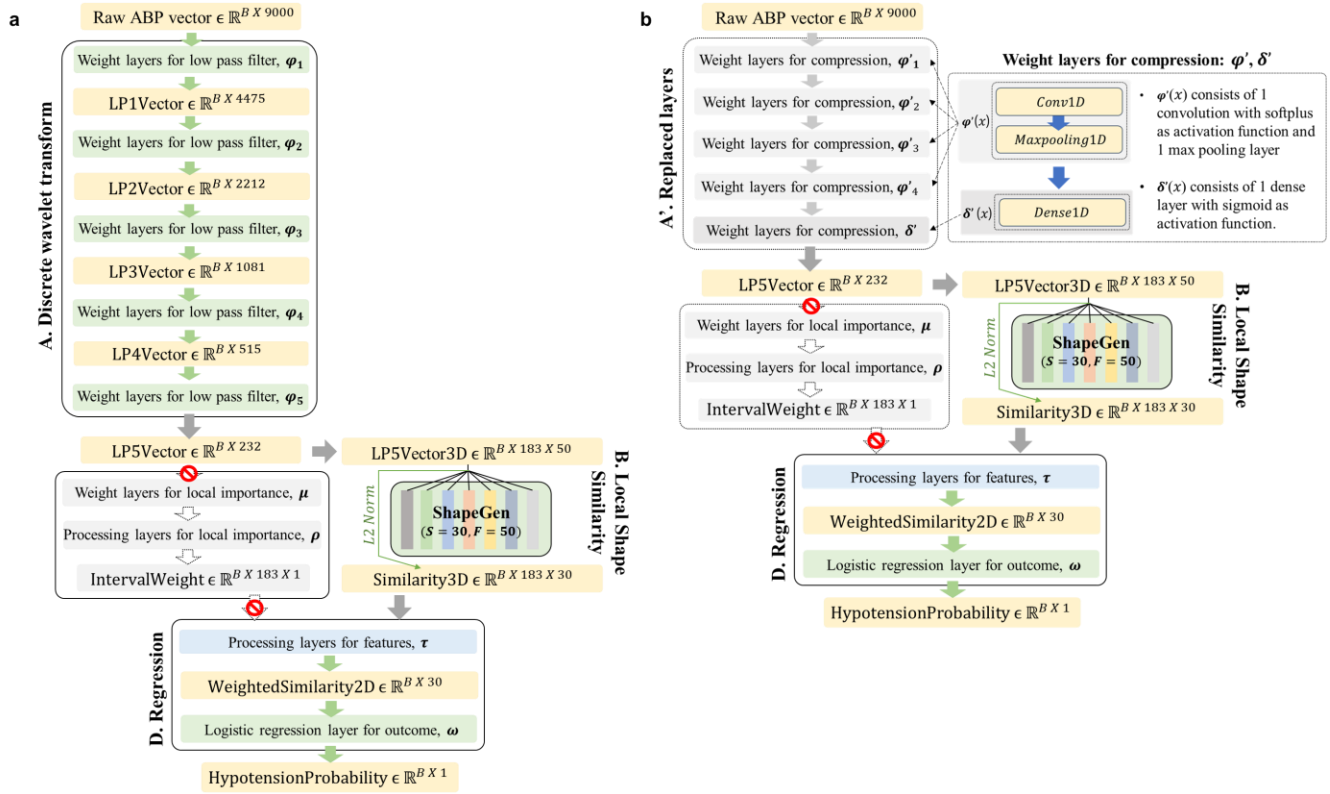


Fig. 6. Architecture of the models for ablation study: a) Ablation model 1 and b) Ablation model 2.

TABLE I
DATA COLLECTION FOR MODEL TRAINING AND INTERNAL VALIDATION

Surgery duration	Total number of cases	Number of hypotension samples	Number of non-hypotension samples	Hypotension rate
0 ~ 3 hours	3,181	75,696	128,685	0.3703
3 ~ 5 hours	3,244	157,056	215,030	0.4221
5 ~ hours	4,029	475,940	496,520	0.4894

tensors from Similarity3D were propagated to the processing layers for features (τ).

In Ablation model 2 (Fig. 6b), the DWT layers were modified in addition to the removal of local importance layers. Herein, five weight layers for low pass filter (φ) in the proposed model were replaced with five weight layers for compression (φ'). Finally, one dense layer (δ') was added to obtain a vector of the same dimension as LP5Vector for further processing.

III. SPECIFICATIONS OF DATA COLLECTION IN AMC DATASET

ABP records of 10,454 patients in Asan Medical Center are stratified with surgical durations of less than 3 hours, 3–5 hours, and greater than 5 hours, as shown in Table I. For each surgery duration, the total number of cases, and the number of hypotension and non-hypotension samples after pre-processing are listed. Hypotension rate indicates the ratio of hypotension samples out of the total number of samples.

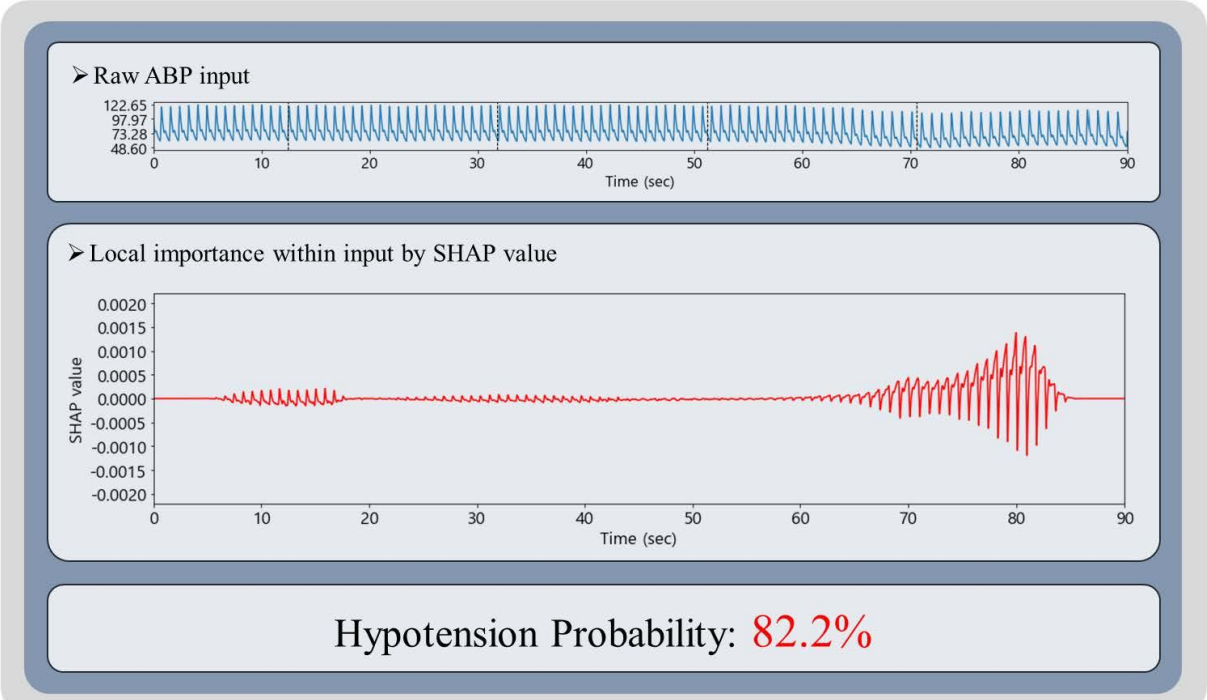
IV. SPECIFICATIONS OF CLINICAL SURVEY

A. Questionnaire and Figures Presented to Participants

The questionnaires and figures presented to the participants were provided in each of the three categories, as shown in Fig. 7, Fig. 8, and Fig. 9. Although this supplementary material contains each sample of the visual summary, the actual survey containing 10 samples for each category was presented.

Thank you for participating in this survey.
The purpose of this survey is to compare and evaluate the clinical feasibility and interpretability of three visual summaries that provide information on intraoperative hypotension occurrence.
Please take a look at the interpretation of the model presented for the predicted result of 10 samples and answer the following questions.

1. Visual summary of Category 1: SHAP
As demonstrated below, 90 s of ABP input, SHAP values, and hypotension probability (%) after 10 min are provided. *SHAP value indicates the magnitude of the impact the corresponding timepoint of ABP input has on the prediction.



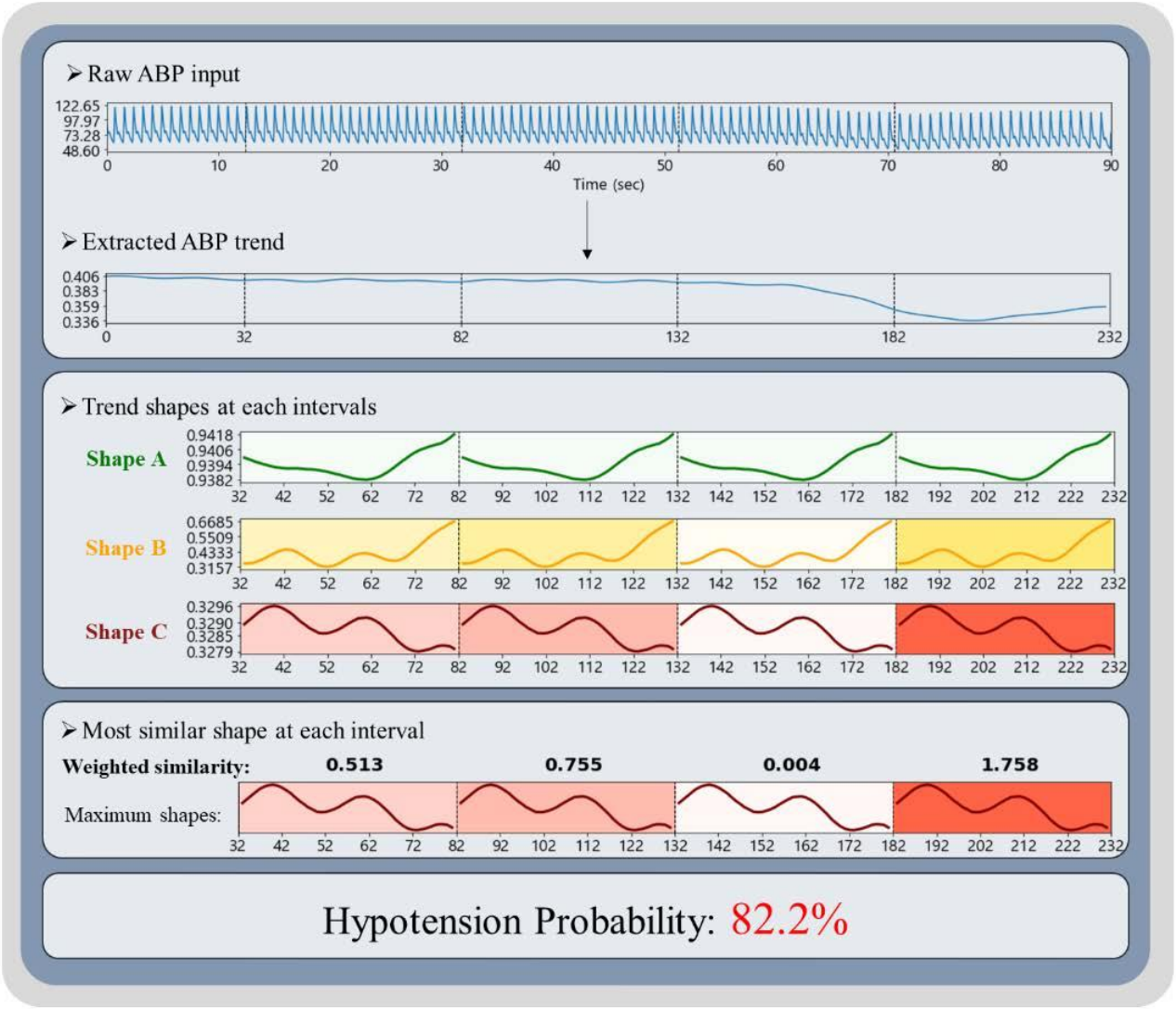
Consider the following 10 samples that predicted hypotension as a result of Category 1.

1) Do you think the results presented in the visual summary of Category 1 are clinically valid?
2) Do you think the SHAP value provided in Category 1 is clinically useful? (Do you think the visual summary of Category 1 provides useful additional information compared to providing hypotension probability alone?)
3) If the visual summary of Category 1 reports high probability of hypotension while actually monitoring a patient under anesthesia, are you willing to follow the report and take action to prevent hypotension?

Fig. 7. Questionnaire and sample visual summary of Category 1.

2. Visual summary of Category 2: Trend shape similarity

As demonstrated below, 90 s of ABP input, extracted ABP trend, information on trend similarity with Shapes A, B, and C, and hypotension probability (%) after 10 min are provided. * If the trend shape is similar to Shapes A or B, hypotension is less likely to occur, and if it is similar to Shape C, hypotension is more likely to occur.



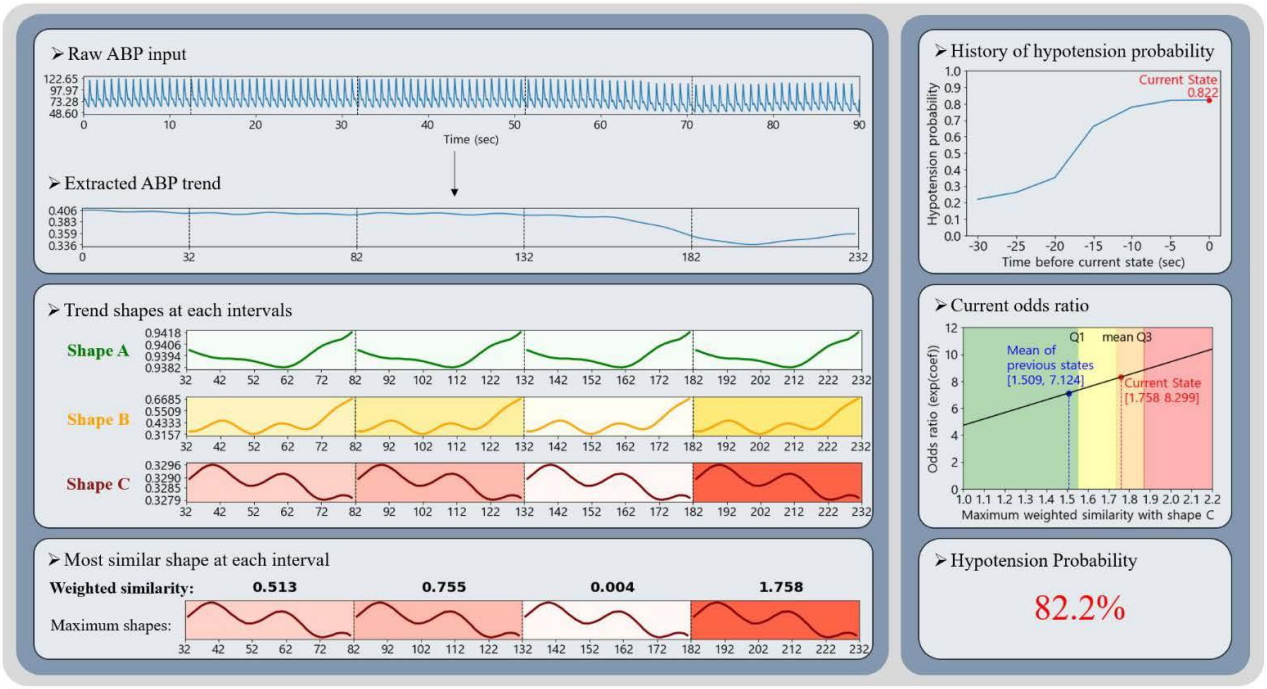
Consider the following 10 samples that predicted hypotension as a result of Category 2.

1) Do you think the results presented in the visual summary of Category 2 are clinically valid?
2) Do you think that the extracted ABP trend and information on trend similarity with Shapes A, B, and C are clinically useful? (Do you think the visual summary of Category 2 provides useful additional information compared to providing hypotension probability alone?)
3) If the visual summary of Category 2 reports high probability of hypotension while actually monitoring a patient under anesthesia, are you willing to follow the report and take action to prevent hypotension?

Fig. 8. Questionnaire and sample visual summary of Category 2.

3. Visual summary of Category 3: Odds ratio of hypotension occurrence and history of predicted probabilities in addition to trend shape similarity

As demonstrated below, history of odds ratio and hypotension probability over 30 s are provided in addition to the information provided in Category 2.



Consider the following 10 cases that predicted hypotension as a result of Category 3.

1) Do you think the results presented in the visual summary of Category 3 are clinically valid?
2) Do you think the history of odds ratio and hypotension probability over 30 seconds additionally provided in Category 3 are clinically useful? (Do you think the visual summary of Category 3 provides useful additional information compared to providing hypotension probability alone?)
3) If the visual summary of Category 3 reports high probability of hypotension while actually monitoring a patient under anesthesia, are you willing to follow the report and take action to prevent hypotension?

Fig. 9. Questionnaire and sample visual summary of Category 3.

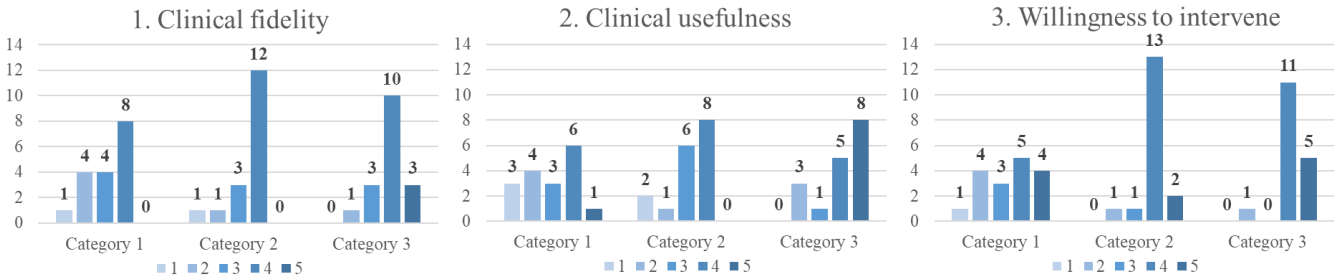


Fig. 10. Distribution of ratings for clinical survey. X-axis for each category indicates the score from 1 to 5 and y-axis indicates the number of patients. The number of patients rated for each 5-point Likert scale are labelled for all assessments.

B. Distribution of Ratings for Clinical Survey

Fig. 10 demonstrates the detailed results of the survey by reporting the number of participants who rated from 1 to 5 for each aspect.

For all three aspects, Categories 2 and 3 showed a dramatic increase compared to Category 1. First, in terms of clinical fidelity, 8 people rated at least 4 for Category 1, whereas 12 and 13 people rated at least 4 for Categories 2 and 3, respectively. Second, in terms of usefulness, 7 people rated at least 4 for Category 1, whereas 14 and 13 people rated at least 4 for Categories 2 and 3, respectively. Finally, in terms of willingness to intervene, 9 people rated at least 4 for Category 1, whereas 15 and 16 people rated at least 4 for Categories 2 and 3, respectively.