OPTIMIZE MEASUREMENT FREQUENCIES OF CLINICAL VARIABLES THROUGH VARIANCE SHAP

Jiacheng Liu
Department of Computer Science
University of Minnesota, Twin Cities
Minneapolis, MN, USA
{liu00520}@umn.edu

Jaideep Srivastava
Department of Computer Science
University of Minnesota, Twin Cities
Minneapolis, MN, USA
{Srivasta}@umn.edu

ABSTRACT

Missingness and measurement frequency are two sides of the same coin. How frequent should we measure clinical variables and conduct laboratory tests? It depends on many factors such as the stability of patient conditions, diagnostic process, treatment plan and measurement costs. The utility of measurements varies disease by disease, patient by patient. In this study we propose a novel view of clinical variable measurement frequency from a predictive modeling perspective, namely the measurements of clinical variables reduce uncertainty in model predictions. To achieve this goal, we propose variance SHAP with variational time series models, an application of Shapley Additive Expanation(SHAP) algorithm to attribute epistemic prediction uncertainty. The prediction variance is estimated by sampling the conditional hidden space in variational models and can be approximated deterministically by delta's method. This approach works with variational time series models such as variational recurrent neural networks and variational transformers. Since SHAP values are additive, the variance SHAP of binary data imputation masks can be directly interpreted as the contribution to prediction variance by measurements. We tested our ideas on a public ICU dataset with deterioration prediction task and study the relation between variance SHAP and measurement time intervals.

1 INTRODUCTION

Researchers have made enormous amount of efforts to make the deep learning black boxes transparent. Both model specific and model agnostic methods have been developed to tackle the challenge of explaining the outputs of deep learning models. Among them, the game theoretic approach of SHAPHuang et al. (2022); Nie et al. (2022) (**SH**apley **A**dditive ex**P**lanations) stands out and become one of the most popular method.

Explanations about the predictive model output alone may not be enough. To gain a holistic view of predictions, we need to understand its variability. For future event predictions, such as patient deterioration prediction, it is also desirable to understand how soon the event will happen. Traditionally, these tasks can be done by training separate models against different targets. However, this approach is at the risk of inconsistent explanation and predictions among different models, thus may be difficult for human to understand. This may cause a problem when trying to translate explanation results to actions, because the causal relation between different tasks are implicit. Therefore, the only valid solution is to use one model for all tasks. Multitask learning is the most appreciated approach in this case. Alternatively, and recently, generative models are also naturally capable of performing multitasks. We propose to use variational generative models to predict patient deterioration, meanwhile, desirable quantities like prediction variability, and acuity of disease(how fast the patient is deteriorating?) can be derived from the model.

Besides the purpose of seeking explanations of black-box models, we believe the explanation of variance is crucial to the question of when and what clinical variables should we collect from patients. The solution to this question could potentially unlock methods to reduce healthcare expenses without compromising the quality of care. The problem becomes even more important in the sce-

nario of pediatric care, where for example, frequent blood draws may do more harm than the benefit gained from test results.

In this paper, we aim to exploit the locality and additivity of SHAP values, and expand model predictions to prediction variability, with the help of variational models and a bit of stochastic calculus. Variational inference methods are powerful generative models which approximate the posterior distribution of assumed hidden, unobserved variables. While in variational models, the hidden states are represented by random variables, we construct explicit and deterministic games of prediction variance, so that the SHAP values can be back propagated to input clinical variables. Further, we argue that additive SHAP values, propagated to carefully handcrafted features, can potentially translate to real actions. We provide such an example in the experiment section, where we discover that the frequency of clinical variable collections are highly correlated prediction variability. The idea is best illustrated by figure1. A variational time series model will be trained for deterioration prediction. At every time step, SHAP value explanations of predicted risk score, prediction variance are generated along with the model output. Jupyter notebooks and codes for the public dataset are available on github[1].

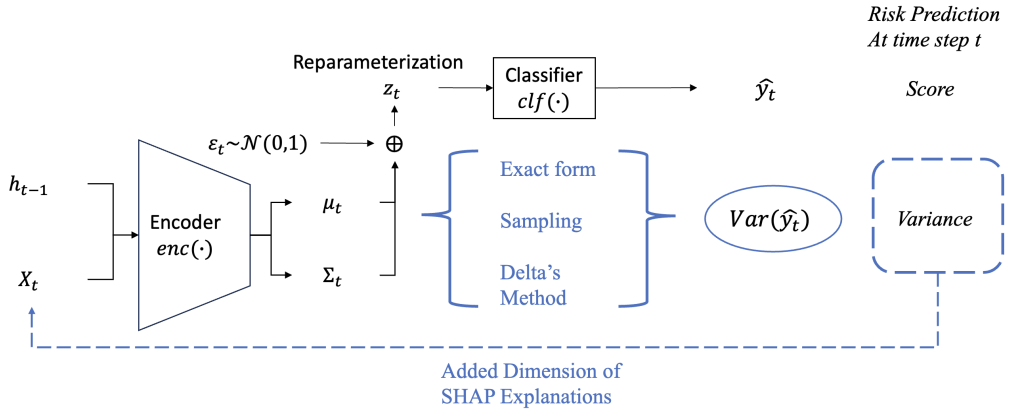

Figure 1: Proposed architecture for explaining prediction variance. Contributions in blue. The prediction variance can be deterministically approximated from the posterior distribution of hidden states via Delta's method, which is essentially a Taylor expansion.

2 PRELIMINARIES

2.1 VARIATIONAL TIME SERIES MODELS

As the name suggests, variational time series models combine variational inference with recurrent structures for time series. We refer the term *"variational time series models"* to any time series models **(i)** of which its hidden states are represented by a set of parameters of some probability distributions and **(ii)** its hidden states are updated by some recurrence mechanism, e.g. recurrent gated unit. A wide range of models fall into this category, such as variational recurrent neural networks(VRNN)Chung et al. (2015); Purushotham et al. (2016), stochastic recurrent neural networks(SRNN)Bayer & Osendorfer (2014); Fraccaro et al. (2016), and variational transformers(VTrans)Shamsolmoali et al. (2023); Lin et al. (2020).

Figure 2 shows a graphic representation of a typical variational time series model structure. Suppose the hidden space at time step t consists of several independently normally distributed variables z_t, parameterized by mean vector μ_t and a diagonal covariance matrix Σ_t. Upon every time step forward, μ_t and Σ_t are inferred by the encoder network from current input x_t and previous recurrent hidden state h_{t-1}. Next, the hidden state vector is drawn from the distribution based on inferred parameters via reparameterization tricksRezende et al. (2014); Kingma & Welling (2013). The task network then uses z_t to make predictions or classifications. As for the recurrent mechanism, both z_t

[1] https://github.com/kanbei7/VarianceSHAP_for_VRNN

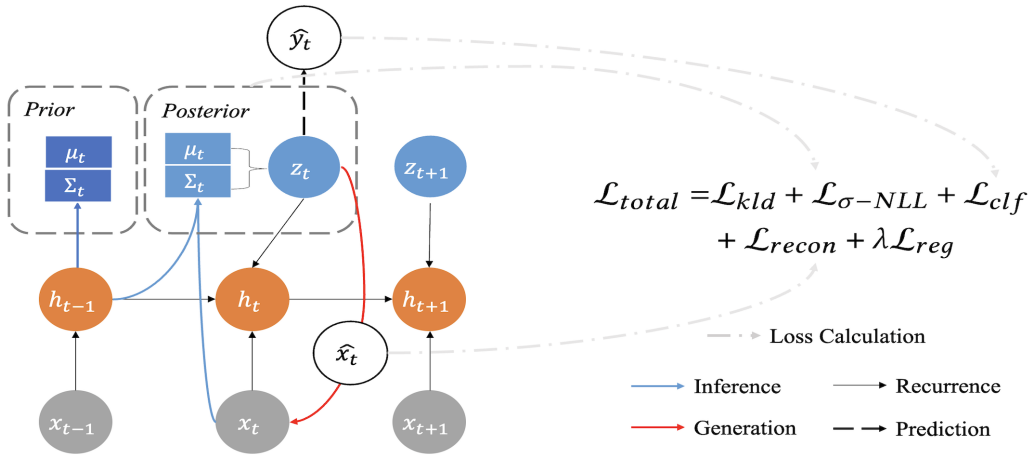

Figure 2: Variational Recurrent Models and Training Loss

and x_t serve as the inputs to the recurrence unit. In this way, the model allows for certain degrees of stochasticity or transition between hidden states. Details of the training steps can be found at appendix A.1.1.

Formally, we denote the input time series as $\mathbf{x} = (x_1, x_2, ..., x_n)$ where n is the length of the sequence, and subscription t is a dummy variable for time step. Each $x_t \in \mathbb{R}^d$, with d being the number of features. The deterministic hidden states from the recurrent model is marked by h_t, while the random variable z_t is drawn from a set of distributions parameterized by μ_t and Σ_t. θ_t is a short hand of the combination of distribution parameters μ_t and Σ_t. $\theta_{t,prior}$ stands for the parameters of prior distribution. $\hat{y}_t$ and y_t denote the predicted score and the ground truth respectively. Though the main task can be of various kind such as classification, prediction or regression, we use $CLF(\cdot)$ for the main task network. Similarly, $RNN(\cdot)$ for recurrence unit, but really it can be any recurrence mechanism like Long-Short Term MemoryHochreiter & Schmidhuber (1997), Gated Recurrent UnitChung et al. (2014) or TransformersHuang et al. (2022); Nie et al. (2022). The naming of other components are straightforward, $ENC(\cdot)$ for encoders, $DEC(\cdot)$ for decoders and $PRIOR(\cdot)$ for prior network. For parameter of distributions, $\mu_t, \Sigma_t \in \mathbb{R}^{z_dim}$. For hidden states, $h_t \in \mathbb{R}^{h_dim}$, $z_t \in \mathbb{R}^{z_dim}$.

$$\begin{aligned} \theta_{t,prior} &= PRIOR(h_{t-1}) \\ \theta_t &= ENC(h_{t-1}, x_t), \theta_t = \mu_t, \Sigma_t \end{aligned} \tag{1}$$

$$z_t = \mu_t + \Sigma_t \epsilon_t, \epsilon_t \sim N(0, 1) \tag{2}$$

$$\hat{y}_t = CLF(z_t) \tag{3}$$

$$h_t = RNN(h_{t-1}, x_t, z_t) \tag{4}$$

Reconstructed $\hat{x_t}$ is given by the decoder. Based on the result of σ-Variational Auto EncodersRybkin et al. (2021), it suffices to just out put $\hat{x_t}$ since the negative log likelihood loss is analytically determined by $\hat{x_t}$ itself. Therefore, there is no need for decoders to produce another set of parameters.

$$\hat{x_t} = DEC(h_t, z_t) \tag{5}$$

2.2 Patient Deterioration Prediction

In this study, we train variational recurrent models to predict the risk of mortality of ICU patients in the next 48 hours. Clinical deterioration (also known as "clinical decompensation") refers to the process during which the patient's condition evolves towards undesirable outcomes. Depending on the context, its meaning varies. In the emergency room, the practice of predicting deterioration is known as "triaging" . Namely, to stratify patients based on their risk of deterioration, so that patients with immediate risk will be prioritized. For patients admitted to the Intensive Care Unit (ICU),

physicians are concerned with unexpected worsening of the disease and risk of mortality. For patients in general wards, clinical deterioration usually results in critical events such as transfer to the ICU or cardiopulmonary arrest. The hope is that early predictions of onsets of clinical deterioration will eventually bring benefits to all stakeholders including patients, physicians, and insurance companies. In the scope of this paper, our models predict ICU transfers for general ward patients and risk of mortality for ICU patients. Figure **??** illustrates the n-step ahead risk of mortality predictions for ICU patients. The prediction is made at every time step.

3 VARIANCE SHAP

Although the hidden variables are modeled by parameterized distributions, the variance game(the game of attribute variance)Galeotti & Rabitti (2021); Colini-Baldeschi et al. (2018), is actually deterministic, because we can explictly and deterministically calculate the variance of $\hat{y} = clf(z)$ given $z \sim N(\mu, \Sigma)$. The only thing we need to do is to wrap the original model so that the prediction variance become the wrapped model's output. Then, SHAP method can be applied. We use $v(\cdot)$ to denote the value of a game. We disregard sampling methods for its computational cost.

For complicated clf functions, we resolve the problem by using Delta's methodOehlert (1992); Hong & Li (2018). To estimate $var[f(z)]$, where $z \sim N(\mu, \sigma)$, notice that

$$f(z) \approx f(\mu) + (z - \mu)f'(\mu) \tag{6}$$

Therefore, the variance can be estimated by,

$$var[f(z)] = var[f(z - \mu)] \cdot [f'(\mu)]^2 = \sigma^2 \cdot [f'(\mu)]^2 \tag{7}$$

4 EXPERIMENTS

Medical Information Mart for Intensive Care, or MIMIC- IV in short, is a large, open-source, deidentified database of hospitalized patientsJohnson et al. (2020; 2023); Goldberger et al. (2000). It contains clinical notes, ECG data, time series of vital signs, laboratory test results and assessment scores. In this study, we use MIMIC4 data for ICU patients which contains about 60,000 ICU stays after data cleaning. The data is aggregated to hourly level time series of varying lengths. The median length of stay is 61 hours, while the mortality rate is 7.8%. We train variational recurrent models to predict the risk of mortality in the next 48 hours. All variables are normalized and sanity checked(for example, heart rate can not be negative, saturation of oxygen must be within 0 to 100, all temperatures share the same units, etc.). The process left 176 time series variables. We pick 10 of them after extensive feature selections. In addition, a mask is associated with each variable indicating whether the value is missing and imputed or actually measured. Log base 24 is also applied to time intervals between measurements. Therefore, there is a total of 30 features per time step. The dataset is split to training, validation and test set, controlling both mortality rate length of stay.

4.1 VARIANCE SHAP AND TIME INTERVAL

To model the frequency of measurements, we simply handcraft features that represents the log time interval from the last valid measurements. In this way, by checking the variance SHAP contribution of these variables, we attempt to answer questions: how frequent we should measure a specific clinical variable? Due to the size of the data, we randomly sample 8,000 patients from training set as background. Figure 3 shows part of the results. Noticeably, we have also observed unexpected patterns for systolic blood pressure measurements. The longer the time interval between two blood pressure measurements, the lower prediction variance there will be.

4.2 AVOIDABLE AND MISSING MEASUREMENTS

For measured variables which contribute (either positively or negatively) little to both predicted risk score and prediction variance, we define these measurements as potentially "*avoidable measurements*". For variables of which its missingness contribute significantly to prediction variance, we

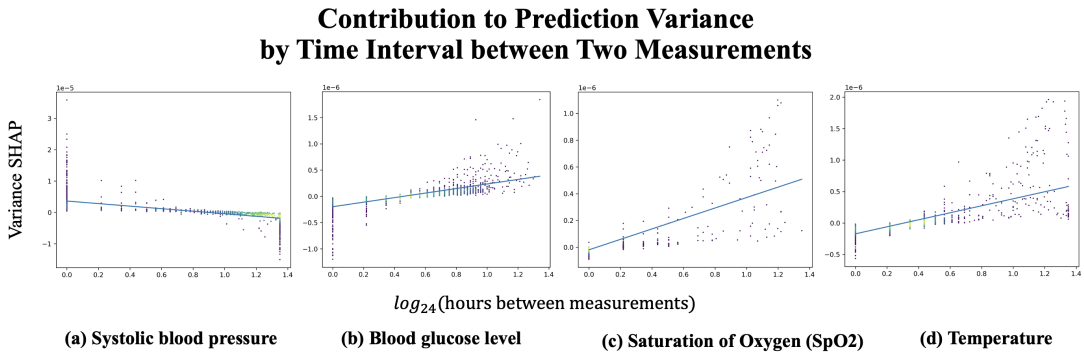

Figure 3: Variance SHAP of measurement time interval. The x-axis are log time interval in hours between two actual measurements. Log of base 24 is applied so that 1 means next measurement will happen a day later and 0 means immediate measurements.

mark these measurements as potentially "*should-have measurements*". The table below shows part of the statistics of sampled test dataset at 72 hours length of stay. We expect to have a better understanding of when more variables and lab test results are included in the prediction and variance explanation model.

5 Discussion

We have acknowledged that there works pointing to the problems of SHAP values. For example, Ismail et al. (2020) points out that SHAP does not work well with time series models. However, we didn't observe SHAP attributions as chaotic as reported in Ismail et al. (2020) (on the same MNIST dataset). Therefore, we did not apply the normalization technique of temporal saliency mapping(TSR).

There are several limitations of this work. First, we are not able to find a fair baseline or ground truth model to compare and validate the efficacy. Second, since the delta's method is an approximation of prediction variance, to gain more accurate estimations, it may be desirable to further expand the Taylor series to include the second order derivatives.Last but not least, clinical validations and further investigations are needed for

For future work, it is intriguing to study the reason behind abnormal patterns. Besides, since SHAP value measure the difference between local feature contribution to expected output, looking into the absolute value of variance contribution may also be helpful in clinical settings. Another potential application is to search for potentially avoidable order of lab tests without compromising the quality of care. Thus the cost can be reduced. This would be useful especially for pediatric care, where frequent blood draws may bring more harm than benefits.

Avoidable and Should-have Measurements for ICU patients

variable	#avoidable	#existing	%	#should-have	#missing	%
Systolic_BP	0	523	0.00%	709	1477	48.00%
Glucose	161	366	43.99%	640	1634	39.17%
Spo2	711	1811	39.26%	140	189	74.07%
Temperature	133	497	26.76%	542	1503	36.06%

Table 1: Avoidable and Should-have Measurements of Sampled patients at time = 72 hours

A APPENDIX

A.1 EXPERIMENT DETAILS

Due to the page limit, we apologize for omitting some details. However, many questions can be answered by our 8-page preprint paper[2] on ArxivLiu & Srivastava (2024). The paper and codes were made availble on Github before this workshop deadline[3].

A.1.1 TRAINING

Since the integral of joint probability $p(x, z)$ over the entire hidden space,

$$\int_{\mathbf{Z}} p(x, z')\, dz'$$

, is intractable, variational are trained on evidence lower bound (ELBo), given by.

$$\begin{aligned} ELBo = & lnp_\theta(x|z,h) \\ & - KLdivergence[p_\theta(z|x,h)||p_{\theta_{prior}}(h)] \end{aligned} \tag{8}$$

Maximizing ElBo is equivalent to minimizing $\mathcal{L}_{kld} + \mathcal{L}_{NLL}$, defined below. Subscription t applies to all variables above, hence is omitted.

The training of variational time series models shares similarities with variational auto encoders (VAE)Rezende et al. (2014); Kingma & Welling (2013); Sohn et al. (2015): a Kullback–Leibler divergence between posterior θ_t and prior $\theta_{t,prior}$ and a regularization loss on θ_t . We adopt the approach in σ-Variational Auto EncodersRybkin et al. (2021) such that:

$$\begin{aligned} \mathcal{L}_{kld} &= KLdivergence[p_{\theta_t}(z_t|x_t,h_t)||p_{\theta_{t,prior}}(h_t)] \\ & p_{\theta_t}(z_t|x_t,h_t) \sim N(\mu_t, \Sigma_t) \\ & p_{\theta_{t,prior}}(z_t|h_t) \sim N(\mu_{t,prior}, \Sigma_{t,prior}) \end{aligned} \tag{9}$$

$$\begin{aligned} \mathcal{L}_{\sigma-NLL} &= NLL(\hat{x_t}, x_t, log(\sigma)) \\ log(\sigma) &= log(\sqrt{(\hat{x_t} - x_t)^2}) \end{aligned} \tag{10}$$

Notice that there are different ways of choosing a prior, depending on the specific problem settings. These two losses are essential components for variational inference.

Additionally, we have the prediction or classification loss from the main task network and the reconstruction loss to train recurrent networks. MSE denotes the mean squared error.

$$\begin{aligned} \mathcal{L}_{clf} &= Cross - Entropy(\hat{y_t}, y_t) \\ \mathcal{L}_{recon} &= MSE(\hat{x_t}, x_t) \end{aligned} \tag{11}$$

Therefore the total loss per time step is given by the following equation.

$$\begin{aligned} \mathcal{L}_{total} = & \mathcal{L}_{kld} + \mathcal{L}_{\sigma-NLL} + \mathcal{L}_{clf} \\ & + \mathcal{L}_{recon} + \lambda\mathcal{L}_{reg} \end{aligned} \tag{12}$$

Depending on whether a complete sequence can be seen at the time of inference, the loss may be average over all the time step. Also, regularization terms $\mathcal{L}_{reg}$ may be appended to stabilize and accelerate training. λ controls the strength of regularization.

A.1.2 FEATURE SELECTION AND BASELINE MODEL

10 features with the least missingness rate are selected. They are:

- Diastolic blood pressure

[2] with a different name, to avoid potential breach of double-blind review rule.

[3] https://github.com/kanbei7/VarianceSHAP_for_VRNN

- Systolic blood pressure
- Arterial Mean blood pressure
- Fraction of inspired O_2
- Glucose, blood
- Heart rate
- PH value, blood
- Respiration rate
- Saturation of Oxygen
- Body temperature

We have compared VRNN models trained on 10, 30, 50 and all time series features(176). No significant performance gap was found when trained on the same random split with various number of features. AUROC and AUPR are reported for 5 different $70-15-15$ random splits (70%training, 15%validation, 15%test) for 10-feature VRNN, 30-feature VRNN in the following table2.

	AUPR		**AUROC**	
exp id	10-feature	30-feature	10-feature	30-feature
1	0.4183	0.4141	0.8194	0.8196
2	0.4158	0.4204	0.8196	0.8224
3	0.4104	0.411	0.814	0.8206
4	0.4149	0.4179	0.8226	0.8225
5	0.4178	0.4193	0.8236	0.8237
mean	0.41544	0.41654	0.81984	0.82176
median	0.4158	0.4179	0.8196	0.8224
std	0.00314531	0.00390551	0.00374540	0.00163799
t-test, two sided				
p-value	0.5011		0.2095	

Table 2: Base variational model performance on random split MIMIC-IV 48-hr ahead deterioration prediction task.

A.1.3 More Results

Notice that for the same clinical feature, different patterns of prediction SHAP - variance SHAP exists, as shown in fig4

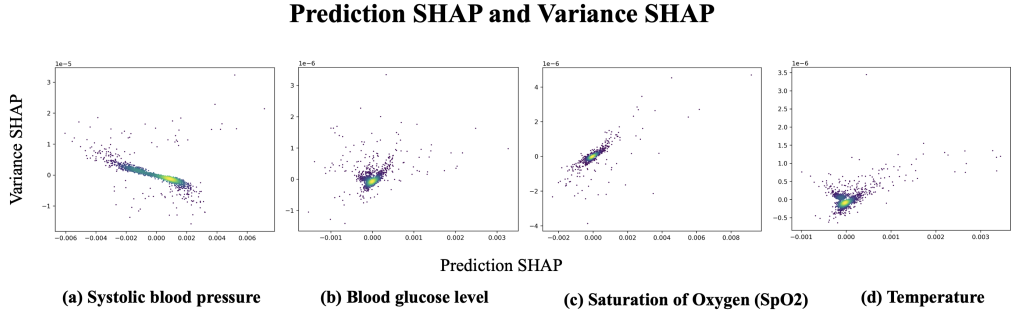

Figure 4: Relation between the contribution to predicted mortality risk (prediction SHAP) and contribution to prediction variance (variance SHAP)

We have observed unexpected patterns for blood pressure measurements. Notice that for blood pressure, the zero score no longer falls into normal reference range. The average blood pressure of

MIMIC-IV ICU population is well above the reference range. Therefore, the interpretation may be different.

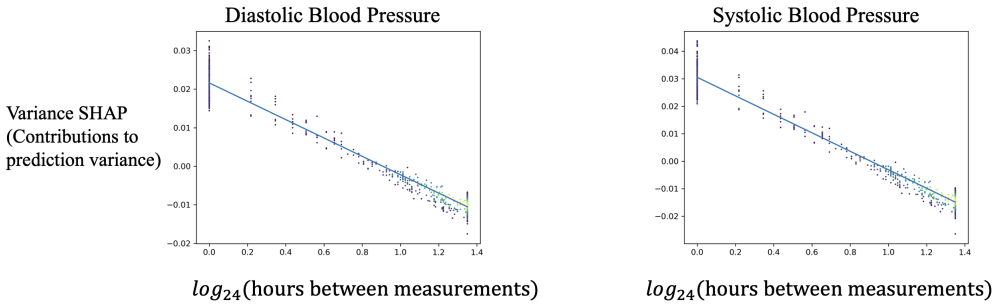

Figure 5: Unexpected patterns of blood pressure. There exist negative correlation predicction variance and between time interval

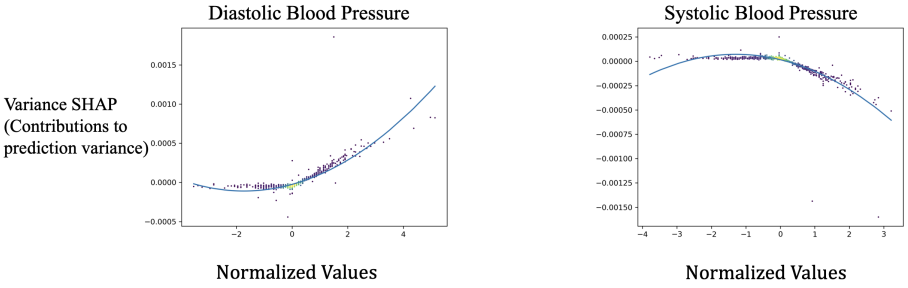

Figure 6: Unexpected patterns of blood pressure.

A.2 Related Work

A.2.1 Shapely Values and DeepSHAP

Since the advent of SHAPLundberg & Lee (2017); Lundberg et al. (2020), it has been extensively applied to many areas, such as geologyAl-Najjar et al. (2023), financeGodin et al. (2023); Duan et al. (2022) and healthcareChen et al. (2022); Lundberg et al. (2018). Recently, SHAP values have been applied to gaussian processChau et al. (2023). The authors have shown that the variance of SHAP value (which is necessary for inferences of gaussian process) is not the SHAP value of the variance game(our focus in this paper). SHAP has also been combined with variational auto encoders (VAE) to explain feature contributionsOlsen et al. (2022); Withnell et al. (2021). We appreciate the beauty of SHAP values since it is model agnostic and flexible to most type of model outputs. To take a step further, if there exists a function calculating the variance of the prediction, SHAP methods can be applied as well!

We also note that there are plenty of approaches other SHAP values, especially for time series dataLim et al. (2021); Castro et al. (2023); Crone & Kourentzes (2010); Taieb & Atiya (2015). However, as argued in the previous section, we claim that SHAP is the best fit for our case. Besides, studies have shown several drawbacks of SHAP, noticeably with entangled time series featuresIsmail et al. (2020). But as we shall see in later sections, with the power of variational time series models, this defect is mitigated by the inference of independent hidden state variables and thus not a major concern to our topic.

A.2.2 PREDICTION UNCERTAINTY AND VARIANCE

The predicted probability score by machine learning models contains two sources of uncertainty, aleatoric and epistemic uncertaintyHüllermeier & Waegeman (2021); Swiler et al. (2009). While the latter one comes from the uncertainty about model parameters, aleatoric originates from data and unobserved factors. With aleatoric uncertainty and variance of prediction scores in focus, Bayesian methodsGal & Ghahramani (2016); Kendall & Gal (2017); Neal (2012) and variational inference methodsKingma & Welling (2013); Rezende et al. (2014) become natural choices for estimating prediction variance, with the assumption that hidden state random variables approximate the posterior distribution of inputs. This fundamental assumption should hold true for every variational generative model, for models to be fully effective. In this study we will focus on how to explain the contributions of input clinical features to prediction variance.

A.2.3 EXPLAINABLE AI FOR TIME SERIES IN HEALTHCARE

In an application area like healthcare, explainability and interpretability are a crucial features to build trustworthy machine learning and AI applications. Di Martino & Delmastro (2022) has a nice summary of recent development of recent Explainable Artificial Intelligence(XAI) in healthcare. SHAP values is the dominant approach in recent studies about explainable healthcare machine learning modelsAng et al. (2021); Ukil et al. (2022),. While few studies have focused on explaining deep learning modelsWithnell et al. (2021); Oviedo et al. (2019), fewer studies have focused on explaining time series in healthcareLundberg et al. (2018); Ivaturi et al. (2021). The most related work is that Withnell et al. (2021) propose to use variational auto encoders to study multi-omics data for cancer diagnostic. However, to the best of our knowledge, we are the first to study contributions of prediction variance and the first to attempt to study the relation between variance SHAP and the frequency of variable measurements.

A.3 MNIST AS AN EXAMPLE

We verify our proposed method by training a unidirectional variational recurrent neural network on MNIST. The training took 10 epochs and achieved an accuracy of 98%. Details are provided in the Jupyter notebook.Figure7,8,9 compare (a)SHAP values of predicted class probabilities, (b) Variance of the prediction SHAP, (c)(proposed)SHAP value of prediction variance attribution. The first thing we noticed is that as expected, the attribution of variance and prediction do not coincide. Namely the model can be very confident on the prediction with a high predicted probability score(high score with low variance), or the model can be sure that the input instance is not one of the target class (low score with low variance). Vice versa, there could be the case where high score and high variance coexist. Additionally, as a note to the discussion in related work, we didn't observe SHAP attributions as chaotic as reported in Ismail et al. (2020) (on the same MNIST dataset). Therefore, we did not apply the normalization technique of temporal saliency mapping(TSR).

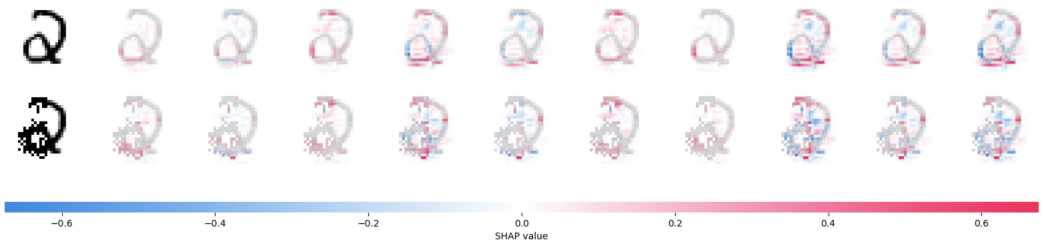

Figure 7: Prediction SHAP value.

REFERENCES

Husam AH Al-Najjar, Biswajeet Pradhan, Ghassan Beydoun, Raju Sarkar, Hyuck-Jin Park, and Adbullah Alamri. A novel method using explainable artificial intelligence (xai)-based shapley additive explanations for spatial landslide prediction using time-series sar dataset. *Gondwana Research*, 123:107–124, 2023.

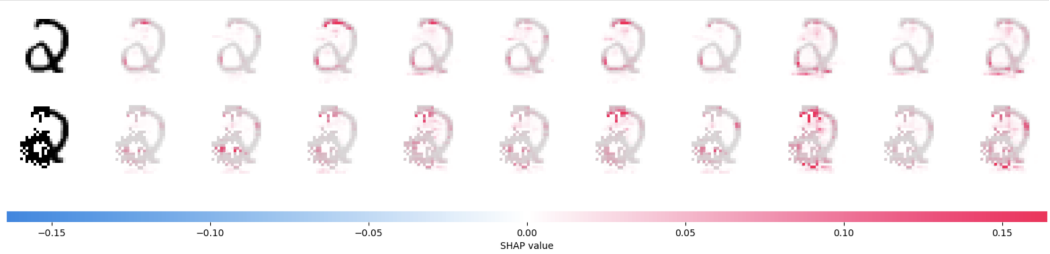

Figure 8: Variance of the prediction SHAP value

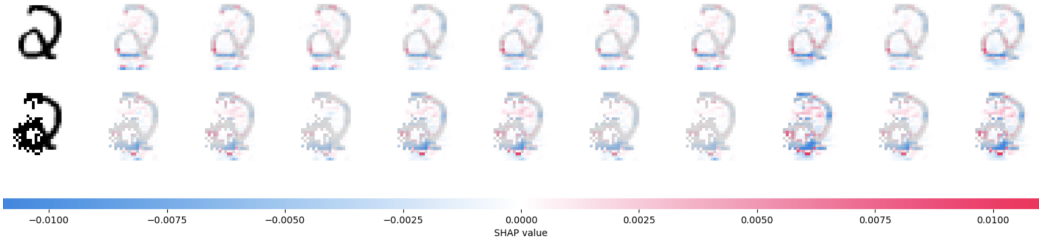

Figure 9: SHAP value of prediction variance attribution.

E T Ang, M Nambiar, Y S Soh, and V Y Tan. An interpretable intensive care unit mortality risk calcula- tor. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biol- ogy Society (EMBC)*, pp. 4152–4158. IEEE, 2021.

Justin Bayer and Christian Osendorfer. Learning stochastic recurrent networks. *arXiv preprint arXiv:1411.7610*, 2014.

Manuel Castro, Pedro Ribeiro Mendes Júnior, Aurea Soriano-Vargas, Rafael de Oliveira Werneck, Maiara Moreira Gonçalves, Leopoldo Lusquino Filho, Renato Moura, Marcelo Zampieri, Oscar Linares, Vitor Ferreira, et al. Time series causal relationships discovery through feature importance and ensemble models. *Scientific Reports*, 13(1):11402, 2023.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *arXiv preprint arXiv:2305.15167*, 2023.

Hugh Chen, Scott M Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature communications*, 13(1):4512, 2022.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. December 2014.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

Riccardo Colini-Baldeschi, Marco Scarsini, and Stefano Vaccari. Variance allocation and shapley value. *Methodology and Computing in Applied Probability*, 20:919–933, 2018.

Sven F Crone and Nikolaos Kourentzes. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. *Neurocomputing*, 73(10-12):1923–1936, 2010.

Flavio Di Martino and Franca Delmastro. Explainable ai for clinical and remote health applications: a survey on tabular and time series data. *Artificial Intelligence Review*, 56(6):5261–5315, October 2022. ISSN 1573-7462. doi: 10.1007/s10462-022-10304-3. URL `http://dx.doi.org/10.1007/s10462-022-10304-3`.

Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4468–4476, 2022.

Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. *Advances in neural information processing systems*, 29, 2016.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pp. 1050–1059, 2016.

Marcello Galeotti and Giovanni Rabitti. On the comparison of shapley values for variance and standard deviation games. *Journal of Applied Probability*, 58(3):609–620, 2021.

Frédéric Godin, Emmanuel Hamel, Patrice Gaillardetz, and Edwin Hon-Man Ng. Risk allocation through shapley decompositions, with applications to variable annuities. *ASTIN Bulletin: The Journal of the IAA*, 53(2):311–331, 2023.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Han Hong and Jessie Li. The numerical delta method. *Journal of Econometrics*, 206(2):379–394, 2018.

Feiqing Huang, Kexin Lu, CAI Yuxi, Zhen Qin, Yanwen Fang, Guangjian Tian, and Guodong Li. Encoding recurrence into transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33:6441–6452, 2020.

Praharsh Ivaturi, Matteo Gadaleta, Amitabh C Pandey, Michael Pazzani, Steven R Steinhubl, and Giorgio Quer. A comprehensive explanation framework for biomedical time series classification. *IEEE journal of biomedical and health informatics*, 25(7):2398–2408, 2021.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. Variational transformers for diverse response generation. *arXiv preprint arXiv:2003.12738*, 2020.

Jiacheng Liu and Jaideep Srivastava. Explain variance of prediction in variational time series models for clinical deterioration prediction, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Radford M Neal. *Bayesian Learning for Neural Networks*. Springer Science and Business Media, 2012.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

Lars HB Olsen, Ingrid K Glad, Martin Jullum, and Kjersti Aas. Using shapley values and variational autoencoders to explain predictive models with dependent mixed features. *Journal of machine learning research*, 23(213):1–51, 2022.

Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L DeCost, Siyu I P Tian, Giuseppe Romano, Aaron Gilad Kusne, and Tonio Buonassisi. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *Npj Comput. Mater.*, 5(1), May 2019.

Sanjay Purushotham, Wilka Carvalho, Tanachat Nilanon, and Yan Liu. Variational recurrent adversarial deep domain adaptation. In *International Conference on Learning Representations*, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective vae training with calibrated decoders. In *International Conference on Machine Learning*, pp. 9179–9189. PMLR, 2021.

Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, Dacheng Tao, and Xuelong Li. Vtae: Variational transformer autoencoder with manifolds learning. *arXiv preprint arXiv:2304.00948*, 2023.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Laura P Swiler, Thomas L Paez, and Randall L Mayes. Epistemic uncertainty quantification tutorial. In *Proceedings of the 27th International Modal Analysis Conference*, 2009.

Souhaib Ben Taieb and Amir F Atiya. A bias and variance analysis for multistep-ahead time series forecasting. *IEEE transactions on neural networks and learning systems*, 27(1):62–76, 2015.

Arijit Ukil, Leandro Marin, and Antonio J Jara. When less is more powerful: Shapley value attributed ablation with augmented learning for practical time series sensor data classification. *Plos one*, 17(11):e0277975, 2022.

Eloise Withnell, Xiaoyu Zhang, Kai Sun, and Yike Guo. Xomivae: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, 22(6):bbab315, 2021.