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

***Abstract*—The monitoring of arterial blood pressure (ABP) in anesthetized patients is crucial for preventing hypotension, which can lead to adverse clinical outcomes. Several efforts have been devoted to develop artificial intelligence-based hypotension prediction indices. However, the use of such indices is limited because they may not provide a compelling interpretation of the association between the predictors and hypotension. Herein, an interpretable deep learning model is developed that forecasts hypotension occurrence 10 min before a given 90 s ABP record. Internal and external validations of the model performance show the area under the receiver operating characteristic curves of 0.9145 and 0.9035, respectively. Furthermore, the hypotension prediction mechanism can be physiologically interpreted using the predictors automatically generated from the proposed model for representing ABP trends. Finally, the applicability of a deep learning model with high accuracy is demonstrated, thereby providing an interpretation of the association between ABP trends and hypotension in clinical practice.**

***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 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 critical for anesthesiologists to minimize the risk of IOH occurrence [1]. Despite this, preemptive measures may not always be taken prior to the onset of IOH, because anesthesiologists may be occupied with responding to unexpected events in a patient during surgery in real time [3]. In such busy environments, artificial intelligence (AI) can minimize the burden on anesthesiologists by predicting the occurrence of IOH [4].

Various attempts have been devoted to developing AI-based IOH prediction models [5]–[9], some of which (e.g., hypotension prediction index (HPI) [6]) have become commercially available. Although these models have successfully provided predictive results, their practical application has several limitations. First, existing commercialized models may not provide a convincing explanation of the IOH predictive mechanism [10]. The lack of interpretability may lead clinicians to ignore or passively respond to warnings provided by the model [10], [11]. Given that experts are more likely to react based on their knowledge and experience rather than on concepts that are not yet well-proven or remain 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 is yet to be sufficiently evaluated [5]. A clinically valid interpretation is crucial to encourage anesthesiologists to actively intervene based on the model [4], [11]. Despite its importance, in-depth discussions of this aspect are limited [13]. Third, the input features utilized in 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 [14], the ABP trend that specifies changes in intravascular volume status over time may be appropriate for both prediction and interpretation. However, raw ABP records and microscopic features of the ABP waveform employed as predictors in previous models may inaccurately reflect the ABP trend [5]–[7].

To address these limitations, we propose a deep learning model that predicts IOH 10 min before its onset with a 90 s ABP record sampled at a rate of 100 Hz. One of the main attributes of this approach is the development of a framework that makes deep learning-based interpretations compatible with statistical hypothesis testing. Accordingly, the deep learning model applied in this framework was designed to be decomposed into two broad parts (Fig. 1). First, the generation part of the model supports statistical operationalization by generating ABP trend shapes, which are the IOH predictors. Second, the regression

---

Manuscript received February 9, 2022; revised August 16, 2022; revised December 14, 2022; revised February 19, 2023. This research was supported by a grant from the National Cancer Center of Korea (grant number: 2010380-2, 2210410-1), and by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C2383, HI18C0022). *Corresponding authors: Junetae Kim (*lyjune0070@gmail.com*) and Sung-Hoon Kim* (shkimans@amc.seoul.kr). Eugene Hwang and Yong-Seok Park contributed equally to this paper.

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

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 the Graduate School of Cancer Science and Policy and the Healthcare AI Team, Healthcare Platform Center, National Cancer Center, Goyang-si, Gyeonggi-do, Republic of Korea.

The source code and synthetic data for this work are available at https://github.com/JunetaeKim/DWT-HPI.

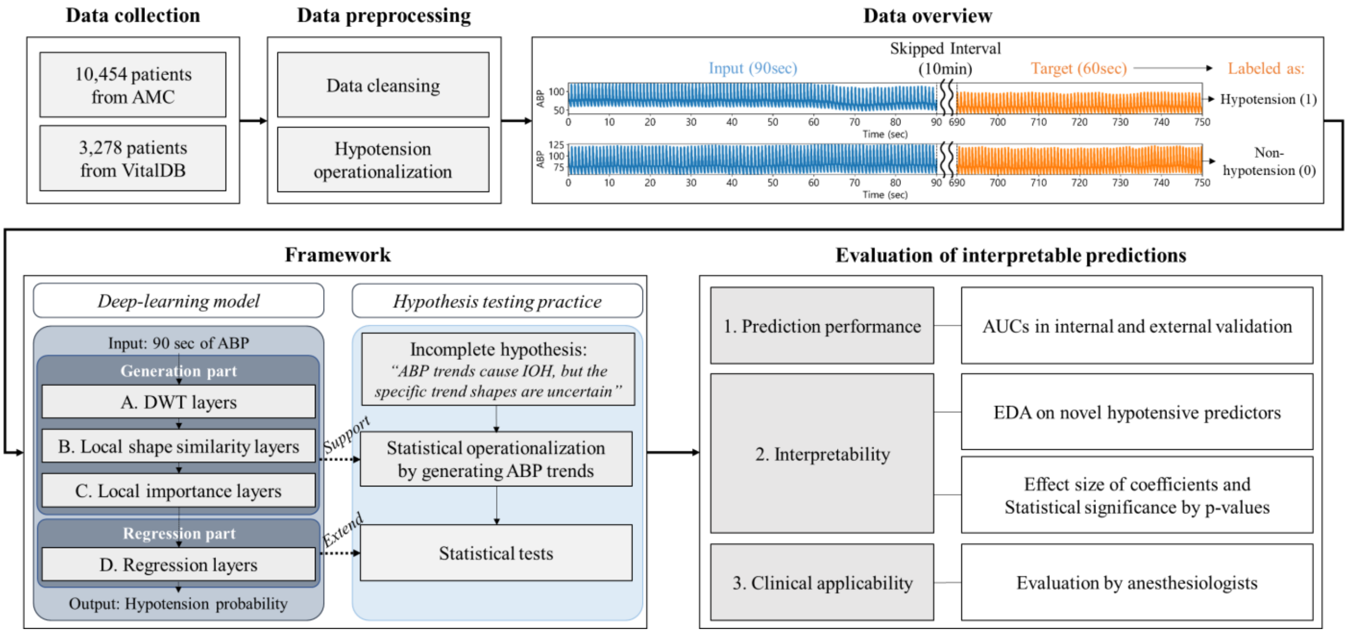

Data collection
10,454 patients from AMC
3,278 patients from VitalDB
Data preprocessing
Data cleansing
Hypotension operationalization
Data overview
Input (90sec)
Skipped Interval (10min)
Target (60sec)
Labeled as:
Hypotension (1)
Non-hypotension (0)
ABP
Time (sec)
Framework
Deep-learning model
Input: 90 sec of ABP
Generation part
A. DWT layers
B. Local shape similarity layers
C. Local importance layers
Regression part
D. Regression layers
Output: Hypotension probability
Hypothesis testing practice
Incomplete hypothesis:
"ABP trends cause IOH, but the specific trend shapes are uncertain"
Support
Statistical operationalization by generating ABP trends
Extend
Statistical tests
Evaluation of interpretable predictions
1. Prediction performance
AUCs in internal and external validation
2. Interpretability
EDA on novel hypotensive predictors
Effect size of coefficients and Statistical significance by p-values
3. Clinical applicability
Evaluation by anesthesiologists

Fig. 1. Research framework.

part extends the generated predictors to be used in a post-hoc statistical analysis for significance testing. Another attribute of this study is the generation of AI-based predictors that can address the physiological basis of hypotension and enhance its clinical fidelity. Because anesthesiologists are likely to be receptive to predictors that provide familiar meaning [11], clinically valid predictors may facilitate their adoption in clinical practice.

The three main contributions of this study can be summarized 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 using 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 operationalized features to be used in both deep learning-based predictions and statistical tests.
- *In-depth evaluation of the interpretable method:* The proposed method was evaluated multidimensionally through a benchmark test using an existing representative method, feedback from anesthesiologists, and theoretical discussions.

The remainder of this paper is organized as follows (Fig. 1). Section 2 introduces related studies on existing IOH prediction models. Section 3 describes the architecture and mathematical details of the model. Section 4 presents the experimental settings. Section 5 presents the results in terms of predictive performance, interpretability, and applicability. Section 6 discusses the results, along with several implications and limitations. Finally, Section 7 provides concluding remarks regarding this research.

II. RELATED WORK

A. Existing Hypotension Prediction Models

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

In addition to ABP-based features, various other clinical features have been used in machine learning algorithms predicting IOH [9], [16], [17]. For instance, the authors of [9] employed features extracted from physiological signals, time-evolving treatment characteristics, and baseline characteristics to predict the IOH 10 min in advance. In [16] and [17], the authors predicted post-induction hypotension (IOH occurrence during the first 20 min after anesthesia) using features from preoperative medications, medical comorbidities, induction medications, and intraoperative vital signs.

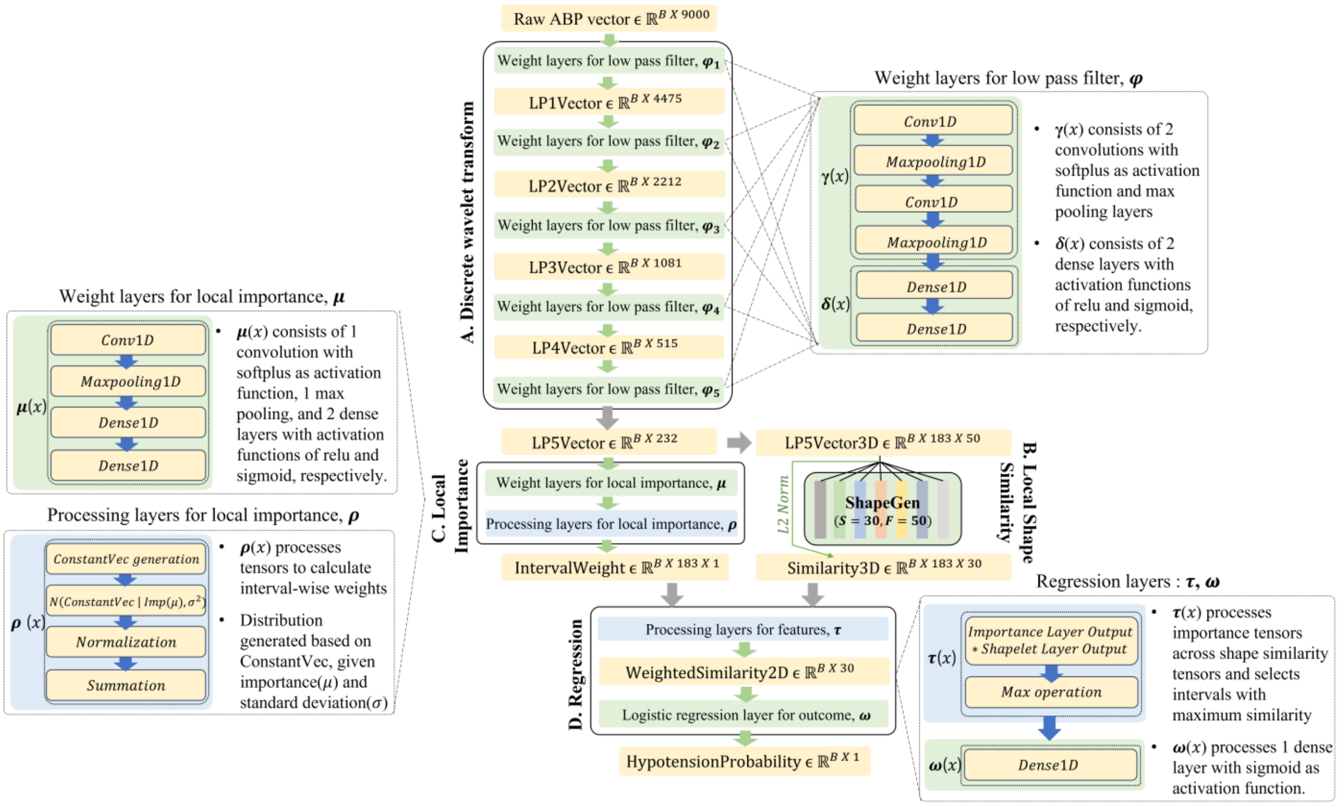

Raw ABP vector ϵ $\mathbb{R}^{B\ X\ 9000}$
A. Discrete wavelet transform
Weight layers for low pass filter, $\boldsymbol{\varphi_1}$
LP1Vector ϵ $\mathbb{R}^{B\ X\ 4475}$
Weight layers for low pass filter, $\boldsymbol{\varphi_2}$
LP2Vector ϵ $\mathbb{R}^{B\ X\ 2212}$
Weight layers for low pass filter, $\boldsymbol{\varphi_3}$
LP3Vector ϵ $\mathbb{R}^{B\ X\ 1081}$
Weight layers for low pass filter, $\boldsymbol{\varphi_4}$
LP4Vector ϵ $\mathbb{R}^{B\ X\ 515}$
Weight layers for low pass filter, $\boldsymbol{\varphi_5}$
LP5Vector ϵ $\mathbb{R}^{B\ X\ 232}$
Weight layers for low pass filter, $\boldsymbol{\varphi}$
$\boldsymbol{\gamma}(x)$
Conv1D
Maxpooling1D
Conv1D
Maxpooling1D
$\boldsymbol{\delta}(x)$
Dense1D
Dense1D
• $\boldsymbol{\gamma}(x)$ consists of 2 convolutions with softplus as activation function and max pooling layers
• $\boldsymbol{\delta}(x)$ consists of 2 dense layers with activation functions of relu and sigmoid, respectively.
LP5Vector3D ϵ $\mathbb{R}^{B\ X\ 183\ X\ 50}$
L2 Norm
ShapeGen
($S = 30, F = 50$)
Similarity3D ϵ $\mathbb{R}^{B\ X\ 183\ X\ 30}$
B. Local Shape Similarity
C. Local Importance
Weight layers for local importance, $\boldsymbol{\mu}$
Processing layers for local importance, $\boldsymbol{\rho}$
IntervalWeight ϵ $\mathbb{R}^{B\ X\ 183\ X\ 1}$
Weight layers for local importance, $\boldsymbol{\mu}$
$\boldsymbol{\mu}(x)$
Conv1D
Maxpooling1D
Dense1D
Dense1D
• $\boldsymbol{\mu}(x)$ consists of 1 convolution with softplus as activation function, 1 max pooling, and 2 dense layers with activation functions of relu and sigmoid, respectively.
Processing layers for local importance, $\boldsymbol{\rho}$
$\boldsymbol{\rho}\ (x)$
ConstantVec generation
$N(ConstantVec \mid Imp(\mu), \sigma^2)$
Normalization
Summation
• $\boldsymbol{\rho}(x)$ processes tensors to calculate interval-wise weights
• Distribution generated based on ConstantVec, given importance(μ) and standard deviation(σ)
D. Regression
Processing layers for features, $\boldsymbol{\tau}$
WeightedSimilarity2D ϵ $\mathbb{R}^{B\ X\ 30}$
Logistic regression layer for outcome, $\boldsymbol{\omega}$
HypotensionProbability ϵ $\mathbb{R}^{B\ X\ 1}$
Regression layers : $\boldsymbol{\tau}, \boldsymbol{\omega}$
$\boldsymbol{\tau}(x)$
Importance Layer Output
∗ Shapelet Layer Output
Max operation
$\boldsymbol{\omega}(x)$
Dense1D
• $\boldsymbol{\tau}(x)$ processes importance tensors across shape similarity tensors and selects intervals with maximum similarity
• $\boldsymbol{\omega}(x)$ processes 1 dense layer with sigmoid as activation function.

Fig. 2. Architecture of proposed hypotension prediction model.

B. XAI and Hypotension Prediction Model Interpretability

Explainable AI (XAI), which discloses black-box characteristics, has received significant attention in the medical field [18]. XAI can be grouped into two broad categories: model-specific and model-agnostic methods [19]–[21]. 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 a neural network [19], [20]. By contrast, model-agnostic methods are applied independently of the model by approximating the relationship between the input and output data. Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) are representative examples [19], [20].

Existing hypotension prediction models focus primarily on predictive performance, presenting either no interpretation [6], [8], [9] or a global interpretation of the relative importance of each predictor in predicting hypotension. The authors in [7] and [16] presented model-specific variable importance based on random forest models, while the authors in [17] demonstrated it based on a stochastic gradient boosting model. The authors of [22] and [23], who trained light-gradient boosting algorithms to predict postoperative hypotension, presented model-agnostic variable importance using SHAP.

In terms of local interpretation based on deep learning models, in [5], a model-specific method of Grad-CAM was applied to a 1D-CNN model to present the temporal importance within a given input in predicting IOH. Conversely, to the best of our knowledge, deep learning-based agnostic methods have rarely been applied to predict hypotension. However, an attempt similar to ours was presented in [19] and [24], in which an electrocardiogram signal was employed as the time-series input to predict arrhythmia. Among the three model-agnostic methods of SHAP, LIME, and Anchor applied in [19], only SHAP was evaluated as having adequate interpretability in signal analysis. Likewise, in [25], the authors 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 in this study was benchmarked against the SHAP only.

III. DESIGN FOR HYPOTENSION PREDICTION MODEL

The architecture of the hypotension prediction model with four broad sections of neural network layers is shown in Fig. 2. In the first section (A in Fig. 2), the input records are compressed into an overall trend of the ABP waveform by applying five levels of a discrete wavelet transform (DWT). Subsequently, vectors with certain shapes are generated to characterize the local intervals within the ABP trend vector, followed by calculations of the similarities between all generated trend shapes and these local intervals (B in Fig. 2). Furthermore, the parameters are trained in the third section to weigh the local intervals (C in Fig. 2). After multiplying the similarity values by the weights, the probability of hypotension occurrence 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 ABP data and the occurrence of hypotension.

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) using convolutional filters [26], [27]. The DWT of a discrete signal record of x with filter h is defined as follows:

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

Whereas the approximation coefficient extracts low-frequency components that represent the overall trend of the signal, the detail coefficient extracts high-frequency components that indicate the regional randomness among the signals [26], [27]. Therefore, the overall trend of the ABP, which functions as a hypotension predictor, was obtained through recursive applications of such a transformation placed only on the approximation coefficient.

2) Implementation of Trend Extraction within the Model

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

A sinc function is a time-domain representation of a low-pass filter, which encloses information below a certain cutoff point, denoted as f_c, within the frequency [28], [29]:

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

In this study, *sinc(x)* and *S*[*n*] in Eq. (2) denote the sinc function and filter, respectively. Because the sinc function is derived from the sine function, the essential gradient for backpropagation can be defined well. Owing to this property, f_c can be trained to reduce hypotension prediction errors. 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 cutoff point during a transition. Thus, the Blackman window, which is defined as follows:

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

was applied to the sinc filter to smoothen the ripples to a value of zero [30]. The length of the filter (*N*) was set based on previous studies that utilized the Blackman window for signal processing [30], [31]. An odd number of 51 was selected for *N* to ensure perfect symmetry in all the values within the filter, which were normalized as follows:

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

As the initial model input, 9,000 raw ABP records were used (batch, 9,000), and the convolution between the filter and record vector was operated in the second-dimension direction. The value of f_c was learned through the weight layers for the low-pass filter (φ), as shown in Fig. 2. Subsequently, the record vector was downsized by averaging all the elements in pairs. This trend extraction procedure was applied in five steps, thereby resulting 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 the variables, and can be guaranteed when the parameters for interpretation (i.e., coefficients) are constant after model training. Essentially, these parameters must not be endogenous. This condition is satisfied when the parameters are not computed from the tensors fed toward the model output [32]. Therefore, multiple vectors representing the local shapes of the ABP trends were trained to be independent constants, unaffected by other tensors (ShapeGen in the shape similarity layers shown in Fig. 2) [33]–[35].

The element sizes in the first (S) and second (F) dimensions of the ShapeGen matrix were set as 30 and 50, respectively. Using the same length as the local shape of the trend, the 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 LP5Vector. The similarities (Similarity3D (B, 183, 30)) were calculated between all 30 local shapes and 183 local intervals of LP5Vector3D as follows:

$$Distance3D_{(b,l.s)} \in \mathbb{R}^{B\times 183\times 30} = \sqrt{\sum_{f=1}^{F=50}\left(LP5Vector3D_{(b,l,f)}\text{-}ShapeGen_{(s,f)}\right)^2}, \qquad (5)$$

$$Similarity3D_{(b,l.s)} \in \mathbb{R}^{B\times 183\times 30} = exp(\text{-}\ Distance3D),$$

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

C. Local Importance Layers

Weighting certain elements in the tensors may accelerate the convergence of model training and enhance model interpretability by highlighting the aspects to be considered during the interpretation [36]. To benefit from these strengths, the model was designed to weigh the temporal positions at which the ABP shape was critical for predicting hypotension.

As the values within the adjacent local intervals of LP5Vector3D were primarily identical owing to the one-unit shift between intervals, the weights were estimated using a Gaussian distribution providing the greatest weight to the most significant interval and symmetrically weak weights to the remaining intervals. Furthermore, by training the local importance as a summation of the probabilities from five Gaussian distributions, the importance was learned with high flexibility [37]. Specifically, if the mean values that determine the location of the Gaussian distribution are scattered, the importance is spread over 183 local intervals. However, if the mean values are close to each other, the importance is concentrated around a particular local interval through a summation of the weights.

The procedure for weighting the local intervals within the model was implemented through the local importance layers (C in Fig. 2). The weight layers for local importance (μ) were trained to obtain five mean values of the Gaussian distributions (*M*). Given each mean value as a trainable parameter and a standard deviation value (σ) of 0.075 as a hyperparameter within a set, the processing layers for local importance (ρ) computed the probabilities according to a discrete random variable (x) of 0 to 183 in the following manner:

$$GausDist_{(b,m,l)} \in \mathbb{R}^{B\times 5\times 183} = N(X|\mu,\sigma^2) \equiv \frac{exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}. \tag{6}$$

All probability values in GausDist were scaled as follows:

$$NormGD_{(b,m,l)} \in \mathbb{R}^{B\times 5\times 183} = \frac{(GausDist\text{-}MinGD)}{(MaxGD\text{-}MinGD)}, \tag{7}$$

where $MaxGD$ indicates $max_{l\in L}GausDist_{b,m,l}$, and $MinGD$ indicates $min_{l\in L}GausDist_{b,m,l}$ with $\forall b \in batch, m \in M$. Thus, the values of the individual distributions are summed across the axis of m, as follows:

$$IntervalWeight_{(b,l)} \in \mathbb{R}^{B\times 183} = \sum_{m=1}^{M=5} NormGD_{(b,m,l)}, \tag{8}$$

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

D. Regression Layers

For models whose outcome is significantly affected by latent variables produced through multiple nonlinear operations, an interpretation of the association between each input and output variable may be very challenging [38]. However, although multiple layers were implemented in this model structure, the association can still be interpreted as the interpretably extracted ABP trend is fed into the final regression layer (D in Fig. 2). For the layer, a logistic regression was employed because the output of the model was a binary variable indicating the occurrence of hypotension.

Each feature used as an independent variable in the logistic regression layer is 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 IntervalWeight by the similarity values in an element-wise manner in the second dimension of Similarity3D. Among all the weighted values, only the maximum value within 183 local intervals was selected in the following manner:

$$IntervalWeight \in \mathbb{R}^{B\times 183} \rightarrow IntervalWeight \in \mathbb{R}^{B\times 183\times 1},$$

$$\begin{aligned} WeightedSimilarity3D_{(b,l,s)} &\in \mathbb{R}^{B\times 183\times 30} \\ &= IntervalWeight_{(b,l,1)} * Similarity3D_{(b,l,s)}, \end{aligned} \tag{9}$$

$$\begin{aligned} WeightedSimilarity_{(b,s)} &\in \mathbb{R}^{B\times 30} \\ &= max_{l\in L} WeightedSimilarity3D_{(b,l,s)}, \end{aligned}$$

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

Using 30 weighted similarity values that represent the most similar local intervals of each shape, the logistic regression can be specified as follows:

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

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

E. Objective Function

The objective function is defined as the summation of the binary cross-entropy and shape loss. The binary cross-entropy minimizes the prediction error (i.e., hypotension), whereas the shape loss generates ABP trend shapes by reducing the Euclidean distances between the generated features and local ABP intervals. Herein, because the similarity values used as the input variables in the logistic regression layer were weighted, the Euclidean distance values were also weighted to compute the shape loss.

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

The implementation of the shape loss computation started with the Distance3D tensor in Eq. (5). Distance3D was expanded to 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 zero. To conduct an element-wise multiplication with Distance4D, MaxNormGD was expanded to a 4D tensor. The summation of all multiplicative products between Distance4D and MaxNormGD along the interval axis was then returned to ShapeLoss3D in the following manner:

$$Distance3D \in \mathbb{R}^{B\times 183\times 30} \rightarrow Distance4D \in \mathbb{R}^{B\times 5\times 183\times 30},$$

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

$$MaxNormGD \in \mathbb{R}^{B\times 5\times 183} \rightarrow MaxNormGD \in \mathbb{R}^{B\times 5\times 183\times 1}, \tag{11}$$

$$\begin{aligned} &ShapeLoss3D_{(b,m,s)} \in \mathbb{R}^{B\times 5\times 30} \\ &= \sum_{l=1}^{L=183} Distance4D_{(b,m,l,s)} * MaxNormGD_{(b,m,l,1)}, \end{aligned}$$

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:

$$ShapeLoss2D_{(b,m)} \in \mathbb{R}^{B\times 5} = min_{s\in S} ShapeLoss3D_{(b,m,s)}, \tag{12}$$

where $\forall b \in batch$ and $m \in M$. The second dimension in ShapeLoss2D represents the distances computed between the five local intervals and the local shapes that are most similar to these local intervals (Fig. 5 in the supplementary document). The average of all the elements in ShapeLoss2D was used as the final shape loss, as follows:

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

The final objective function of the model is as follows:

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

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

where ω is the matrix of all network weights.

IV. EXPERIMENTAL SETTING

A. Data Processing

1) Data Acquisition

The ABP data records were obtained from two independent medical institutions and processed according to the model settings. Data from 10,454 patients who were treated at Asan Medical Center, stratified by surgical durations of less than 3 h, 3–5 h, and greater than 5 h, were processed into 1,548,927 samples. The numbers of patients in each group were 3,181, 3,244, and 4,029, respectively. Subsequently, the entire set of processed data was randomly sampled into a training dataset of 600,000 samples and internal validation dataset of 60,000 samples. The study conducted on 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 owing to the minimal risk imposed on the study participants. In addition, 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 [39]. Patient information collected from the two institutions is listed in Table I.

2) Data Cleansing

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

The second filtering procedure removed 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 were identified at every cycle of the ABP waveform. ABP values within a cycle were then resampled to obtain the same vector length (i.e., 100 data points) using a fast Fourier transform [40]. Any cycle with ABP records deviating by at least 15% from the mean value of the 20 adjacent cycles was considered a noise candidate. For these candidates, 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 deviating by at least 15% from the mean values of consecutive cycles were excluded. Finally, subsets with a series of at least 15 min in length without discontinuity in the data were included in the training dataset.

TABLE I

PATIENT CHARACTERISTICS FOR INTERNAL AND EXTERNAL VALIDATION

	VitalDB (n = 3,278)	**AMC (n = 10,454)**
Age (y)	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 are reported for age, weight, and anesthesia duration, whereas sex is reported as counts.

3) Hypotension Operationalization

The model outcomes (1 = hypotension or 0 = nonhypotension) were operationalized based on the mean arterial pressure (MAP $= \frac{\text{SystolicPressure+2*DiastolicPressure}}{3}$) [41].

Based on previous research, the periods during which all MAPs were maintained at less than 65 mmHg for at least 1 min were defined as the occurrence of hypotension [5], [6]. By contrast, periods during which all MAPs were maintained at >75 mmHg for at least 1 min were operationalized as nonhypotension [6]. The remaining periods with MAP values of between 65 and 75 mmHg were considered a “gray zone” [6]. Only definite hypotension (i.e., MAP <65 mmHg) and nonhypotension (i.e., MAP >75 mmHg) events were used to train the prediction model. However, because the sensitivity of the model accuracy is worth validating according to the gray zone during external validation, three additional tests with various MAP thresholds applied to nonhypotension samples were conducted. Accordingly, in terms of the external validation set, nonhypotension was redefined as 1) samples with at least one incidence of MAP greater than 65 mmHg within a 1 min period, or samples in which all MAPs were greater than 2) 65 or 3) 70 mmHg for at least 1 min.

Because hypotension cases are less common than nonhypotension cases, the label distribution was adjusted using random sampling with a maximum limit of 30 nonhypotension cases per patient. This final procedure mitigated any imbalance in the distribution, which may adversely affect the convergence in model training [42].

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 shown in Fig. 2 were estimated using a gradient descent-based optimizer [43]. Although the association between the local ABP shapes and hypotension occurrence can be interpreted based on these coefficient values, a deep learning model without p-values cannot provide statistical significance for this relationship [44]. Therefore, an additional logistic regression analysis based on the Newton–Raphson optimizer was applied to obtain the p-values of these coefficients.

The Newton–Raphson method approximates the solution to f(x) = 0 by identifying a tangent from the current value and updating the value to the point where the tangent meets the x-axis [45], [46]. The Newton–Raphson method was applied to the maximum likelihood estimation in the logit model. Using this method, the parameters that lead the first-derivative of the likelihood function to a value of zero were iteratively estimated until convergence was achieved [46]. Finally, the statistical significance of the coefficients (θ) can be determined based on the hypothesis test through a regression analysis, in which the null hypothesis states that each coefficient equals 0. In this model, the similarity values calculated from the WeightedSimilarity layer after training the proposed model were employed as independent variables.

In a regression analysis, high correlations between independent variables may lead to biased coefficient estimates [47]. Hence, data preprocessing was applied to address the high correlations between all pairs of the 30 weighted similarity values. In particular, only one trend shape was selected from the groups with correlations exceeding 0.8, thus allowing only distinct shapes that were not highly correlated with the others to be analyzed further. Subsequently, only the weighted similarity values corresponding to the representative shapes are included as independent variables in the logistic regression model, as shown in Eq. (10).

After fitting the regression model, only statistically significant ($p < 0.01$) shapes were selected, and the signs and effect sizes of their coefficients were analyzed. In addition, the consistency between the selected coefficients (θ) and those estimated using the deep learning model (β) was investigated.

2) *Exploratory Data Analysis based on Descriptive Statistics*

An exploratory data analysis (EDA) was conducted for a multidimensional review of the proposed model and predictor. First, the 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 nonhypotension) [47]. In this study, a confusion matrix was computed based on the samples with the maximum similarity to each shape.

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

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

C. Evaluation of Applicability of Proposed Interpretable Method

1) *Interpretability Comparison with SHAP*

The interpretability of the proposed model was further evaluated by benchmarking the SHAP. Given that SHAP explains model predictions by assigning weights to certain features [48], the absolute value of SHAP represents the contribution of a feature to the prediction. The SHAP values in this study were computed by approximating the expected gradients [49] with respect to 9,000 temporal ABPs and indicated the influence of the temporal positions within the input ABPs on the prediction of hypotension. The SHAP values were compared with those of the proposed method to explore the aspects of our interpretable approach with greater interpretable power.

2) *Assessment based on Survey of Anesthesiologists*

The proposed interpretation method was evaluated in three aspects by surveying anesthesia experts [50]. 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 question is whether providing interpretative predictors along with the predicted probability of hypotension is useful in clinical practice (i.e., clinical usefulness). The third aspect is the willingness of the clinician to intervene based on the predictive information (i.e., willingness to intervene).

A group of 17 board-certified anesthesiologists from the 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) a set of trend shape similarities, the odds of hypotension occurrence, and the history of predicted probabilities.

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, based on a five-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 the questions and a sample visual summary for each category are shown in Figs. 7, 8, and 9 in the supplementary document.

D. Ablation Study

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

V. RESULTS

A. Prediction Performance

The hypotension prediction performance of the proposed model was evaluated using a training set, internal validation set, and external validation set. The areas under the receiver operating characteristic curves (AUCs) of the three sets with nonhypotension, defined as MAPs being above 75 for at least a 1 min period, are reported in Table II-a as 0.9152, 0.9145, and

0.9035, respectively. The AUCs of the external validation set with nonhypotension, defined as 1) at least one MAP value above 65, 2) all MAPs 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 decreased slightly as more indefinite samples for nonhypotension were included in the dataset, the overall 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 scores are reported. Nonhypotension samples are defined as all MAP values being 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

AUC scores are reported based on the external validation set. Definitions of nonhypotension samples are presented in the first row.

B. Evaluation of Interpretability

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

Among the 30 shapes generated, four were found to have low correlations with each other. Thus, the coefficient values corresponding to these four shapes within the regression layer (ω) were analyzed further. Among the four coefficients estimated using the Newton–Raphson estimator, three were statistically significant, whereas one was not within the 5% significance level. The details of the coefficients for each estimation method, including the p-values and effect sizes, are shown in the proposed model column in Table III. Shapes A and B, which had negative coefficients, were associated with nonhypotension, 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 nonhypotension than shape B is. These results are consistent in both the gradient descent and Newton–Raphson optimizers, which may ensure a robust interpretation of the association between the shapes and hypotension in the deep learning model.

2) Results of Exploratory Data Analysis

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

Fig. 3a presents the overall magnitude of the values of the three statistically significant local ABP shapes extracted from the ShapeGen layer. In terms of the overall trend in values,

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 nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension
Actual nonhypotension	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 nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension
Actual nonhypotension	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 nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension	Predicted as nonhypotension	Predicted as hypotension
Actual nonhypotension	7,038 (0.192)	4,561 (0.125)	–	–	–	–
Actual hypotension	2,133 (0.058)	22,891 (0.625)	–	–	–	–

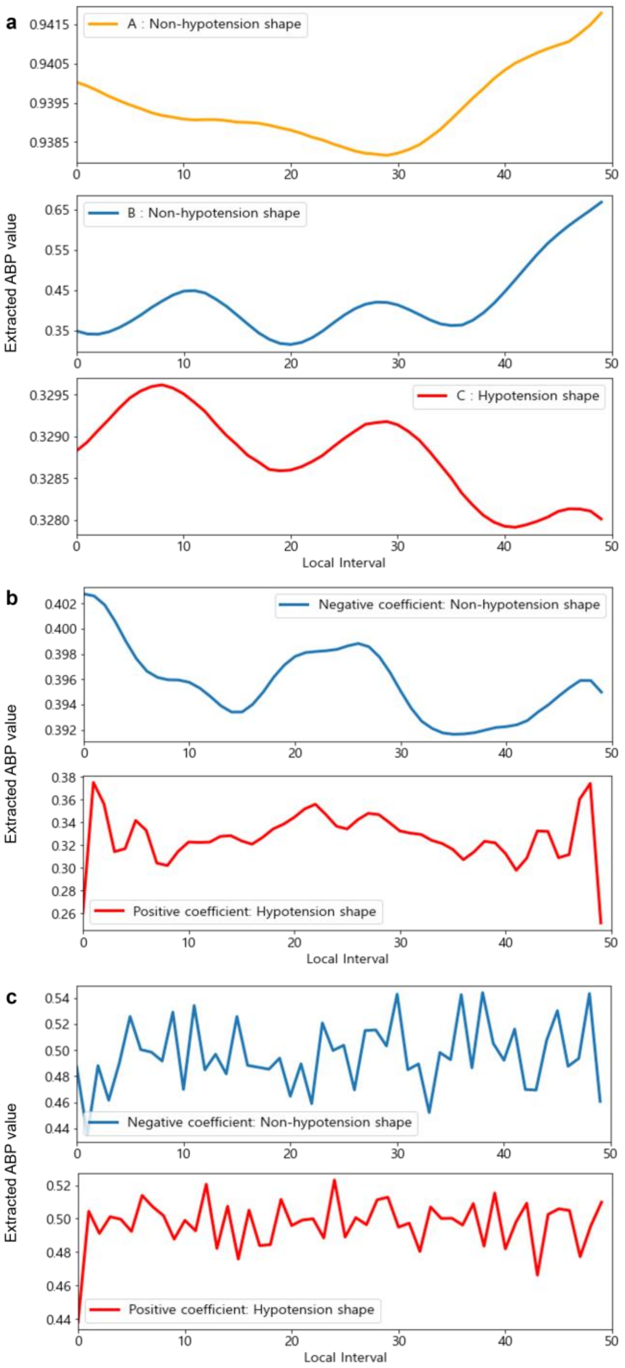


Fig. 3. Representative shapes of local ABP trends: a) Proposed model; b) Ablation model 1; and c) Ablation model 2.

shapes A and B gradually increased over time, whereas shape C gradually decreased. Based on the definition of hypotension, in which low MAP values persist for more than 1 min, these results suggest that shape C is associated with the development of hypotension. In addition, the magnitude values of shape A were greater than those of shape B, which may suggest that shape A has a stronger relationship to nonhypotension. Thus, these results demonstrate a reasonable association between hypotensive development and the morphological characteristics of each ABP trend.

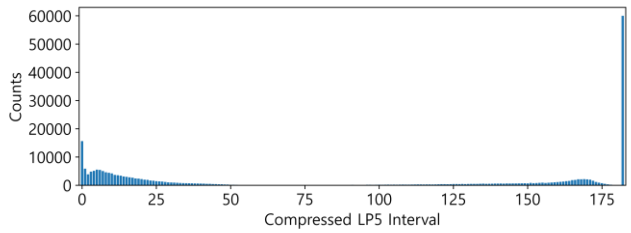


Fig. 4. Counts for each local interval that is influential in the prediction of hypotension.

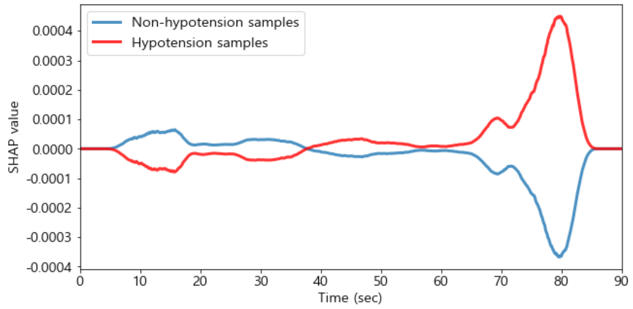


Fig. 5. Mean SHAP values of hypotension and nonhypotension samples.

Finally, Fig. 4 shows the distribution of weights that influence hypotension prediction across the 183 local intervals. Because weights were assigned by the five parameters for every sample, the distribution was formed using 300,000 values from the 60,000 validation samples. As shown in Fig. 4, approximately 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 for predicting hypotension.

C. Evaluation of Proposed Interpretable Method

1) Model Interpretability Based on SHAP

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

2) Survey Results of Anesthesiologists

Table V presents the average and standard deviation values of the survey questionnaires based on three aspects according to the three categories. For all assessments, the mean scores increased sequentially from Categories 1 to 3. Specifically, the average scores of all three questionnaires in Categories 2 and 3, which evaluated the interpretable components of the proposed model, were 3 or higher. Compared with 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 exhibited AUCs of 0.9090, 0.9087, and 0.8923 for the training, internal validation, and external validation sets, respectively, which were approximately 0.006–0.01 lower than those of the proposed model (Table II-a). Although two statistically significant shapes were identified (Table III), the coefficient signs for nonhypotension estimated based on the gradient descent and Newton–Raphson methods were inconsistent. Moreover, the trend shapes illustrated in Fig. 3b may not be associated with the development of hypotension. In particular, the gradual decline in the ABP trend in the shape statistically associated with nonhypotension and the irregular pattern of the ABP trend in the hypotension-related shape are not clinically rational. These results suggest that these factors have little explanatory power as predictors of hypotension.

Ablation model 2 yielded 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 a significance level of 0.05 were generated (Table III). However, because all samples had maximum similarity in shape A, as presented in Table IV, the representativeness of the shapes that account for hypotension prediction is lacking. Moreover, irregular values that fluctuate over time are almost random walks and are unlikely to provide a plausible interpretation for hypotension prediction (Fig. 3c). Thus, ablation model 2, whose AUC was as high as that of the proposed model, had poor explanatory power.

VI. DISCUSSION

A. Performance Verification of New Hypotensive Predictor

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

Similarly, the proposed model, using local ABP trends as a new predictor, exhibited excellent performance compared with existing models. In particular, in the internal and external validations, the AUCs of the model were 0.9145 and 0.9035, respectively, which were similar to those of the previous models (Table II-a). These results suggest that the local ABP trends generated through deep learning layers can be an effective hypotension predictor.

Moreover, an additional test was conducted for rigorous performance evaluation. In particular, unlike previous studies, in which the performance was assessed based only on well-behaved samples [5], [6], this study evaluated the model performance with the inclusion of ill-behaved samples, which accounted for a gray zone of MAP values for nonhypotension labeling. In the external validation of these samples, the AUC was reported to be 0.8337 to 0.9035, as the MAP value threshold for defining hypotension varied from 65 to 75 mmHg (Table II-b). Although it tends to be less accurate in a rigorous setting, an AUC of 0.8337 indicates good performance. Thus, these results suggest that compressed local ABP trends are robust predictors 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 correspond to the model interpretations based on (1) SHAP values, (2) trend shape similarity, and (3) a set of trend shape similarities, odds of hypotension occurrence, and history of predicted probabilities, respectively. Average and standard deviation values are reported for each category.

B. Clinical Applicability of Proposed Interpretable Method

In addition to performance verification, the clinical applicability of the proposed interpretable method was verified based on three aspects of the survey results. First, the predictors generated by the proposed model may have high clinical fidelity. From a theoretical perspective, it has been well established in physiology that changes in cardiac preload are an important cause of intraoperative hypotension [51]. In particular, cardiac preload can be indicated by the respiratory variability of the ABP waveform [51], which is an important piece of information that anesthesiologists can obtain by monitoring ABP waveforms. As the low-frequency component of ABP is caused primarily by respiratory variations [14], ABP trends can generally play an important role as precursors of hypotension [52]. Accordingly, trends extracted from 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 strong association between the proposed predictors and the knowledge of clinicians (Table V). Further, 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 than the conventional method for understanding hypotension prediction.

Second, the higher ratings of clinical usefulness in Categories 2 and 3 compared with those of Category 1 (Table V) may indicate that the proposed interpretable method provides more useful information than SHAP. As a common feature between SHAP and the proposed interpretable method, both indicate the importance of certain temporal positions for predicting hypotension within a given 90 s of ABP data. However, the proposed method provides additional information regarding 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 in enabling anesthesiologists to acquire information intuitively.

Finally, the highest scores in Category 3 in terms of willingness to intervene (Table V) suggest that the history of predicted hypotension probability and the odds ratio of

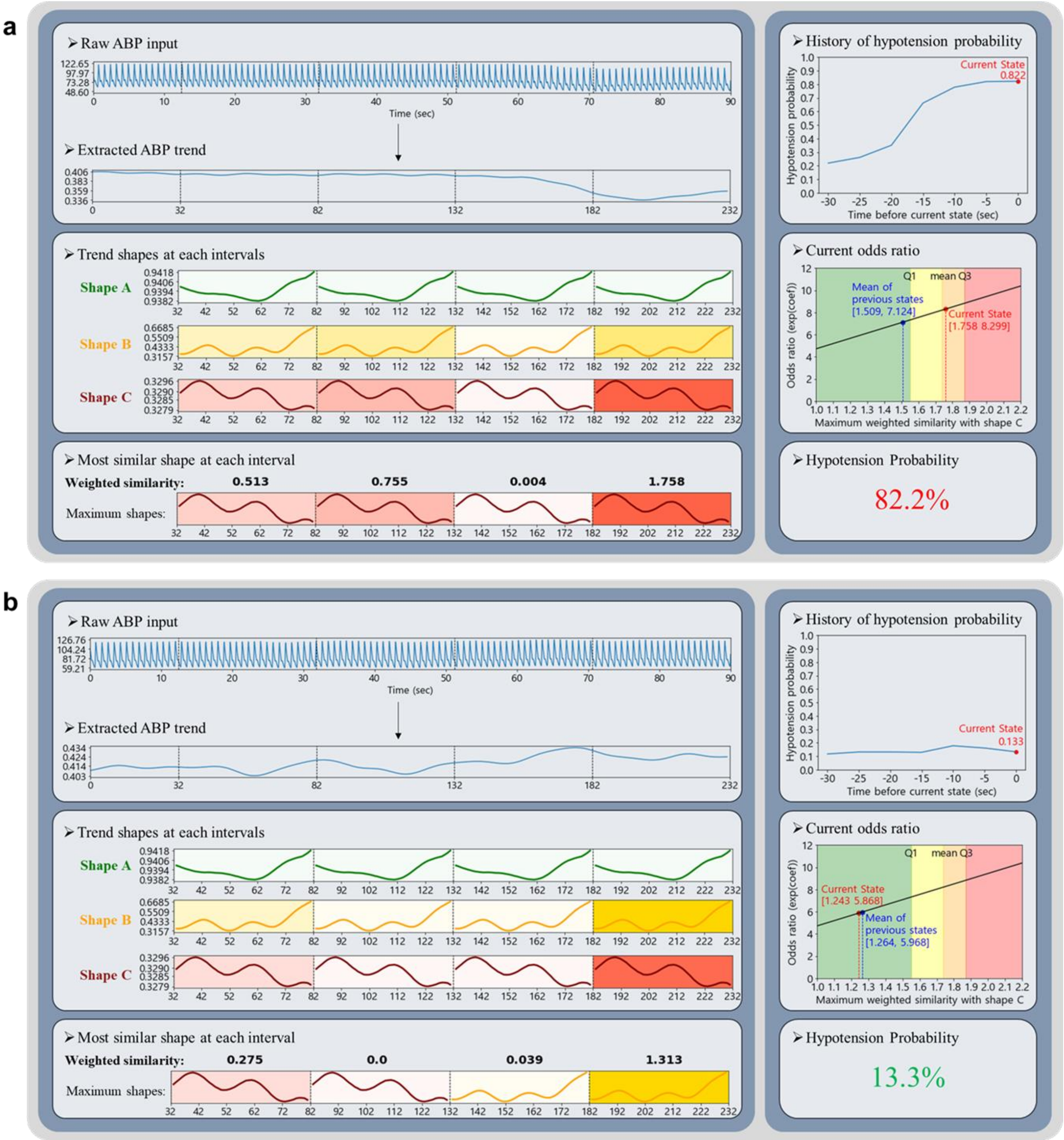


Fig. 6. Visual summary of interpretable hypotension prediction: a) hypotension and b) nonhypotension sample cases.

hypotension occurrence provide convincing information regarding the prognostic signs of hypotension. In particular, because an interpretation based on the odds of logistic regression is one of the most widely accepted methods [47], 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 information already familiar to them.

C. Scenario-based Clinical Guidelines

To encourage the practical adoption of the proposed interpretable method, an end-to-end process is visualized in Fig. 6, starting from the raw ABP input data and displaying the final probability computation of hypotension.

In the first section, the raw ABP and its trend, extracted using DWT, are presented. The extracted ABP trend is more readable compared with the raw ABP trend in terms of overall changes. Because the information on respiratory variation is captured by the ABP trend, observing the trend dynamics in blood pressure levels may help anesthesiologists monitor the precursors of hypotension.

Second, the weighted similarities of the three significant ABP local trends (Fig. 3a) with four consecutive local intervals are presented and differentiated by color. This is a local interpretation of the extracted ABP trend and may yield critical information by foreshadowing the future blood pressure state. In particular, local morphological changes in the hypotensive precursor allow anesthesiologists to identify a transition from a high to low or from a low to high risk of hypotension over time. 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, because regression analysis is conducted under the assumption that independent variables are not highly correlated with each other, the association between hypotensive development and each generated ABP shape can be interpreted independently [53], [54]. This implies that the relationship between hypotension and ABP shape of interest can be interpreted while holding all other ABP shapes constant (all else being equal). In Fig. 6, the odds of hypotension occurrence are plotted to illustrate the shift in the odds from the mean of the previous states to the current state given the maximum similarity value with shape C among the four intervals (Table III). This 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 or transition to normal blood pressure).

Finally, the upper-right section indicates the change in hypotension probability every 5 s from 30 s before the current state. Here, the probability of hypotension in 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 nonhypotension conditions illustrates the clinical intuitiveness of the proposed interpretable method.

D. Theoretical Discussion of Contributions

This study contributes to the incorporation of clinical practices into the development of XAI, essentially providing insights that can bridge the gap between AI-based interpretations and medical practices. The contributions are discussed from the perspective of technology adoption theories, starting with the identification of the limitations of existing interpretable methods in medicine.

Traditionally, new knowledge has not been well-integrated into medical practice [55]. Among the various reasons for this, changes proposed without considering clinical needs can be a major obstacle to adopting emerging AI-based concepts [55]. Recently, various attempts have been made to apply XAI to medical prediction problems, thereby improving interpretability [18], [21]. However, when limiting these applications to the realm of signal data, evidence that the existing methods can address clinical needs is lacking. In particular, interpretable methods, such as SHAP, just reveal important temporal positions in a signal for predicting an event, as shown in Fig. 5 [24], [56]–[58]. However, previous studies have neglected in-depth discussions on whether the presentation of temporal importance satisfies such needs. In addition, our survey results (Table V), in which SHAP-based interpretations were rated the lowest by anesthesiologists, call into question how seriously previous studies have considered the need for clinical practice.

Given that change often entails an additional burden [55], low-value innovations further increase the barriers to the adoption of new concepts. Thus, efforts should be directed toward the development of interpretable methods that are acceptable to practitioners in the field. Previous research has theorized that perceived usefulness and compatibility with current practices are essential factors for medical technology adoption [59]. Another well-established theoretical study conceptualized that job relevance, defined as the capability to support one’s task, positively affects perceived usefulness [60]. Accordingly, this study attempted to incorporate some important practices relevant to hypotension monitoring into the development of an AI-based interpretation framework.

First, in the proposed method, the clinical rationale for hypotension detection in the interpretable mechanism was addressed. As discussed previously, the relationship between ABP trends and the IOH can be well explained in terms of physiology [14], [52], [61]. As physiological evidence is a key factor in clinical decision-making [62], this relationship can be a fundamental premise for anesthesiologists to detect hypotension. Thus, the interpretation of ABP trends provided by the proposed method is compatible with the practice of hypotension monitoring, which has a positive effect on perceived usefulness. Moreover, the proposed method only highlights trends in ABP data, thereby facilitating the understanding of anesthesiologists regarding physiological changes. It is more efficient for anesthesiologists to understand the presented trends than to cognitively isolate the trend from the raw ABP data and then try to understand them. Increased work efficiency has been theorized to have a positive effect on perceived usefulness [59], thus suggesting that our method is highly useful for anesthesiologists.

Second, a framework was proposed 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 shown in Fig. 1 may assist anesthesiologists in building hypothesis. In particular, even if anesthesiologists have hypothetical ABP trend shapes that cause IOH, they are likely to be uncertain regarding the specific shapes for statistical operationalization. Given this incomplete hypothesis, this section completes the operationalization by generating interpretable ABP trend shapes. This completion was attributed to the DWT layers implemented within the generation part. In general, the typical nonlinear learning structure of a deep learning model may generate features that do not account for domain specificity [63]. However, our DWT layers preserve the original nature of the ABP waveform when generating trend shapes. Thus, the property of the part that 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 is scalable for statistical significance testing. Statistical models may have advantages over common interpretable methods such as SHAP because they provide a generalizable interpretation. In particular, SHAP values change according to the given input data [20], thus rendering difficulty in generalizing the interpretation. However, because the p-values and odds are constant regardless of the given input data, a logistic regression can provide a generalizable interpretation [47]. Generalizability is extremely important in clinical practice for evidence-based decision-making [18]. Nevertheless, existing studies on XAI have tended not to consider this important practice seriously. By contrast, the proposed framework, which integrates AI-based and statistical interpretations, enhances generalizability, thereby suggesting its high compatibility with major clinical practice.

Third, the guidelines for the proposed method were presented and evaluations of the method from clinicians were received, which have often been neglected in previous studies. The insufficient guidance provisioning of previous methods may be due to the paucity of clinically relevant information because domain specificity was not well incorporated into them [12], [63], [64]. By contrast, the proposed interpretable method can provide new types of clinically relevant information by addressing physiological mechanisms. However, even new advanced concepts in medicine may not propagate without proper guidance [55]. Accordingly, scenario-based guidelines were provided as a visual summary outlining how the information output by the proposed method can be well integrated into current practices. Furthermore, because the ABP trend shapes generated by our model are new, an evaluation of their relevance to the physiological basis is essential. As the physiological basis is close to the qualitative domain, it was assessed by a group of experts in this study [62]. Although the evaluation was conducted 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 [55], [65]. In particular, the majority adopt new medical concepts through interaction with a minority that has already embraced the concepts [55]. This suggests that the evaluation of a new concept by a small number of early adopters can have a positive impact on future dissemination. Interestingly, the high evaluation of the relevance of ABP trend shapes to the physiological basis in this study may indicate that anesthesiologists have become aware of its potential usefulness during the 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 technology acceptance theory, providing insight into the direction of future research.

E. Limitations and Future Work

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

The second limitation concerns the computational inefficiency of learning the 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 same position in another sequence, the similarity between two sequence data is inflexibly measured [66]. Thus, redundant shapes that are highly correlated with each other may be produced. This limitation may be alleviated by employing dynamic time warping (DTW) distance, which allows a nonlinear alignment between two sequence datasets to account for the similarity of local shapes that are adjacent but different in their shapes [66]–[68]. However, 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 [67] should be encouraged.

The last limitation is the lack of flexibility in *post-hoc* interpretation, which is caused by the dependence of the model-specific interpretation on the model structure [20]. For instance, because the ABP trend shapes were generated based on 30 s of ABP in this study, the association between hypotension development and ABP trends at different time lengths, such as 40 s, cannot be interpreted. Therefore, the models must be redesigned to address the association between factors and outcomes based on different interests. Given the recent complexity of the model structures used in deep learning and the need for large amounts of data for training, the training of all models according to the interest of each interpretation is computationally inefficient. However, the hasty generalization that model-agnostic interpretations are superior to model-specific interpretations should be avoided. Rather, studies on how the two methods can be combined and used together for 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 groundbreaking support in the medical field. The hypotension prediction model proposed in this study focuses on the local trends of ABP in interpreting how certain shapes of local trends are associated with hypotension, along with good predictive power. To facilitate ABP monitoring in clinical practice, we expect that more empirical validation of hypotension prediction models will be conducted within 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.

[2] A. Gregory *et al.*, “Intraoperative Hypotension Is Associated With Adverse Clinical Outcomes After Noncardiac Surgery,” *Anesth. Analg.*, vol. 132, no. 6, pp. 1654–1665, Jun. 2021.

[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.

[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.

[5] S. Lee *et al.*, “Deep learning models for the prediction of intraoperative hypotension,” *Br. J. Anaesth.*, vol. 126, no. 4, pp. 808–817, Apr. 2021.

[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.

[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.

[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," *J. Med. Internet Res. Medical Informatics*, vol. 9, no. 9, p. e31311, Sep. 2021.

[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:," *Anesth. Analg.*, vol. 130, no. 5, pp. 1157–1166, May 2020.

[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.

[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.

[12] S. Kundu, "AI in medicine must be explainable," *Nat. Med.*, vol. 27, no. 8, pp. 1328–1328, Aug. 2021.

[13] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli, "Applications of Deep Learning and Reinforcement Learning to Biological Data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2063–2079, Jun. 2018.

[14] 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.

[15] S. Mao and E. Sejdić, "A Review of Recurrent Neural Network-Based Methods in Computational Physiology," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–21, 2022, doi: 10.1109/TNNLS.2022.3145365.

[16] 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.

[17] 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.

[18] 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. [Online]. Available: http://arxiv.org/abs/1712.09923

[19] 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 *Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 4211–4222.

[20] A. Carrillo, L. F. Cantú, and A. Noriega, "Individual Explanations in Machine Learning Models: A Survey for Practitioners," 2021, *arXiv:2104.04144*.

[21] E. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Jan. 2021.

[22] N. Prasad and K. Palla, "The Role of Context in the Prediction of Acute Hypotension in Critical Care," in *SAIL: Symposium on Artificial Intelligence for Learning Health Systems*, 2020, vol. 10, p. 6.

[23] 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.

[24] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, "Explainable AI decision model for ECG data of cardiac disorders," *Biomed. Signal Process. Control*, vol. 75, p. 103584, May 2022.

[25] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "An Empirical Evaluation of AI Deep Explainable Tools," in *IEEE Globecom Workshops*, Dec. 2020, pp. 1–6.

[26] 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, 613.

[27] S. K. Khare and V. Bajaj, "Time–Frequency Representation and Convolutional Neural Network-Based Emotion Recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 2901–2909, Jul. 2021.

[28] 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.

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

[30] S. M. S. Alam and T. Hasan, "Performance Analysis of FIR Filter Design by Using Optimal, Blackman Window and Frequency Sampling Methods," *Int. J. Electr. Comput. Sciences*, vol. 10, no. 01, p. 6, 2010.

[31] 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.

[32] M. Akrout, C. Wilson, P. C. Humphreys, T. Lillicrap, and D. Tweed, "Deep Learning without Weight Transport," 2020, *arXiv:1904.05391*.

[33] K. Huang, S. Wu, F. Li, C. Yang, and W. Gui, "Fault Diagnosis of Hydraulic Systems Based on Deep Learning Model With Multirate Data Samples," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6789–6801, Jan. 2022.

[34] Y. Huang, G. G. Yen, and V. S. Tseng, "Snippet Policy Network V2: Knee-Guided Neuroevolution for Multi-Lead ECG Early Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2022, doi: 10.1109/TNNLS.2022.3187741

[35] Z. Yang, A. Zhang, and A. Sudjianto, "Enhancing Explainability of Neural Networks Through Architecture Constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2610–2621, Jun. 2021.

[36] Z. Niu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[37] P. Michel, O. Levy, and G. Neubig, "Are Sixteen Heads Really Better than One?," in *Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[38] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[39] 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.

[40] 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.

[41] H. Lee *et al.*, "Deep Learning Model for Real-Time Prediction of Intradialytic Hypotension," *Clin. J. Am. Soc. Nephrol.*, vol. 16, no. 3, pp. 396–406, Mar. 2021.

[42] M. M. Rahman and D. N. Davis, "Addressing the Class Imbalance Problem in Medical Datasets," *Int. J. Mach. Learn. Cybern.*, pp. 224–228, 2013.

[43] 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.

[44] 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, [Online]. Available: http://arxiv.org/abs/2010.01823

[45] A. Ben-Israel, "A Newton-Raphson method for the solution of systems of equations," *J. Math. Anal. Appl.*, vol. 15, no. 2, pp. 243–252, Aug. 1966.

[46] S. A. Czepiel, "Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation," 2002, [Online]. Available: http://czep.net/stat/mlelr.pdf

[47] G. A. Morgan, J. J. Vaske, J. A. Gliner, and R. J. Harmon, "Logistic Regression and Discriminant Analysis: Use and Interpretation," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 42, no. 8, pp. 994–997, Aug. 2003.

[48] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[49] S. Lundberg, "slundberg/shap." Jun. 07, 2022. [Online]. Available: https://github.com/slundberg/shap

[50] 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," *Int. J. Med. Inform.*, vol. 98, pp. 56–64, Feb. 2017.

[51] X. Monnet, P. E. Marik, and J.-L. Teboul, "Prediction of fluid responsiveness: an update," *Ann. Intensive Care*, vol. 6, p. 111, Nov. 2016.

[52] 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.

[53] L. Tanzi, P. Piazzolla, and E. Vezzetti, "Intraoperative surgery room management: A deep learning perspective," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 16, no. 5, p. e2136, 2020.

[54] M. A. Poole and P. N. O'Farrell, "The Assumptions of the Linear Regression Model," *Trans. Inst. Br. Geogr.*, no. 52, pp. 145–158, 1971.

[55] D. M. Berwick, "Disseminating Innovations in Health Care," *JAMA*, vol. 289, no. 15, p. 1969, Apr. 2003.

[56] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey," 2021, *arXiv: 2104.00950*.

[57] 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.

[58] 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.

[59] 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," *Int. J. Med. Inform.*, vol. 77, no. 5, pp. 324–335, May 2008.

[60] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Management Sci.*, vol. 46, no. 2, pp. 186–204, Feb. 2000.

[61] 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.

[62] M. R. Tonelli, "Integrating evidence into clinical practice: an alternative to evidence-based approaches," *J. Eval. Clin. Pract.*, vol. 12, no. 3, pp. 248–256, 2006.

[63] 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.

[64] 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.

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

[66] 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.

[67] X. Cai, T. Xu, J. Yi, J. Huang, and S. Rajasekaran, "DTWNet: a Dynamic Time Warping Network," in *Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[68] F. Rivest and R. Kohar, "A New Timing Error Cost Function for Binary Time Series Prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 174–185, Jan. 2020.

Eugene Hwang received a BS degree in chemical and biomedical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 2018. She is currently a PhD candidate in the Department of Management Engineering, KAIST College of Business, Seoul, Korea. Her research interests include artificial intelligence and medical informatics.

Yong-Seok Park received a BSc degree and an MD in medicine from the University of Ulsan, College of Medicine, Seoul, Korea in 2009, and an MSc degree from the Graduate School, University of Ulsan, College of Medicine, Seoul, Korea in 2013. He is currently a clinical assistant professor at the Department of Anesthesiology and Pain Medicine, Asan Medical Center, Seoul, Korea. His research area includes general anesthesia, neuroanesthesia, and bio-signal analysis.

Jin-Young Kim received a BS degree in data information and computer science from Pyeongtaek University, Gyeonggi-do, South Korea in 2016. He is currently pursuing an MS degree in biomedical engineering at the University of Ulsan and is a researcher at Asan Medical Center, Seoul, Korea. His research areas include data engineering, electronic medical record analysis, and bio-signal analysis.

Sung-Hyuk Park received a BS degree in applied mathematics from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 2005 and a PhD in management engineering from KAIST College of Business, Seoul, Korea in 2011. He is currently an assistant professor in the Department of Management Engineering, KAIST. His research interests include business analytics, recommender systems, and artificial intelligence.

Junetae Kim received a BBA degree from Hanyang University, Seoul, Korea in 2013, and a PhD in management information systems from the Korea Advanced Institute of Science and Technology (KAIST), Seoul, Korea, in 2018. He is currently an assistant professor at the Graduate School of Cancer Science and Policy (GCSP), National Cancer Center, Goyang-si, Korea. Prior to moving to GCSP in 2019, he worked on both the smart IT team and the strategy marketing team at the Samsung Electronics' Device Solutions Division. His research fields include the development of artificial intelligence models with biosignals and genetic data.

Sung-Hoon Kim received an MD degree from the Hanyang University School of Medicine, Seoul, Korea, in 2005, and a PhD from the University of Ulsan College of Medicine, Seoul, Korea, in 2013. He also studied statistical informatics and received a BSc degree from the Korea National Open University in 2017. He is currently an associate professor in the Department of Anesthesiology and Pain Medicine, Asan Medical Center, Seoul, Korea. His research areas include bio-signal analysis and anesthesia outcomes.

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 Blackman window, as presented in Fig. 1. The frequency cutoff value, which determines the level of frequency to be filtered, was trained using a deep learning model.

Utilizing the above filter, Fig. 2 demonstrates how the input data are processed within the first section of the layers in the proposed model. Because the ABP records utilized in this model were sampled at 100 Hz, 90 s of ABP records consisted of 9,000 values. Starting with a raw vector having 9,000 values, five levels of the discrete wavelet were applied 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 dimensionally by 183 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 is calculated, as illustrated in Fig. 3. Distance3D and Simiarlity3D illustrate the distance and similarity between certain intervals and generated features, respectively.

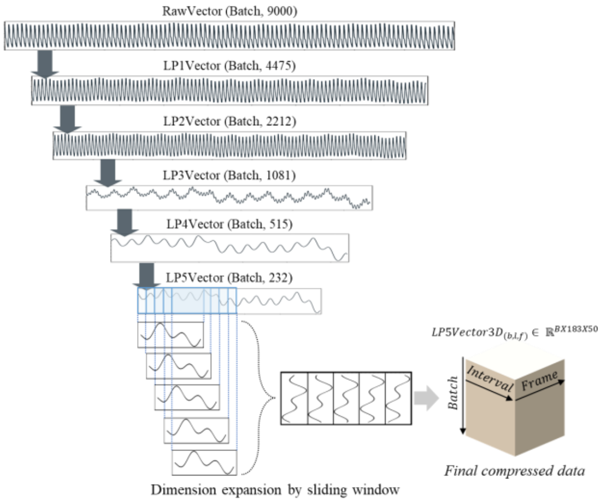
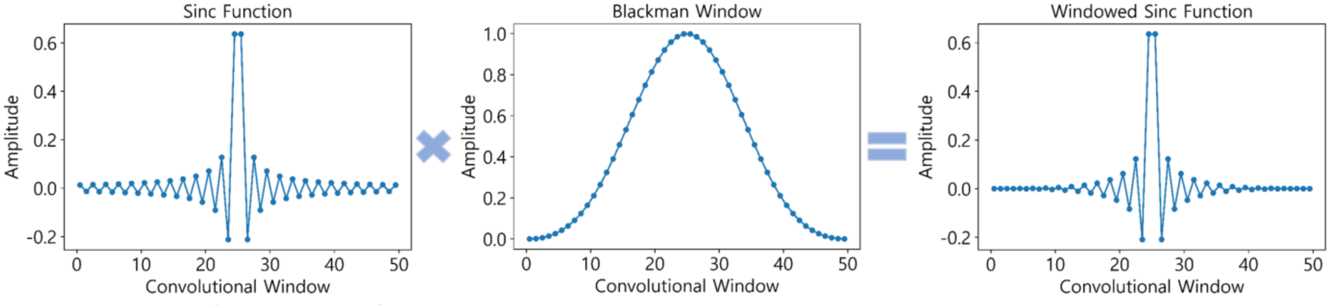


Fig. 2. Example of tensor flows through a discrete wavelet transform

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. Five Gaussian distributions, which weight the local intervals, were formed based on the five μ parameters. Next, the probability values of the five Gaussian distributions over the compressed vector were summed as a single set of weight values. Subsequently, multiplication of the weight values and similarity values was applied along the similarity axis denoted in Fig. 4. Then, only the maximum value along the similarity axis was propagated during the prediction.



$$S[n] = sinc\left(2f_c\left(n - \frac{N-1}{2}\right)\right)$$

$$(sinc(x) = \frac{\sin(\pi x)}{\pi x})$$

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

$$h[n] = S[n] * w[n]$$

$$h_{normalized}[n] = \frac{h[n]}{\sum_{i=0}^{N-1} h[i]}$$

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

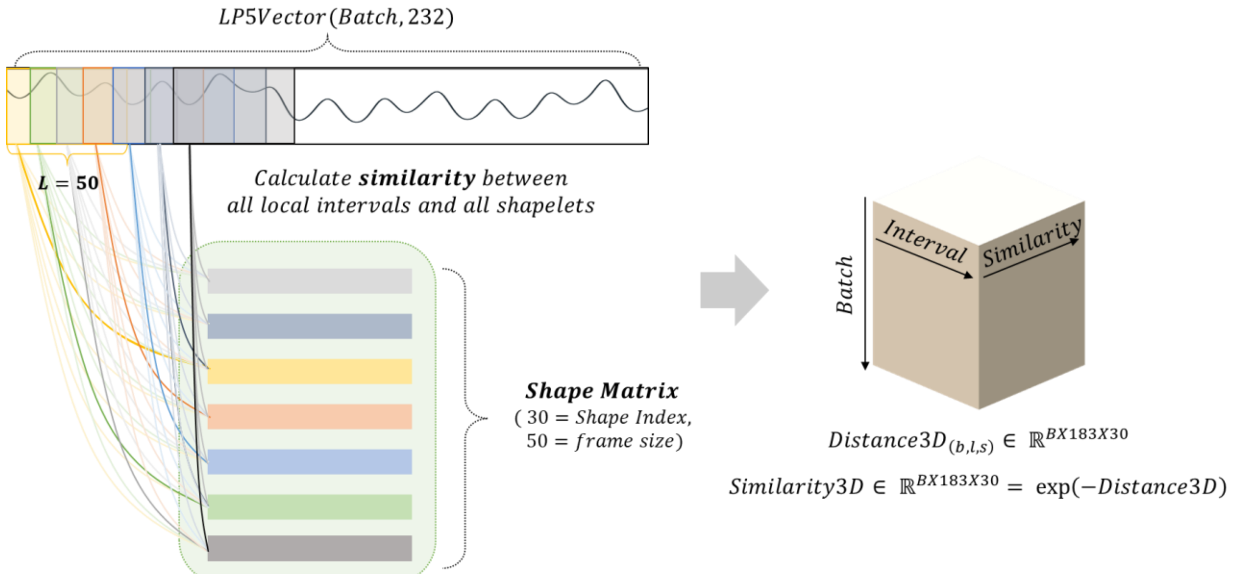

$LP5Vector(Batch, 232)$
$L = 50$
Calculate **similarity** between
all local intervals and all shapelets
Shape Matrix
(30 = Shape Index,
50 = frame size)
Batch
Interval
Similarity
$Distance3D_{(b,l,s)} \in \mathbb{R}^{BX183X30}$
$Similarity3D \in \mathbb{R}^{BX183X30} = \exp(-Distance3D)$

Fig. 3. Local shape similarity layers.

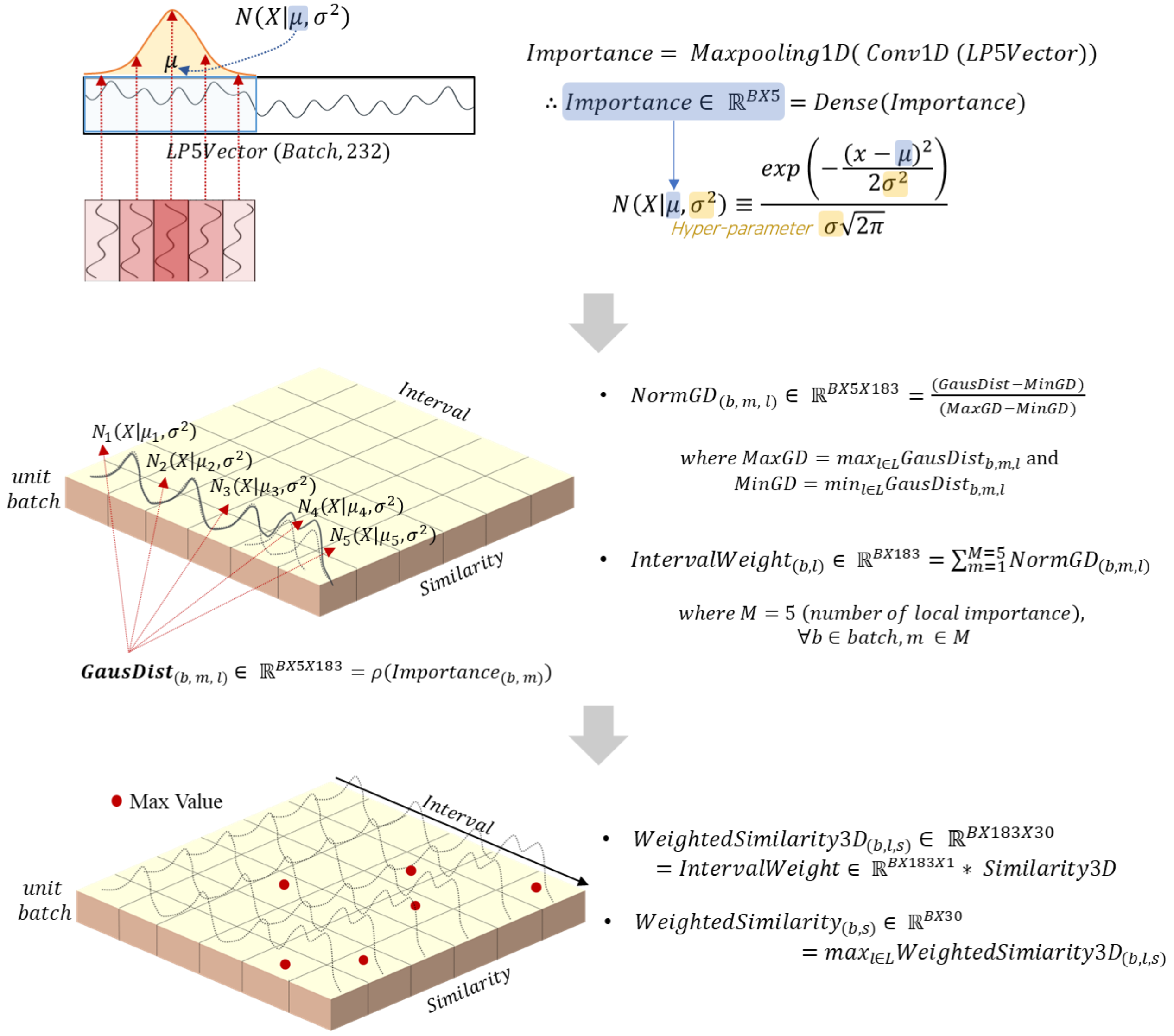

$N(X|\mu,\sigma^2)$
μ
$LP5Vector\ (Batch, 232)$
$Importance = \ Maxpooling1D(\ Conv1D\ (LP5Vector))$
$\therefore Importance \in \ \mathbb{R}^{BX5} = Dense(Importance)$
$N(X|\mu,\sigma^2) \equiv \frac{exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}$
Hyper-parameter
Interval
$N_1(X|\mu_1,\sigma^2)$
$N_2(X|\mu_2,\sigma^2)$
$N_3(X|\mu_3,\sigma^2)$
$N_4(X|\mu_4,\sigma^2)$
$N_5(X|\mu_5,\sigma^2)$
unit batch
Similarity
$GausDist_{(b,m,l)} \in \ \mathbb{R}^{BX5X183} = \rho(Importance_{(b,m)})$
• $NormGD_{(b,m,l)} \in \ \mathbb{R}^{BX5X183} = \frac{(GausDist-MinGD)}{(MaxGD-MinGD)}$
where $MaxGD = max_{l\in L}GausDist_{b,m,l}$ and
$MinGD = min_{l\in L}GausDist_{b,m,l}$
• $IntervalWeight_{(b,l)} \in \ \mathbb{R}^{BX183} = \sum_{m=1}^{M=5} NormGD_{(b,m,l)}$
where $M = 5$ (number of local importance),
$\forall b \in batch, m \in M$
● Max Value
Interval
unit batch
Similarity
• $WeightedSimilarity3D_{(b,l,s)} \in \ \mathbb{R}^{BX183X30}$
$= IntervalWeight \in \ \mathbb{R}^{BX183X1} * \ Similarity3D$
• $WeightedSimilarity_{(b,s)} \in \ \mathbb{R}^{BX30}$
$= max_{l\in L}WeightedSimiarity3D_{(b,l,s)}$

Fig. 4. Local importance layers.

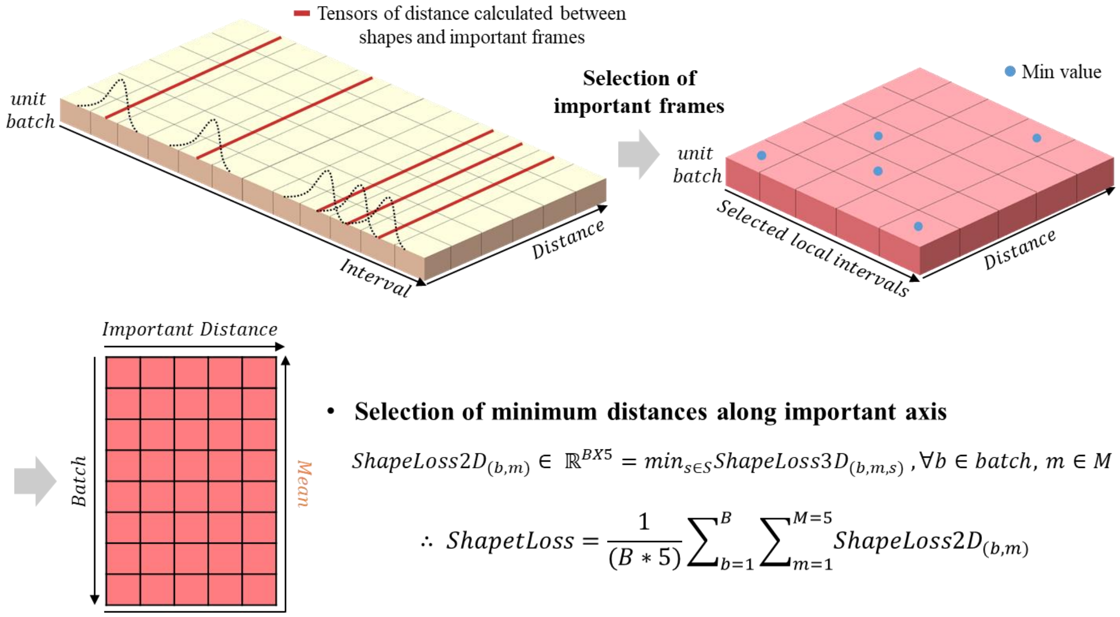

Tensors of distance calculated between shapes and important frames
unit
batch
Interval
Distance
Selection of important frames
Min value
unit
batch
Selected local intervals
Distance
Important Distance
Batch
Mean
Selection of minimum distances along important axis
$ShapeLoss2D_{(b,m)} \in \mathbb{R}^{BX5} = min_{s \in S} ShapeLoss3D_{(b,m,s)}, \forall b \in batch, m \in M$
$\therefore ShapetLoss = \frac{1}{(B*5)} \sum_{b=1}^{B} \sum_{m=1}^{M=5} ShapeLoss2D_{(b,m)}$

Fig. 5. Shape loss function.

D. Objective Function

Fig. 5 shows 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 denoted in Fig. 5 was considered during the back-propagation update. Finally, all five values at the batch level were averaged into a single loss, which was minimized during the 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 the

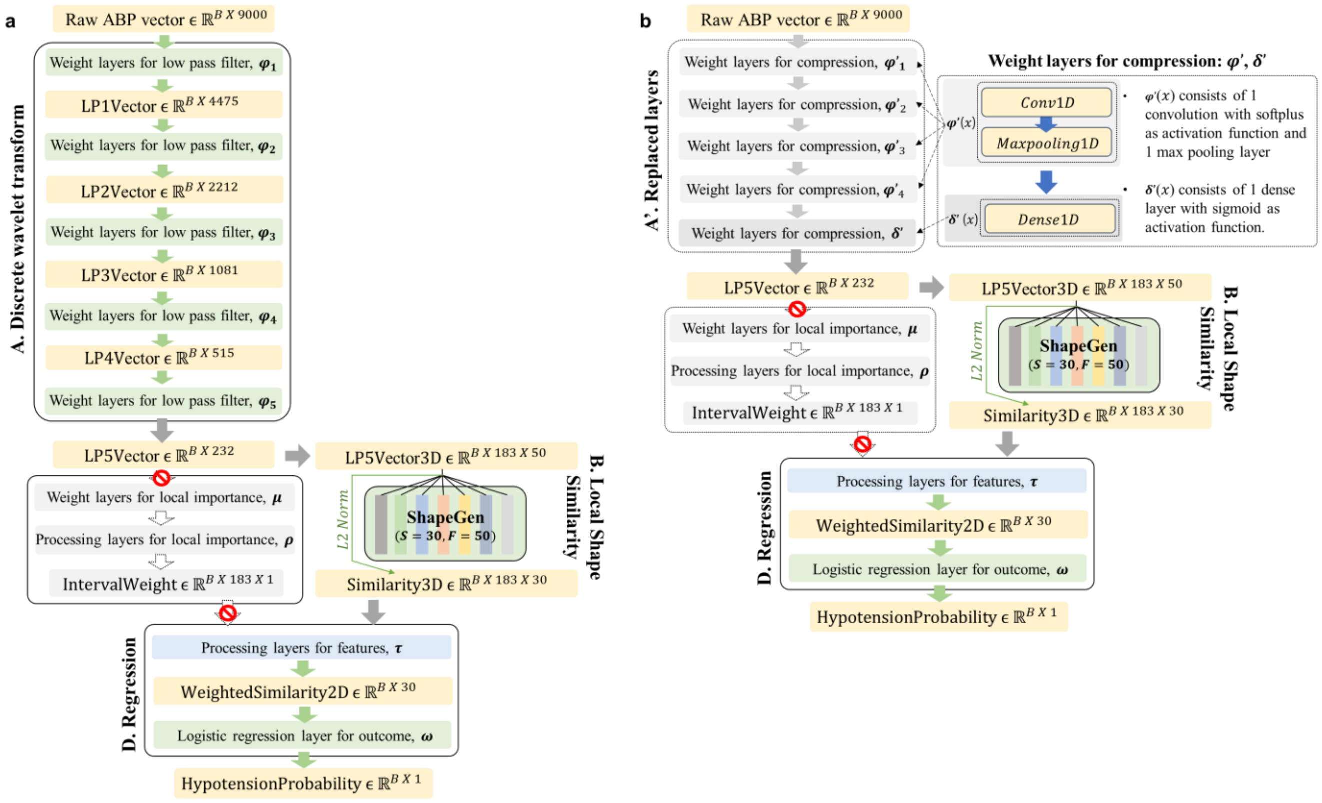

a
Raw ABP vector ϵ $\mathbb{R}^{B X 9000}$
A. Discrete wavelet transform
Weight layers for low pass filter, φ_1
LP1Vector ϵ $\mathbb{R}^{B X 4475}$
Weight layers for low pass filter, φ_2
LP2Vector ϵ $\mathbb{R}^{B X 2212}$
Weight layers for low pass filter, φ_3
LP3Vector ϵ $\mathbb{R}^{B X 1081}$
Weight layers for low pass filter, φ_4
LP4Vector ϵ $\mathbb{R}^{B X 515}$
Weight layers for low pass filter, φ_5
LP5Vector ϵ $\mathbb{R}^{B X 232}$
LP5Vector3D ϵ $\mathbb{R}^{B X 183 X 50}$
B. Local Shape Similarity
Weight layers for local importance, μ
Processing layers for local importance, ρ
IntervalWeight ϵ $\mathbb{R}^{B X 183 X 1}$
L2 Norm
ShapeGen
$(S = 30, F = 50)$
Similarity3D ϵ $\mathbb{R}^{B X 183 X 30}$
D. Regression
Processing layers for features, τ
WeightedSimilarity2D ϵ $\mathbb{R}^{B X 30}$
Logistic regression layer for outcome, ω
HypotensionProbability ϵ $\mathbb{R}^{B X 1}$
b
Raw ABP vector ϵ $\mathbb{R}^{B X 9000}$
A'. Replaced layers
Weight layers for compression, φ'_1
Weight layers for compression, φ'_2
Weight layers for compression, φ'_3
Weight layers for compression, φ'_4
Weight layers for compression, δ'
Weight layers for compression: φ', δ'
$\varphi'(x)$
Conv1D
Maxpooling1D
$\varphi'(x)$ consists of 1 convolution with softplus as activation function and 1 max pooling layer
$\delta'(x)$
Dense1D
$\delta'(x)$ consists of 1 dense layer with sigmoid as activation function.
LP5Vector ϵ $\mathbb{R}^{B X 232}$
LP5Vector3D ϵ $\mathbb{R}^{B X 183 X 50}$
B. Local Shape Similarity
Weight layers for local importance, μ
Processing layers for local importance, ρ
IntervalWeight ϵ $\mathbb{R}^{B X 183 X 1}$
L2 Norm
ShapeGen
$(S = 30, F = 50)$
Similarity3D ϵ $\mathbb{R}^{B X 183 X 30}$
D. Regression
Processing layers for features, τ
WeightedSimilarity2D ϵ $\mathbb{R}^{B X 30}$
Logistic regression layer for outcome, ω
HypotensionProbability ϵ $\mathbb{R}^{B X 1}$

Fig. 6. Architecture of models used 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 the features (τ).

In Ablation model 2 (Fig. 6b), the DWT layers were modified in addition to the removal of local importance layers. Herein, the five weight layers for the low-pass filter (φ) used in the main 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.

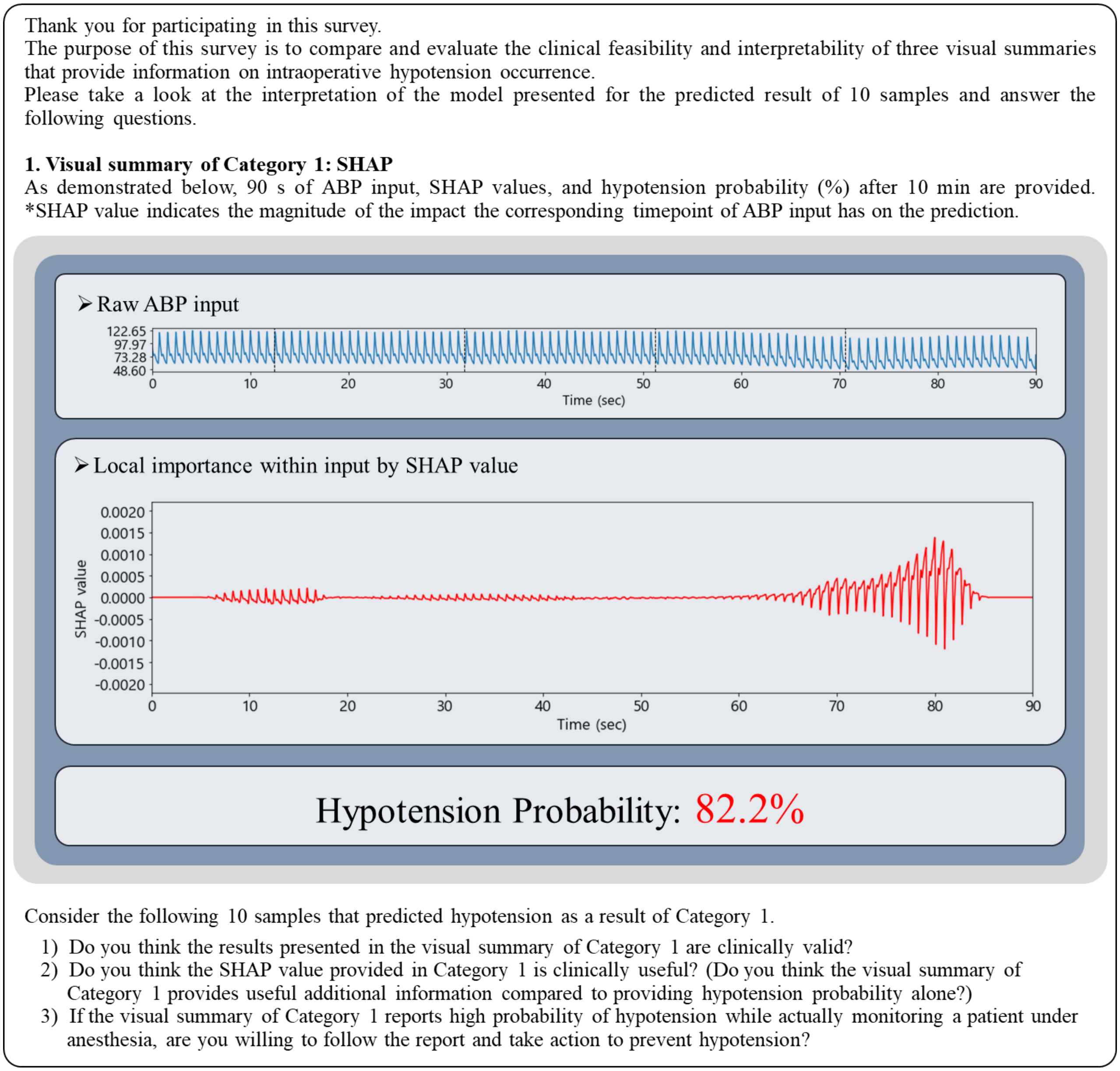
III. SPECIFICATIONS OF DATA COLLECTION IN AMC DATASET

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

IV. SPECIFICATIONS OF CLINICAL SURVEY

A. Questionnaire and Figures Presented to Participants

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



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.

Hypotension Probability: 82.2%

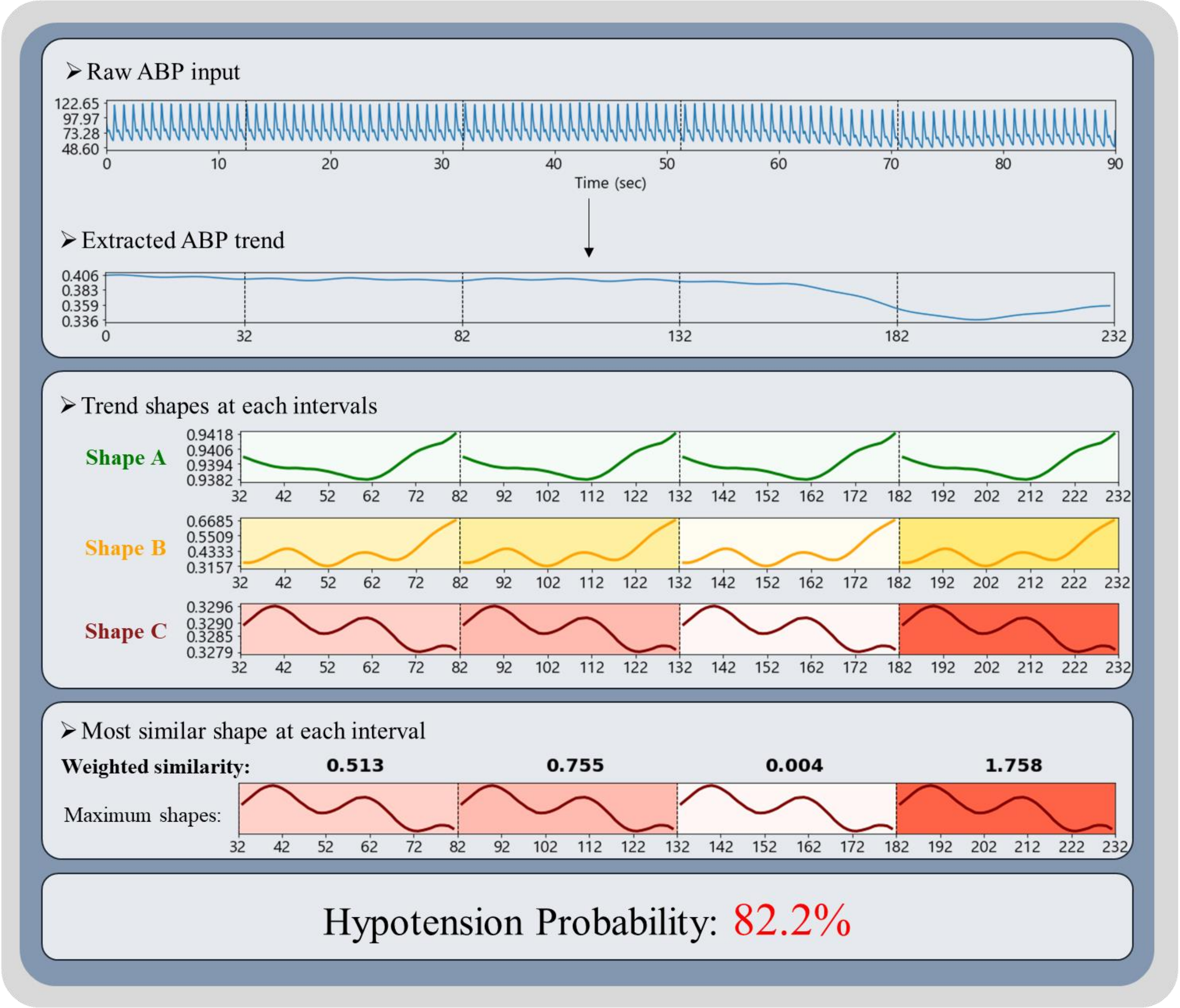
Consider the following 10 samples that predicted hypotension as a result of Category 1.

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.



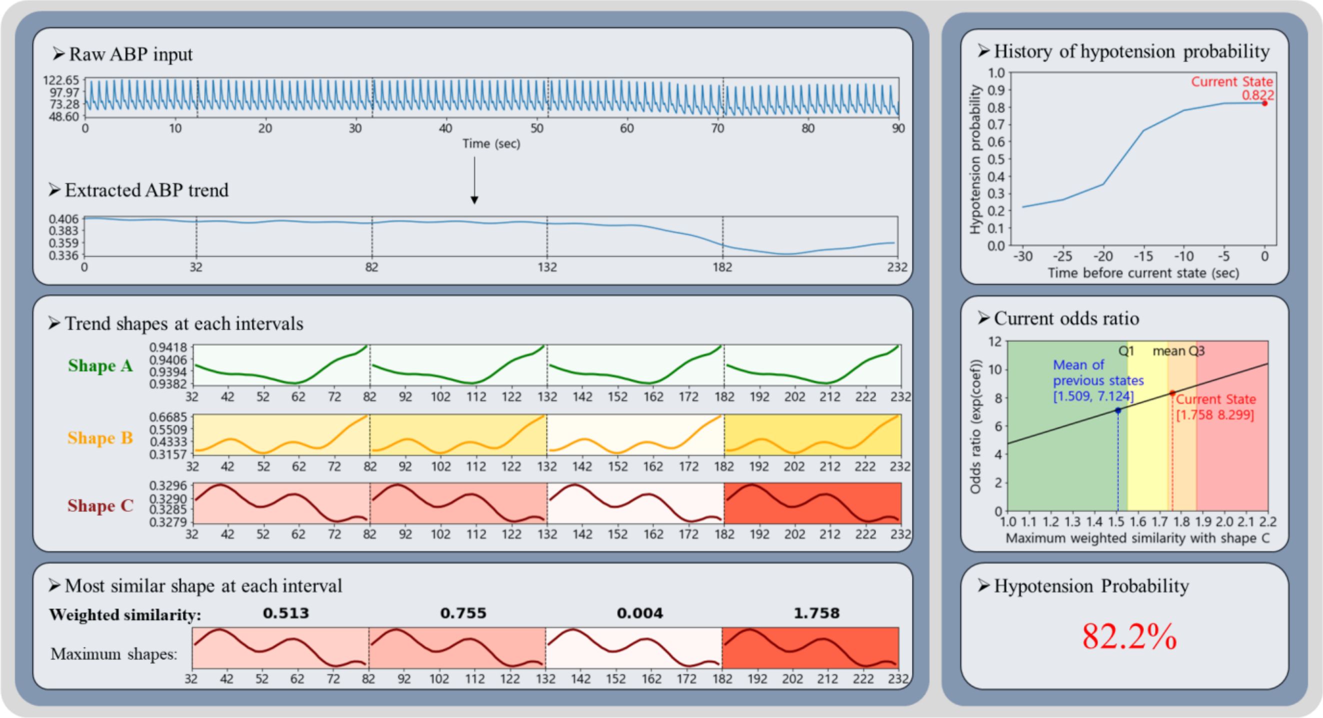
Consider the following 10 samples that predicted hypotension as a result of Category 2.

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.

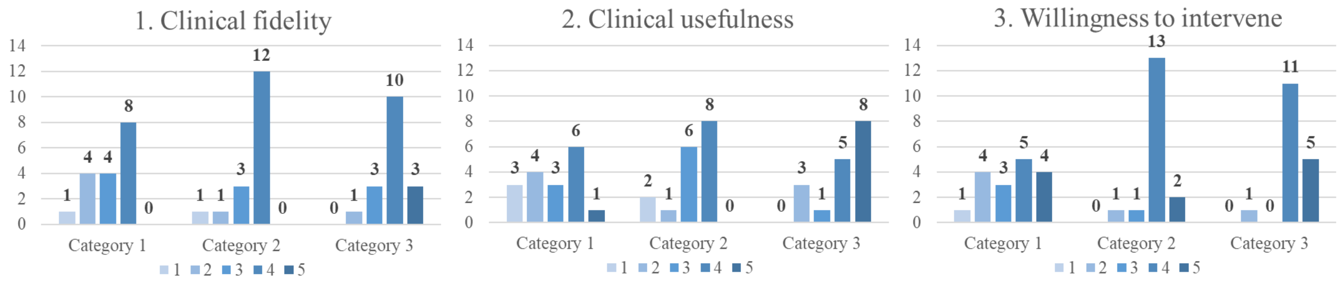


Fig. 10. Distribution of ratings for clinical survey. The x-axis for each category indicates a score of 1 to 5, and the y-axis indicates the number of patients. The number of patients rated for each 5-point Likert scale are labelled for all assessments.

B. Distribution of Ratings for Clinical Survey

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

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