Uncertainty-Aware Deep Ensembles for Reliable and Explainable Predictions of Clinical Time Series

Kristoffer Wickstrøm, Karl Øyvind Mikalsen, Michael Kampffmeyer, Arthur Revhaug, and Robert Jenssen

***Abstract*—Deep learning-based support systems have demonstrated encouraging results in numerous clinical applications involving the processing of time series data. While such systems often are very accurate, they have no inherent mechanism for explaining what influenced the predictions, which is critical for clinical tasks. However, existing explainability techniques lack an important component for trustworthy and reliable decision support, namely a notion of uncertainty. In this paper, we address this lack of uncertainty by proposing a deep ensemble approach where a collection of DNNs are trained independently. A measure of uncertainty in the relevance scores is computed by taking the standard deviation across the relevance scores produced by each model in the ensemble, which in turn is used to make the explanations more reliable. The class activation mapping method is used to assign a relevance score for each time step in the time series. Results demonstrate that the proposed ensemble is more accurate in locating relevant time steps and is more consistent across random initializations, thus making the model more trustworthy. The proposed methodology paves the way for constructing trustworthy and dependable support systems for processing clinical time series for healthcare related tasks.**

***Index Terms*—deep learning, ensembles, interpretability, uncertainty, time series**

I. Introduction

CLINICAL data stored in electronic health records (EHRs) contain valuable information that can be used for e.g. diagnosis support [1]. The type of data stored in EHRs can vary between a number of different modalities, for instance free text (e.g. nursing notes) or clinical time series (e.g. blood or temperature measurements). Recent advances in machine learning have shown how such information can be extracted from EHRs and used to construct data-driven algorithms, which can serve as support systems that aid medical practitioners in decision making [2]. Particularly, systems based on deep neural networks (DNNs) have shown promising results on a number of tasks such as mortality prediction [2], detection of infections [3], and patient treatment trajectory prediction [4].

While DNNs often provide accurate predictions, they have no inherent mechanism for explaining what influenced the predictions. This has been noted on numerous occasions, and has resulted in DNNs often being referred to as black-boxes [5]. A recent study found that it is crucial to; 1) know the subset of features deriving the model outcome and 2) provide a measure of uncertainty for predictions for creating trustworthy machine learning-based support systems [6]. These mechanisms are not built into DNNs, but recent advances in explainable artificial intelligence (XAI) have made great leaps in developing methods that provide interpretations for the prediction of a model [7].

One promising method for XAI is based on the so-called class activation mapping (CAM) method [9]. CAM was originally developed for image data, but has also been shown to be applicable for temporal data such as clinical time series [8, 10]. By utilizing DNNs with a particular processing structure, CAM can assign a score to each time step in an input that indicates how important that time step is for a given prediction. This score will be referred to as a relevance score. Explainable methods that provide a notion of uncertainty are lacking, but needed to provide trustworthy and dependable support systems. For instance, if a clinical measurement is identified as being highly relevant for a prediction, how certain is this relevance score? Or if several DNNs are trained from different random initialization, will the same measurements be highlighted as relevant across the different models? Such questions cannot be answered within the standard framework of DNNs. If explanations are accepted without taking uncertainty into account it might results in an unjustified belief in the explanations.

In this work we propose a deep ensemble approach to model uncertainty in explainability for DNN-based predictions of clinical time series. A collection of DNNs are trained independently, each producing a prediction and a relevance score for each time step. The uncertainty in the relevance scores is computed by taking the standard deviation across the relevance scores produced by each model in the ensemble. Intuitively, time steps that all models indicate as relevant will be highlighted as certainly relevant. Time steps that only one or some models highlight as relevant will be highlighted as relevant, but with a high degree of uncertainty. To the best of our knowledge, such a deep ensemble approach for uncertainty in explainable DNNs for clinical tasks has not been previously explored. The proposed approach is validated on synthetic data and on two clinically relevant tasks; myocardial infarction detection in echocardiograms (ECGs) and surgical site infection (SSI) in blood measurements of C-reactive protein (CRP). Experiments show how the deep ensemble is more accurate at locating relevant time steps and more consistent. Consistent means that the model highlights similar time steps as relevant

K. Wickstrøm, KØ. Mikalsen, M. Kampffmeyer, and R. Jenssen are with the UiT Machine Learning Group at the Dept. of Physics and Technology, UiT the Arctic University of Norway, Norway, Tromsø, NO-9037, e-mail: kwi030@uit.no.

A. Revhaug is with the Dept. of Clinical Medicine, UiT the Arctic University of Norway, Tromsø, Norway.

KØ. Mikalsen is also with the Dept. of Gastrointestinal Surgery, University Hospital of North Norway (UNN), Tromsø, Norway.

M. Kampffmeyer and R. Jenssen are also with the Norwegian Computing Center, Dept. SAMBA, P.O. Box 114 Blindern, NO-0314 Oslo, Norway.

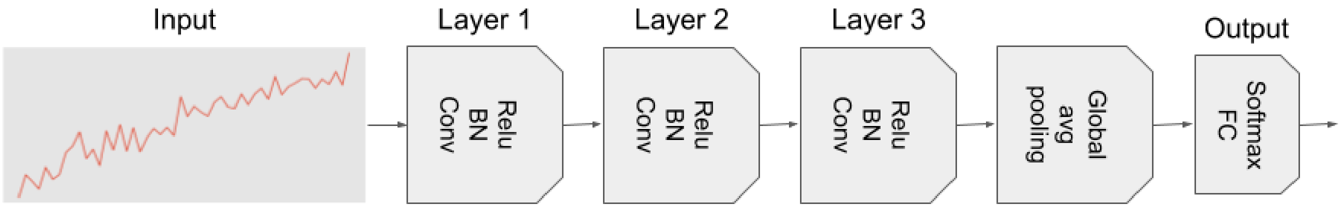


Fig. 1. Figure illustrates the architecture of the FCN proposed by Wang *et al.* (2017) [8]. Layer 1, 2, and 3 are convolutional blocks that consists of a convolutional operation, followed by batch normalization and a ReLU activation function. Block 1, 2, and 3 contains 128, 256, and 128 filters of size 7, 5, and 3, respectively. The three convolutional blocks are followed by a global average pooling layer that averages over the time dimension of the out output of the last convolutional block. Lastly, the output layer of the model is a fully connected layer followed by a softmax activation function. At each convolutional block, the input is zero-padded such that the length of time series does not diminish. This makes FCNs particularly suited for utilizing the CAM technique for explainable CNNs, as a relevance score can be computed for each time step in the time series.

when retrained from different initializations. Further, the value of the uncertainty measures obtained by the proposed methodology is demonstrated through several qualitative experiments. Although in this work the method is illustrated with the CAM approach, it is applicable for any explainability technique for DNNs. The proposed approach paves they way for increasing trustworthiness of DNNs and can be an important component in constructing dependable and transparent decision support systems.

II. Related Work

Recently there has been a great increase in methods for creating explainable DNNs. This section will describe the recent works that are most closely related to this paper. For a more comprehensive review the reader is referred to a summary of XAI by Samek *et al.* (2017) [7] and an overview of XAI in healthcare by Tonekaboni *et al.* (2019) [6].

Many approaches have been proposed to create interpretable decision support systems based on DNNs. Zhang *et al.* (2018) proposed a model based on recurrent neural networks (RNNs) that could learn an interpretable deep representation that was personalized for each patients [11]. This was achieved through the attention mechanism [12], which was used to indicate the relative importance of different features to the personalized embedding of a patient. Assaf and Schumann (2019) proposed a gradient-based interpretablity approach for convolutional neural networks (CNNs) that handles multivariate time series through a two-stage approach [10]. They demonstrate how their approach can be used to explain which features during a time interval are important for a given prediction. Tonekaboni *et al.* (2020) introduced a method that automatically assigns an importance values to each features at each time step by simulating counterfactual trajectories given previous observations [13]. However, apart from Wickstrøm *et al.* (2020) [14] who proposed an approach for providing uncertainty measures for input feature importance in computer vision tasks, the issue of uncertainty in input feature importance have, to the best of our knowledge, not been explored in the context of EHRs.

III. Fully Convolutional Networks

Several recent works have shown that CNNs can achieve state-of-the-art performance on time series classification tasks, and are usually easier to train than RNNs [8, 15]. In this paper we use a network similar to the fully convolutional network (FCN) proposed by Wang *et al.* (2017) [8], which has demonstrated good performance on time series classification benchmarks [8]. The FCN consists of three convolutional blocks, each consisting of a convolution operation, batch normalization [16] and a rectified linear unit (ReLU), a global average pooling operation and a fully connected layer followed by a softmax activation function for the output layer. The convolutional blocks and the pooling layer can be considered the encoder part of the model, while the fully connected layer combined with the softmax function constitute the classifier of the model. An illustration of the model is shown in Figure 1. The first convolutional block consists of 128 filters with size 7, the second of 256 filters with size 5, and the third convolutional block has 128 filters with size 3. An important component of the FCN is that the input is zero-padded such that the length of the time series does not change through the three convolutional blocks. This is vital for the interpretablity technique that is discussed in Section IV. The global average pooling operation, which takes the average over the entire time dimension, summarizes the content of the filtered time series and produces a single value for each filter in the last convolutional block.

IV. Class Activation Mapping

CAM is an interpretability method designed for CNNs that highlights class-specific relevance scores in the input data [9]. CAM have recently shown encouraging results for tasks involving non-clinical and clinical time series [8, 10]. Let $w_{c,k}$ denote the weight connecting the k^{th} filter in the last convolutional block with the neuron corresponding to class c in the output layer, and $z_{k,t}$ denote the activation at time step t produced by the k^{th} filter in the last convolutional block of the FCN. Then, the input to the final softmax function in the output layer can be expressed as:

$$g_c = \sum_{k=1}^{K} w_{c,k} \sum_{t=1}^{T} z_{k,t} = \sum_{t=1}^{T} r_{c,t}, \tag{1}$$

where K is the number of filters in the last convolutional block, T is the length of the time series, and $r_{c,t}$ denotes the relevance score of time step t for class c:

$$r_{c,t} = \sum_{k=1}^{K} w_{c,k} z_{k,t}. \tag{2}$$

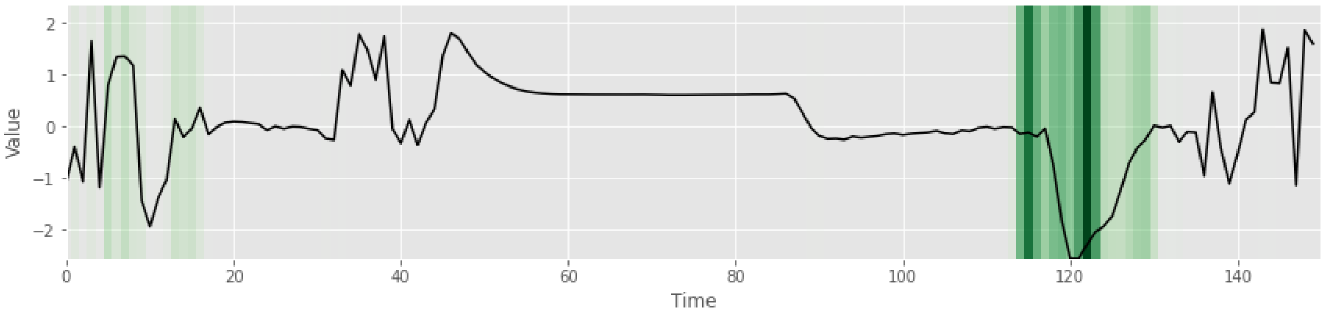


Fig. 2. Illustration of CAM technique for CNN interpretablity. Green indicates positive relevance to the prediction. The figure displays an example of a time series that is characterized by a downward dip in the initial of final periods of the time series. This example is correctly classified by the FCN, and the CAM highlights the time steps associated with the downward dip towards the end of the time series as the most relevant features for the prediction.

Equation 1 and 2 show that the relevance score at a given time step can be directly related to the input of the softmax function in the output layer, i.e. what produces the prediction of the model. Equation 2 also illustrates why the zero-padding in the FCN makes the architecture particularly suited for CAM. Since the output of the last convolutional block has the same length as the input, a relevance score for each time step can be computed. The CAM method can be understood as a weighted linear sum of the presence of different patterns in the input data, which can be used to identify the input regions most relevant to the particular category [9]. While it is possible to also use CAM with CNNs that reduce the length of the time series during processing through different upsampling procedures, it does complicate the procedure significantly. Furthermore, many interpretability techniques only consider the features that have a positive relevance for a given prediction, for instance the guided backpropagation method [17]. This is to provide clearer and unambiguous explanation for a prediction. While negative relevance scores can in some cases provide additional information, they can also be difficult to interpret. In binary classification problems, such as those considered in this work, the positive class typically makes up a homogeneous group with some shared defining characteristic (e.g. an elevate blood measurement). In contrast, the negative class (e.g. control patients) can be very different and typically makes up a highly heterogeneous group that can be difficult to interpret. Removing all negative relevance scores can be achieved by modifying Equation 2 as follows:

$$r_{c,t} = \max\Big(0, \sum_{k=1}^{K} w_{c,k} z_{k,t}\Big). \tag{3}$$

To end this section, an example illustrating the CAM approach is presented. The CAM interpretability technique is illustrated by training a FCN on the UMD dataset [18]. This dataset has three classes, one characterized by a downward dip in the initial or final period of the time series, one characterized by an upward dip in the initial or final period of the time series, and one characterized by no dip. Figure 2 shows an example that belongs to the first class, because of the significant dip towards the end of the time series, and is correctly classified by the FCN. Figure 2 displays the relevance scores produced for the first class by the CAM, where green indicates that a time step is relevant for classifying the sample to the first class. Figure 2 clearly indicates that the FCN is focusing on the downward dip towards the end of the time series, which is the defining characteristic of this class.

V. Ensembles for Uncertainty Estimation in Explainability

In this work, we propose an ensemble of FCNs for classification of EHRs. An ensemble is comprised of a set of separately trained classifiers which are combined to provide predictions when presented with new data. Ensemble approaches have been widely used in machine learning [19, 20] as they offer a number of advantageous properties. Ensembles are typically more accurate, consistent and have lower variance than their single model counterparts [19, 20]. Furthermore, they enable estimation of uncertainty in predictions and are simple and applicable for most tasks [19, 20]. The limitation of ensembles approaches are usually computational, but in the context of EHRs classification data is often limited and smaller models can usually be utilized [21]. Previous work on deep ensembles have demonstrated how they can increase accuracy and provide reliable uncertainty estimates [22]. Furthermore, as the amount of data can be limited for medical tasks [21], the training procedure can be unstable. However, an ensemble of classifiers are known to tackle such issues well [23]. The uncertainty measures are obtained by computing the standard deviation across the relevance scores for each model in the ensemble.

To calculate the ensemble mean and standard deviation of the relevance scores, let each CNN be parametrized by a set of parameters $\{\theta_1, \cdots, \theta_M\}$, where M is the number of models in the ensemble. Then, when considering the positive relevance scores from Equation 3, the mean relevance across the ensemble is defined as:

$$\boldsymbol{\mu}_{\mathbf{r}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{r}_{\theta_m}, \tag{4}$$

and the standard deviation across the ensemble is defined as:

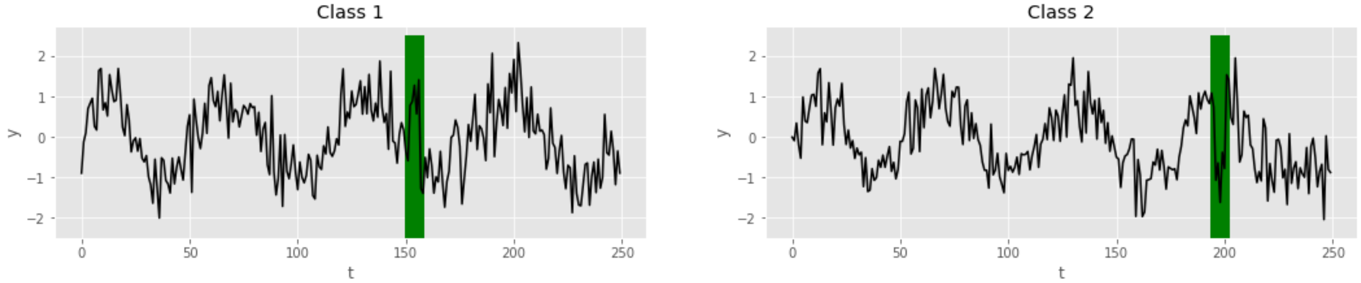


Fig. 3. Illustration of the synthetic data constructed for quantitative analysis of the proposed ensemble method. The leftmost plot shows an example from class 1, where an upwards spike is the defining characteristic. The rightmost plot shows an example from class 2, where a downward spike is the defining characteristic. The characteristic spikes are highlighted in the plots.

$$\boldsymbol{\sigma}_{\mathbf{r}} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\mathbf{r}_{\theta_m} - \boldsymbol{\mu}_{\mathbf{r}})^2}. \tag{5}$$

In Equation 4 and 5, $\mathbf{r}_{\theta_m}$ refers to the relevance scores for a given sample $\mathbf{x}$ provided by the model parametrized by the parameter set θ_m. Note that relevance scores should be scaled to the same range for them to be comparable. In all experiments we scale the relevance scores between 0 and 1 using the following equation:

$$\tilde{\mathbf{r}}_c = \frac{\mathbf{r}_c - \min(\mathbf{r}_c)}{\max(\mathbf{r}_c) - \min(\mathbf{r}_c)}. \tag{6}$$

A. Uncertainty Filtered Relevance Scores

For challenging datasets, several models in the ensemble might disagree on which time steps are most relevant for a prediction. In such cases, it might be desirable to only consider the time steps that most models in the ensemble agrees on, i.e. to filter out the uncertain relevance scores. With that in mind, a modification of the relevance scores is proposed, which will be referred to as uncertainty filtered relevance scores. For a given sample, the uncertainty filtered relevance scores are calculated by considering the standard deviation for each time step across all models in the ensemble. If the standard deviation is below some threshold the relevance scores are kept as it is. If the standard deviation is above some threshold the relevance scores are set to zero. This can be formulated as:

$$\tilde{\mu}_{r_t} = \begin{cases} \mu_{r_t} & \text{, if } \sigma_{r_t} < \epsilon \\ 0 & \text{, else} \end{cases}. \tag{7}$$

The threshold, ϵ, can be picked to fit the particular data in question. If only time steps with high certainty should be considered, a low threshold can be chosen. Or if uncertainty is not a concern, a high threshold can be selected. For this work, a simple but intuitive heuristic is proposed for setting the value of ϵ. Let the threshold be set to the mean of the standard deviation across all time steps of a given samples, that is:

$$\epsilon = \frac{1}{T}\sum_{t=1}^{T}\sigma_{r_t}. \tag{8}$$

This approach will ensure that the most uncertain relevance scores will be filtered out and will adapt the threshold to each specific sample.

VI. Synthetic Data for Quantitative Assessment of Explainability

A challenging aspect of XAI is that it is inherently qualitative, which makes quantitative assessment difficult. This is because for real datasets it is rarely known exactly what time steps are important. Therefore, to evaluate the proposed methodology, a synthetic dataset is constructed in such a way that the relevant time steps are known in advance. A time series classification task with two classes is constructed using the Python TimeSynth package[1], following the example of Tonekaboni *et al.* (2020) [13]. The data is constructed to resemble the periodic nature of ECG measurements in the presence of noise. Class number one is characterised by an upward spike, and class number two is characterised by a downward spike. The spike spans five time step, which are labeled as relevant time steps for the sample. Additionally, the two time steps preceding and succeeding the spike are also labeled as relevant time steps. These time steps are also chosen to be relevant as the change from no-spike to spike and vice versa is also important for the model to pick up. In total, there are 9 relevant points in each time series. Each time series consists of 250 time steps sampled from a univariate sinusoidal wave with a frequency of 0.2. Gaussian noise with zero mean and a variance of 0.5 is added at each time step. An example of the data and each class can be seen in Figure 3.

a) Calculating relevance accuracy: To calculate how accurate a model is a locating relevant time steps, we compare the k most relevant time steps for a prediction with the known most relevant times steps. For a given sample i, let $Y_i = \{y_1^{(i)}, \cdots, y_{N_r}^{(i)}\}$ denote the set of relevant points, where N_r is the number of known relevant time steps in the time series, and $R_i(k)$ denote the k most relevant time steps for the prediction of the model. The order of the relevance scores is not taken into account here, as all the relevant points are considered equally important in this case. Perfect relevance accuracy is achieved when all elements of Y_i are contained in $R_i(k)$, preferably with as few ks as possible. For a given

[1] https://github.com/TimeSynth/TimeSynth

sample i, Relevance accuracy can be calculated by dividing the cardinality of the intersection of $R_i(k)$ and Y_i by the cardinality of Y_i, which can be expressed as:

$$\text{Relevance accuracy} = \frac{|R_i(k) \cap Y_i|}{|Y_i|}, \tag{9}$$

where $\cap$ is the intersection of two sets and $|\cdot|$ is the cardinality of a set. Note that the relevant points are ranked from least to most relevant before they are compared to Y_i. Therefore, simply highlighting all time steps as relevant will not result in a high relevance accuracy score, the model needs to highlight some time steps as being more important than the others.

b) Calculating relevance consistency: For a model to be trustworthy, it should indicate mostly the same time steps as relevant for its prediction when trained from a different initialization. Models that highlight the same time steps as relevant when trained from different initialization will be referred to as consistent. Relevance consistency can be calculated in a similar way as relevance accuracy was computed. For a given sample, let $R_i^{(m)}(k)$ and $R_i^{(n)}(k)$ denote the k most relevant time steps for the prediction of two models trained from different initialization. Relevance consistency is computed by counting the number of shared elements of the two sets. As with relevance accuracy, the order is not taken into account. This computation can be formulated as, for a given sample, to compute the cardinality of the intersection of the set of the k most relevant time steps for the prediction of two models for then to divide by k. For a given sample i, computing the relevance consistency across M models can then be achieved by:

$$\text{Relevance consistency} = \frac{1}{M} \sum_{m,n=1}^{M} \frac{|R_i^{(m)}(k) \cap R_i^{(n)}(k)|}{k}. \tag{10}$$

Similarly as for Equation 9, the ranking of the relevance scores is important. To achieve high relevance consistency, the model must highlight the same time steps as more relevant than other time steps, even when trained from a different random initialization.

VII. Experiments and Discussion

Several experiments are conducted that demonstrate the benefits of the proposed methodology for creating explainable support systems based on DNNs. First, the proposed approaches are validated on synthetic data. Next, the relevance consistency of the ensemble approach is validated on both synthetic and clinical data. Further, the proposed approaches are used to determine what inputs are important for classifying ECGs as a normal heartbeat or a myocardial infarction (heart attack). A similar experiment is also conducted for identifying patients with surgical site infection in blood measurements of CRP. Table I provides an overview of the different properties of all datasets used to evaluate the proposed methodology. The table shows, for each dataset, the number of samples, the class distribution, the length of the time series, what each time step represents and what type of measurement that is considered. The FCN described in Section III is used for all tasks, and is trained using a cross-entropy loss and the Adam optimizer [24]. For all experiments in this section, all ensembles are composed of 10 FCNs. To evaluate the performance of the classifiers we compute four metrics on the test data of each dataset; precision, recall, negative predictive value (NPV), and specificity. These metrics are chosen to reflect typical challenges when evaluating performance in classification of clinical time series, such as unbalanced data and false positives. The code used in this manuscript is available at https://github.com/Wickstrom/TimeSeriesXAI.

TABLE I
Description of the three datasets used in experiments, including the number of training and test samples (N_{TR}, N_{TE}), the length of the time series (T), what each time step represents, and the type of data (Data). The table also displays the number of samples for each class (C=0 and C=1) in both the training and test data.

	Synthetic	**ECG200**	**SSI**
N_{tr} (C=0)	250	69	520
N_{tr} (C=1)	250	31	185
N_{te} (C=0)	250	64	130
N_{te} (C=1)	250	36	48
T	250	96	20
TS	·	microseconds	days
Data	·	heartbeat	C-reactive protein

A. Validation on Synthetic Data

Following the procedure described in Section VI, a training set of 500 samples is generated, along with a separate test set of 500 samples. The performance of a single FCN and an ensemble of FCNs on the synthetic test data, each trained for 150 epochs, is displayed in Table II. A Monte Carlo permutation test with 10000 permutations is conducted to test for significance. Results indicate that the ensemble has higher precision and produce less false positives.

a) Relevant time steps accuracy: The relevance accuracy of the single and ensemble model for different values of k are presented in Table III, where the results are averaged over 10 independent training runs. As described in Section VI, the data is constructed in such a way that there are 9 time steps that are considered relevant, inspired by Tonekaboni *et al.* (2020) [13]. Therefore, we start by evaluating the relevance accuracy at $k = 9$ and above. Table III shows that the FCN is capable of identifying the relevant samples in time series with high relevance accuracy. Furthermore, the results show that the deep ensemble is more accurate at identifying relevant samples compared to single models, and also has much less variability in their prediction. A Monte Carlo permutation test with 10000 permutations is conducted to test for significance, and the difference between the single and the ensemble model is significant for most ks at a significance level of 0.01.

b) Most highlighted relevant time steps: A priori, it is not obvious which time steps the model will highlight most frequently as being relevant for predictions. However, it is desirable that the model use the known relevant time steps for its prediction. To evaluate which time steps are most frequently

TABLE II
EVALUATION OF CLASSIFICATION PERFORMANCE OF SINGLE AND ENSEMBLE MODEL ON THE TEST DATA OF DIFFERENT DATASET. BOLD NUMBER INDICATE STATISTICAL SIGNIFICANCE AT A SIGNIFICANCE LEVEL OF 0.01.

Dataset	Precision		Recall		NPV		Specificity	
	Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble
Synthetic	.983±.011	**.991±.001**	.968±.015	.962±.002	.968±.014	.963±.002	.983±.028	**.992±.001**
ECG200	.819±.018	.814±.023	.741±.044	**.755±.035**	.862±.020	.867±.016	.908±.010	.903±.014
SSI	.947±.069	**.978±.015**	.922±.037	.936±.041	.973±.013	.976±.013	.981±.029	.993±.006

used to make a prediction, we define a relevance ratio (RR). For a given sample i and a known relevant time step $y_j^{(i)}$, the RR is expressed as:

$$RR_j = \frac{1}{N_{te}} \sum_{i=1}^{N_{te}} I(y_j^{(i)}, R_i(k)), \tag{11}$$

where I is an indicator function that evaluates to 1 if the known relevant sample $y_j^{(i)}$ is among the k most relevant points $R_i(k)$ and N_{te} is the number of test samples. If $RR_j = 1$, then the known relevant time step $y_j^{(i)}$ is included in $R_i(k)$ for all samples. It is expected that a random time step will be included $N_{te}(k/T)$ number of times among the most relevant time steps, which can be considered a lower bound for how many times a time step should be included.

Figure 4 displays which time steps are most frequently highlighted as relevant for different values of k. In Figure 4, the two initial and two final bars correspond to the two initial and two final relevant time steps, while the five bars in the middle correspond to the spike in the synthetic data. Results indicate that the five time steps corresponding to the spike are most frequently highlighted as being relevant for the prediction. The figure also shows that the ensemble model uses the known relevant time steps much more frequently than other non-relevant time steps.

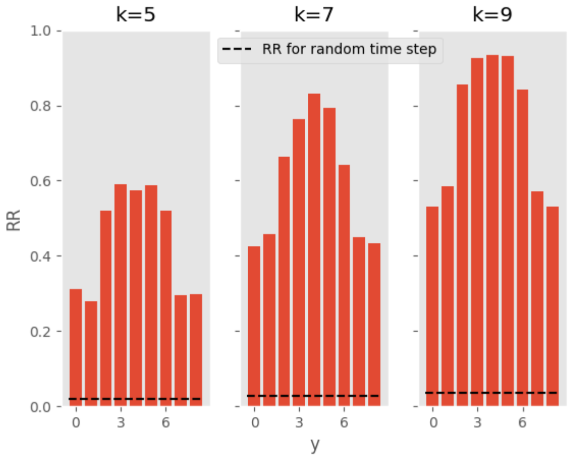


Fig. 4. Figure shows the number of times a known relevant time step is included in the top k most relevant time steps across all test samples in the synthetic dataset. Results indicate that the time steps at the middle of the spike is the most frequently highlighted as relevant for a prediction, and that the known relevant time steps are more frequently highlighted as being important compared to non-relevant time steps.

TABLE III
RELEVANCE ACCURACY OF FCN FOR RELEVANCE SCORES ON SYNTHETIC DATA. SCORES ARE AVERAGED OVER 10 RUNS FOR BOTH THE SINGLE AND ENSEMBLE MODEL. BOLD NUMBER INDICATE STATISTICAL SIGNIFICANCE AT A SIGNIFICANCE LEVEL OF 0.01.

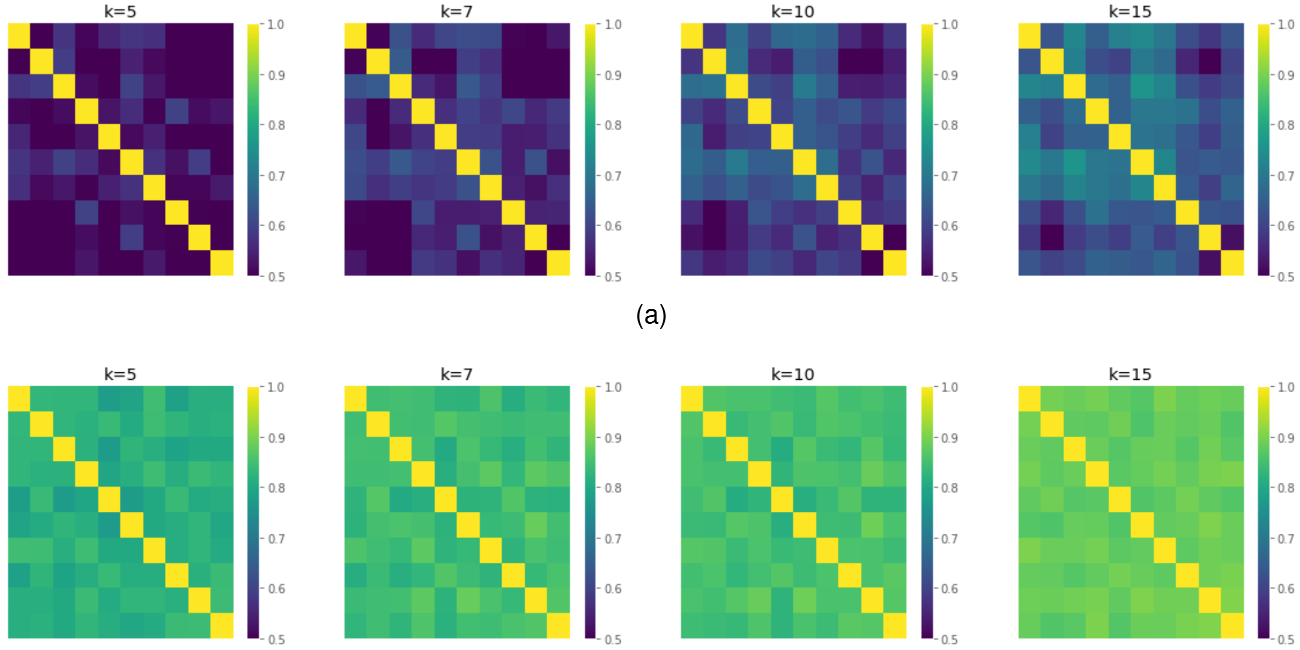
Top k	Single	Ensemble
k=9	.703 ± 0.014	**.736 ± .014**
k=10	.765 ± 0.013	**.800 ± .011**
k=11	.821 ± 0.012	**.855 ± .011**
k=12	.868 ± 0.010	**.899 ± .007**
k=15	.935 ± 0.007	**.950 ± .001**

B. Relevance consistency in Relevance Scores

The relevance consistency of single and ensemble models is evaluated on both the synthetic data described in the previous section and on the ECG200 dataset [18, 25]. The ECG200 dataset consists of ECGs that traces the electrical activity recorded during a single heartbeat. It is obtained from the UCR time series classification archive [18] and has a predefined training and test split. The task is to discriminate between normal heartbeats and those associated with myocardial infarction, also known as a heart attack. Table IV shows the relevance consistency of 10 single and 10 ensemble models on both datasets. Results demonstrate that the ensemble approach is far more consistent than the single models, and has much lower variability in its scores. This is particularly prominent for the more challenging ECG200 data, where the single models have difficulties with agreeing on the what time steps are relevant compared to the ensemble approach. A Monte Carlo permutation test with 10000 permutations is conducted to test for significance in both datasets, and the difference between the single and the ensemble model is significant for all ks at a significance level of 0.01. Furthermore, Figure 5 shows the relevance consistency between the 10 single models and 10 ensemble models on the ECG200 dataset. The figure corroborate the quantitative results in Table IV that show how the ensemble approach is more consistent than single models.

TABLE IV
RELEVANCE CONSISTENCY OF RELEVANCE SCORES AVERAGED OVER 10 SINGLE AND 10 ENSEMBLE MODELS ON SYNTHETIC DATA AND ECG TIME SERIES. BOLD NUMBER INDICATE STATISTICAL SIGNIFICANCE AT A SIGNIFICANCE LEVEL OF 0.01.

Top k	Synthetic		ECG200	
	Single	**Ensemble**	**Single**	**Ensemble**
k=5	.80 ± .07	**.93 ± 0.02**	.57 ± .15	**.82 ± 0.07**
k=7	.85 ± .05	**.94 ± 0.02**	.60 ± .15	**.84 ± 0.07**
k=10	.89 ± .04	**.95 ± 0.01**	.63 ± .14	**.85 ± 0.06**
k=15	.91 ± .04	**.97 ± 0.01**	.67 ± .13	**.88 ± 0.05**



(a)

(b)

Fig. 5. Relevance consistency across 10 single and ensemble models for different number of k relevant time steps on the ECG200 test data. Top row shows relevance consistency of single models and bottom row shows relevance consistency in ensemble models. Figure displays how ensemble models regularly highlight the same time steps as relevant for their prediction, while the single models have much more variability in what time steps are being indicated as relevant for their prediction.

C. Myocardial Infarction Detection

The proposed approach for measuring uncertainty in the relevance scores is validated on the ECG200 data described above. The performance of a single FCN and an ensemble of FCNs on the test data of the ECGO200 dataset, each trained for 150 epochs, is displayed in Table II. A Monte Carlo permutation test with 10000 permutations is conducted to test for significance. Results show that the ensemble is more capable of identifying positive samples. Figure 6a and 6b show the mean and standard deviation of the relevance scores across all models in the ensemble for a myocardial infarction case. This sample was correctly classified by the ensemble as belonging to the myocardial infarction class. Figure 6a indicates that there are three regions of electrical activity that influenced the prediction of the model. First, a steep incline in the initial ST-period, second a slight decline after a peak, and lastly, a peak towards the end. However, Figure 6b shows that there are several regions where the models in the ensemble disagrees on the relevance of different time steps. By using the proposed method of uncertainty filtered relevance scores, Figure 6c is created. Here, the uncertain relevance scores are removed and only the certain scores remain. Now, Figure 6c shows that there is only one region of highly certain and relevant time steps, which is the initial incline in electrical activity in the ST-period. This rapid change is electrical activity is also associated with the myocardial infarction class [25], which suggests that the ensemble is able to capture clinically relevant features in the input data.

Lastly, we present an example where the ensemble classifies a sample correctly as a heart attack while the single model makes an error. Figure 7 displays the relevance scores for both the single and ensemble model for this particular example. While the relevance scores have similarities, notice that the single model emphasis the importance of some initial time steps, while the ensemble model indicates that the decline after the ST-period is the most relevant part of the time series. The single model also indicate this part of the time series as important, but less important than the initial part of the time series. This shows how the ensemble is capable of filtering out time steps that might not be important and focus on clinically relevant features, which in this case leads to a correct classification.

D. Surgical Site Infection Detection

The next task the proposed methodology is validated on is surgical site infection in measurements of CRP. The dataset consists of 883 patients that have undergone a gastrointestinal surgical procedure at the Department of Gastrointestinal Surgery at the University Hospital of North Norway in the years 2004 - 2012 [26]. Of the 883 patients, 232 are infected while the rest are control patients. 80 % of the data were used for training and 20 % were used as an independent test set. This split is conducted at each independent training run to obtain a cross-validated evaluation of performance. The performance of a single FCN and an ensemble of FCNs on the test data of the SSI dataset, each trained for 150 epochs, is displayed in Table II. A Monte Carlo permutation test with 10000 permutations is conducted to test for significance. Results display that the ensemble has higher precision compared to single models.

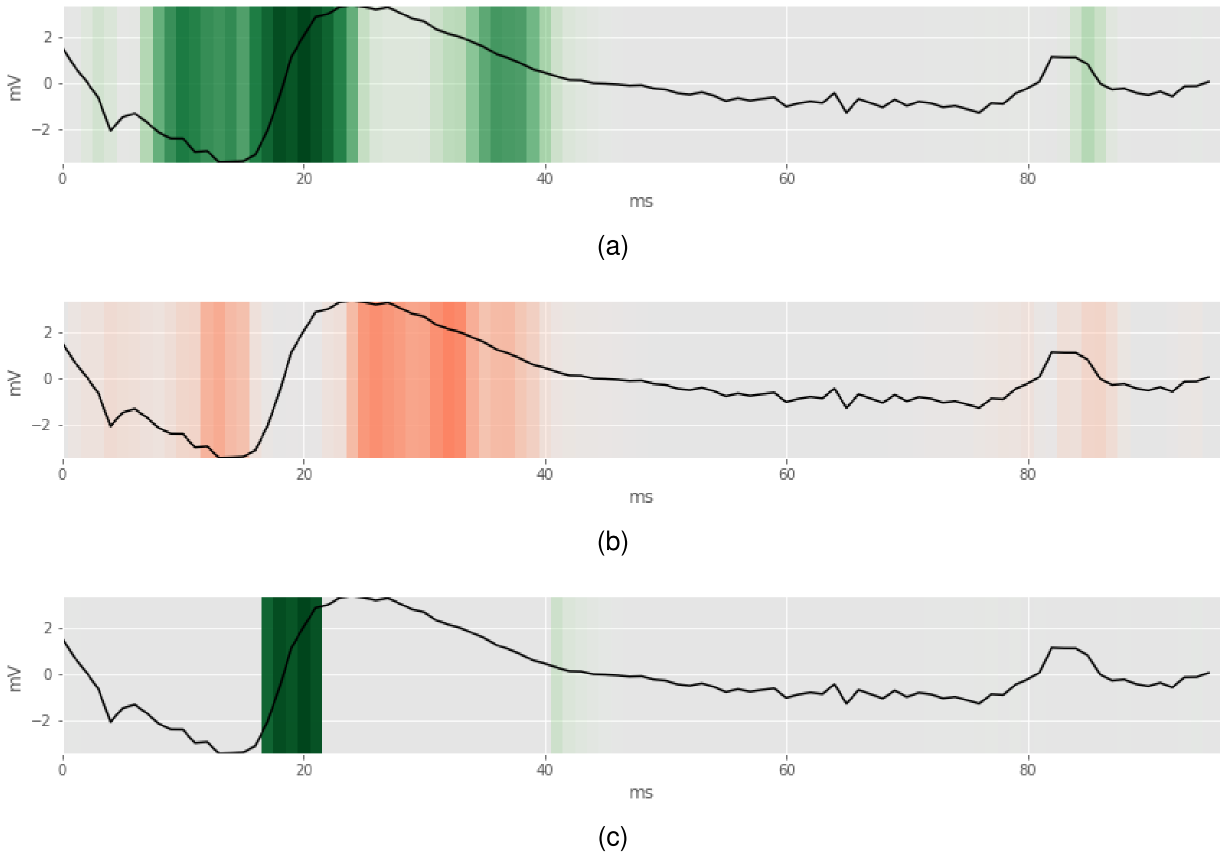


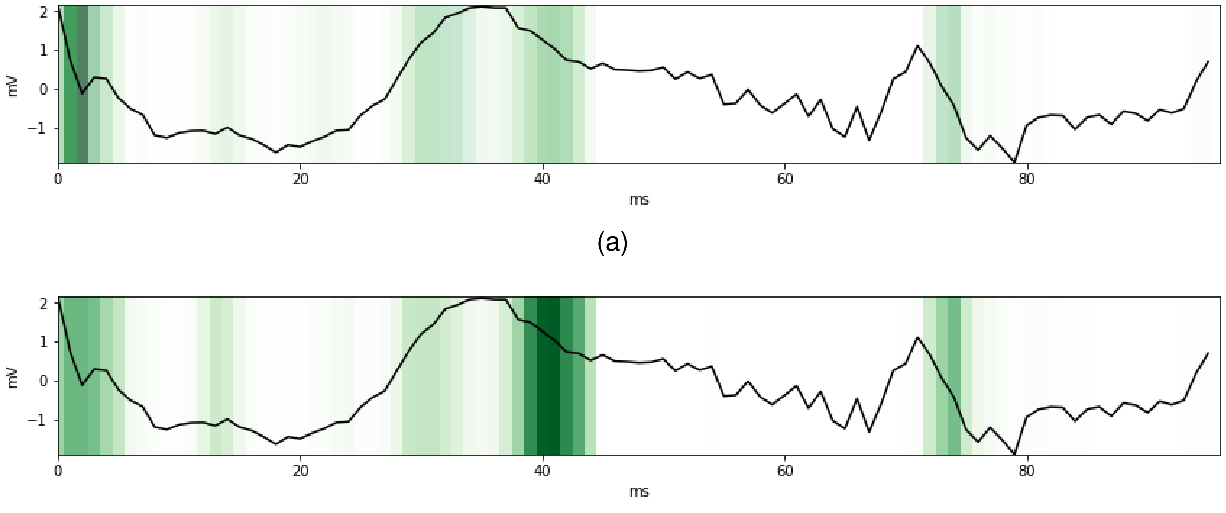
Fig. 6. Example from the myocardial infarction class that was correctly classified by the deep ensemble. From top to bottom: mean relevance scores across all models in ensemble (a), standard deviation across all models in ensemble (b), and uncertainty filtered relevance scores obtained using the proposed method in Section V-A (c). (a) shows that there are several regions of relevant time steps, but (b) indicates that there is a degree of uncertainty associated with several of those regions. (c) shows only the certainly relevant samples, where the uncertain time steps are filtered out using the proposed methodology.

Figure 9 displays an example of a patient that contracted an infection and was correctly classified by the deep ensemble. Figure 9a and 9b show the mean and standard deviation of the relevance scores across all models in the ensemble. Figure 9a indicates that there are several time step deemed relevant by the ensemble, and particularly the rise in CRP is indicated as highly relevant for the prediction. However, Figure 9b shows that a number of these these time steps have a high degree of uncertainty associated with them. Particularly, the relevance of the central plateau and final parts of the CRP measurement is something that the models in the ensemble disagree on. Figure 9c displays the standard deviation filtered mean relevance scores, which indicates that only the incline around day 13 is the only certainly relevant part for the prediction of infection for this patient.

The typical development of CRP for a patient that has undergone surgery but does not contract an infection is an initial postoperative increase followed by a steady decline. For patients that do get an infection, CRP typically sees an increase again some days postoperatively after the initial decline. The correlation between CRP and the risk of infection has been noted in previous studies [27, 28]. Figure 8 shows the median CRP at each time steps for each class of all samples in the training dataset. This illustrates how CRP is typically higher for infected patients, and they tend to have an increase in CRP after the initial decline after surgery. Figure 9c shows the incline at 12-13 days after surgery are indicated as the certainly relevant time steps for the prediction of the surgical site infection class. As described in this paragraph, such a pattern is closely connected with a possibility of developing an infection, which suggests that the deep ensemble uses clinically relevant features to make its prediction.

VIII. Conclusion

In this work a deep ensemble approach for explainable CNNs was proposed. The proposed method was evaluated on both synthetic and real world data. Results demonstrate that deep ensembles are capable of finding relevant features in clinical time series and that by modeling the uncertainty in relevance scores more understandable and trustworthy explanations can be provided. A novel thresholding approach was proposed and demonstrated. While only one thresholding was investigated in this work, we believe that different thresholding strategies could be applicable, which is an interesting line of research for future works. The contributions of this work can enable the construction of explainable decision support systems that are more trustworthy and more accurate than previous systems based on deep learning.



(a)

(b)

Fig. 7. Example for the heart attack class where a single model fails to detect the heart attack while the ensemble correctly classifies the patient. The figure shows the relevance scores for the single model (a) and the ensemble model (b) for the prediction of the heart attack class. While there are similarities between the relevance scores, the single model puts more emphasis on some initial time steps while the ensemble focuses on time steps related to the ST-period.

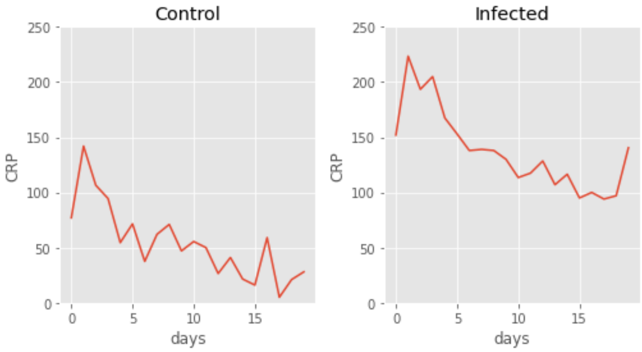


Fig. 8. Median CRP at each time steps for each class of all samples in the training dataset. Figure shows how control patients usually have an increase in CRP after surgery but declines to a low value towards the end of the time series. The infected patients typically have a higher value of CRP and also tend to have an increase in CRP towards the end of the time series.

Acknowledgment

The authors would like to thank Jonas Nordhaug Myhre, Luigi Tomasso Luppino, and Ahcene Boubekki for helpful discussions during both the conception and development of this work.

References

[1] T. Khaleghi, A. Murat, S. Arslanturk, and E. Davies, "Automated surgical term clustering: A text mining approach for unstructured textual surgery descriptions," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2019.

[2] H. Harutyunyan, H. Khachatrian, D. Kale, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, 03 2017.

[3] A. S. Strauman, F. M. Bianchi, K. Ø. Mikalsen, M. Kampffmeyer, C. Soguero-Ruiz, and R. Jenssen, "Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks," in *2018 IEEE EMBS BHI*, 2018.

[4] H. Duan, Z. Sun, W. Dong, K. He, and Z. Huang, "On clinical event prediction in patient treatment trajectory using longitudinal electronic health records," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2019.

[5] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *ArXiv*, vol. abs/1610.01644, 2017.

[6] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *MLHC*, 2019, pp. 359–380.

[7] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2019.

[8] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 IJCNN*, 2017, pp. 1578–1585.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE CVPR*, 2016.

[10] R. Assaf and A. Schumann, "Explainable deep neural networks for multivariate time series predictions," in *IJCAI*, 2019, pp. 6488–6490.

[11] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes, "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record," *IEEE Access*, vol. 6, pp. 65 333–65 346, 2018.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, Y. Bengio and Y. LeCun, Eds., 2015.

[13] S. Tonekaboni, S. Joshi, D. Duvenaud, and

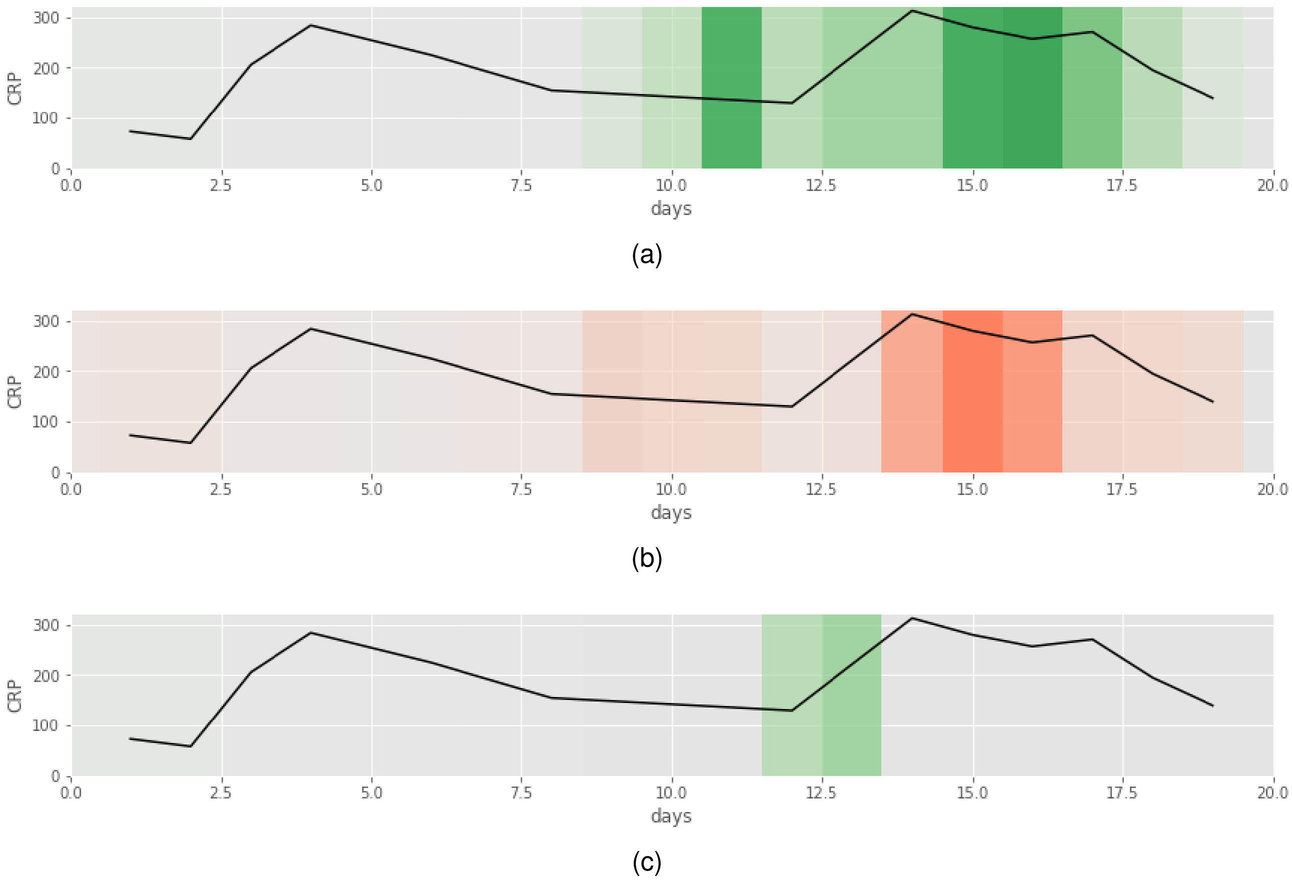


Fig. 9. Example from the infection class that was correctly classified by the deep ensemble. From top to bottom: mean relevance scores across all models in ensemble (a), standard deviation across all models in ensemble (b), and uncertainty filtered relevance scores obtained using the proposed method in Section V-A (c). (a) shows that there are several time steps highlighted as important for the prediction, but (b) shows that there is a degree of uncertainty associated with several of them. (c) shows that the only certainly relevant time steps are those associated with the CRP incline around day 12-13, a pattern known to correlated with the risk of contracting an infection.

A. Goldenberg, "Explaining time series by counterfactuals," 2020. [Online]. Available: https://openreview.net/forum?id=HygDF1rYDB

[14] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Medical Image Analysis*, vol. 60, p. 101619, 2020.

[15] I. Fawaz *et al.*, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, pp. 1936–1962, 2020.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[17] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[18] H. Dau *et al.*, "The ucr time series archive," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. JAS-2019-0114, p. 1293, 2019.

[19] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artif. Int. Res.*, vol. 11, no. 1, p. 169–198, Jul. 1999.

[20] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, pp. 1–39, 02 2010.

[21] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeuRIPS*, 2017, p. 6405–6416.

[23] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, p. 123–140, Aug. 1996.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd ICLR*, 2015.

[25] R. Olszewski, "Generalized feature extraction for structural pattern recognition in time-series data," Ph.D. dissertation, Carnegie Mellon University, 2001.

[26] Mikalsen *et al.*, "Learning similarities between irregularly sampled short multivariate time series from ehrs," in *ICPR*, 2016, p. 1–6.

[27] P. Singh, I. Zeng, S. Srinivasa, D. Lemanu, A. Connolly, and A. Hill, "Systematic review and meta-analysis of use of serum c-reactive protein levels to predict anastomotic leak after colorectal surgery," *The British journal of surgery*, vol. 101, pp. 339–346, 03 2014.

[28] Soguero-Ruiz *et al.*, "Data-driven temporal prediction of surgical site infection," *AMIA Symposium*, p. 1164—1173, 2015.