Utilizing Temporal Patterns for Estimating Uncertainty in Interpretable Early Decision Making

Mohamed F. Ghalwash[*]
Temple University
Philadelphia, PA
mohamed@temple.edu

Vladan Radosavljevic[†]
Yahoo Labs
Sunnyvale, CA
vladan@yahoo-inc.com

Zoran Obradovic
Temple University
Philadelphia, PA
zoran.obradovic@temple.edu

ABSTRACT

Early classification of time series is prevalent in many time-sensitive applications such as, but not limited to, early warning of disease outcome and early warning of crisis in stock market. For example, early diagnosis allows physicians to design appropriate therapeutic strategies at early stages of diseases. However, practical adaptation of early classification of time series requires an easy to understand explanation (interpretability) and a measure of confidence of the prediction results (uncertainty estimates). These two aspects were not jointly addressed in previous time series early classification studies, such that a difficult choice of selecting one of these aspects is required. In this study, we propose a simple and yet effective method to provide uncertainty estimates for an interpretable early classification method. The question we address here is "*how to provide estimates of uncertainty in regard to interpretable early prediction*." In our extensive evaluation on twenty time series datasets we showed that the proposed method has several advantages over the state-of-the-art method that provides reliability estimates in early classification. Namely, the proposed method is more *effective* than the state-of-the-art method, is *simple* to implement, and provides *interpretable* results.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

Keywords

Interpretability, earliness, uncertainty, time series, reliability

1. INTRODUCTION

Time series early classification models aim to predict the label of the entire time series by observing the phenomenon for very short time and as soon as enough data is available. If the available data is insufficient to make an accurate classification, more data, which might be expensive, is required.

The framework of time series early classification is illustrated in Figure 1. The model looks into a portion of l observations of the time series T of length L (where $l < L$) and determines the label of the entire time series without observing the rest of the time series. If the method can not classify the time series at time l, the observation segment is enlarged and the process is repeated with the aim of predicting the class label of the entire time series.

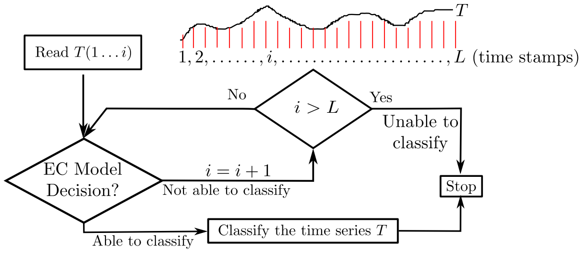


Figure 1: Framework of Early Classification (EC) of time series.

Early classification of time series is important in many applications. For example, early diagnosis could save patients' lives by allowing administration of treatment before the diseases are fully manifested [6, 3]. In such applications it is highly desirable to have easily interpretable results, as physicians aim to understand why a prediction is made. Moreover, in order to decide whether the available data is sufficient for the model to make an accurate prediction, uncertainty estimates of the predictions would be provided by the model. For example, in medical applications, providing uncertainty estimates would assist physicians in optimizing therapy. We illustrate the importance of providing uncertainty estimates using the following example.

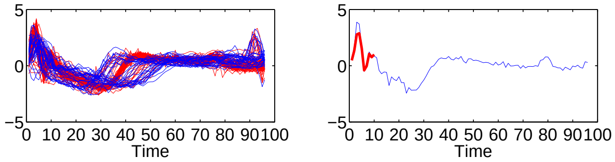
Example 1. Time series from the red and blue classes from a medical dataset (ECG dataset) are shown in Figure 2a. The time series are very similar to each other such that it is hard to distinguish between the classes using human eyes. Using only the principle of early classification without the notion of uncertainty, the blue time series in Figure 2b could be incorrectly classified as the red class at time point 12 (this is done using the interpretable early classification method described later in Section 2). Since the method does not provide uncertainty estimates, there is no clue indicating how accurate the classification is. However, if the method provides high uncertainty estimates with that classification decision, then additional data would be required in order to provide a more confident decision.

[*] Ain Shams University, Cairo, Egypt.

[†] This study was conducted while the author was a postdoctoral associate at Prof. Obradovic's laboratory at Temple University.

KDD'14, August 24-27, 2014, New York, New York, USA.

ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623694.



(a) (b)

Figure 2: (a) Time series from the blue and red classes. (b) A blue time series is early incorrectly classified as a member of the red class. Additional time and data is needed to provide a more confident decision.

Table 1: Three properties are used to categorize time series classification methods: interpretability (I), earliness (E), and uncertainty estimate (U).

	LDS/SVM [1]	HMM [10]	Shapelet [16]	ECTS [14]	EDSC [15]	PR [12]	QDA [11]
I	X	X	✓	X	✓	X	X
E	X	X	X	✓	✓	✓	✓
U	X	X	X	X	X	✓	✓

Therefore, the early classification method is needed to provide uncertainty estimates with the classification decision such that confident results are obtained as early as possible. Although earliness, interpretability, and uncertainty estimates are highly desirable properties in many time series classification applications, to the best of our knowledge, existing methods are limited to addressing at most two of these aspects. Table 1 summarizes properties of several methods (discussed later in Section 2) that successfully addressed some of these aspects. There is not yet a time series classification method that simultaneously provides these three aspects.

In this study, we propose a *simple but effective* method that satisfies all three properties. This is achieved by extending a recently proposed interpretable early classification method, called early distinctive shapelet classification (EDSC), to estimate the temporal uncertainty associated with the prediction.

In Section 2, existing methods for time series classification that address some of the three aspects shown in Table 1 are reviewed. The EDSC method is summarized in Section 3 followed by a description of our proposed temporal uncertainty estimation method provided in Section 4. Extensive evaluation results on benchmark datasets (twenty time series datasets) are presented in Section 5. Finally, the conclusions and future work is given in Section 6.

2. RELATED WORK

In this section we discuss the existing time series classification methods that address some aspects of earliness, interpretability, and uncertainty estimates.

2.1 Adapting time series classification methods for early classification

Several time series classification methods are *adapted* for early classification by applying the method at each time point. Such methods are inflexible, as for methods trained on time series of length t, the prediction is always done at the t^{th} time point, which limits the applicability of the model, e.g. in medical applications, patients may develop diseases at different times, which this type of methods can not handle appropriately.

Examples of methods in this category are [1] and [10]. In [1] a linear dynamical system and support vector machines are used to model time series for gene expression while in [10], a method that utilizes hidden Markov models with less states than the time points is proposed. These methods are evaluated at each time point and the best results were obtained when using the entire time series. These results provide evidence that there is a need for methods designed specifically for early classification.

Shapelets, which are subsequences of time series, have been used as features representing the characteristics of time series [16]. A perfect shapelet of a class is the one representing all time series in that class but not covering any time series in other classes. These shapelets are used for interpretable time series classification where they are able to handle challenges such as drifts and changes, which often occur in medical domain [8] and which are usually addressed by boosting algorithms [5]. Although the shapelet method provides interpretability, it provides neither uncertainty estimate nor early classification.

2.2 Early classification

The problem of early classification of time series is formulated recently [13, 14]. A novel concept of minimum prediction length is introduced in the early classification of time series (ECTS) method [13, 14]. ECTS makes early predictions and retains accuracy comparable with that of the 1-nearest neighbor classifier using the entire time series. However, ECTS does not extract patterns from the training data; thus, users may not be able to gain insights from the classification results.

The drawback of ECTS has been resolved in a method called early distinctive shapelet classification (EDSC) [15]. The EDSC method extracts local shapelets which distinctly manifest the target class locally, and are effective for early classification. However, the EDSC method does not provide uncertainty estimates.

2.3 Early classification with uncertainty estimate

A method is proposed to represent the patient risk (PR) as a time series and estimates the uncertainty as the distance between the evolving approximate daily risk of a patient and the hyperplane learned by SVM [12]. The PR model requires labels for each time point in the time series, not just a label for the entire time series, and therefore can not be compared to our method. The state-of-the-art method for early classification of signals using a quadratic discriminant analysis (QDA) classifier was developed recently [11]. QDA provides a reliability bound on the classifier's decision for every time point. The disadvantage of these methods is that they are not interpretable and can be used only as "black box" classifiers.

3. BACKGROUND: EDSC

In this section we briefly describe the early distinctive shapelet classification (EDSC) method for interpretable early classification and the details of EDSC are described in the subsequent subsections [1]. EDSC extracts discriminative shapelets for early classification. In this approach shapelets, which are subsequences of time series and thus are highly interpretable, have been used as features representing the characteristics of the time series [16].

Example 2. Suppose we have a dataset of time series for 3 unhealthy subjects (red) and 3 healthy subjects (blue) as illustrated in Figure 3. The extracted shapelet is the time series segment that represents the characteristics of the class (drawn as solid lines).

Given a time series dataset D where each time series is associated with a label, the task is to classify the time series as early as possible. The EDSC method addresses this problem in 4 steps:

[1] For more details about EDSC, the reader is referred to [15]

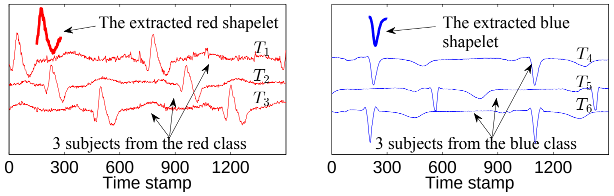


Figure 3: Three subjects from the red (blue) class are shown in the left (right) panel, respectively. The discriminative shapelet is represented as solid line. Clearly, the shapelet represents the characteristics of its class.

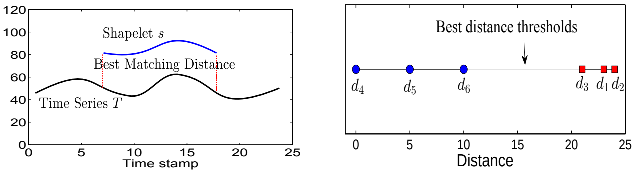
1. extracts all shapelets of different lengths, where for each shapelet a distance threshold is learned such that the shapelet discriminates between classes,
2. ranks the shapelets using a utility function that incorporates earliness and accuracy of the shapelet,
3. prunes the shapelets by selecting the top shapelets that cover the entire dataset,
4. classifies unknown time series based on the closest matching shapelet.

3.1 Extracting all shapelets

The shapelet is defined as $S = (s, l, \delta, c)$ where s is a time series subsequence of length l, c is the class label of the shapelet which is called the target class. The other classes ($\overline{c}$) are called the non-target classes. δ is a distance threshold which needs to be learned. To compute the distance threshold, we compute the distances between the subsequence s and all time series in the dataset.

To compute the distance between a subsequence s of length l and a time series T of length L (where $l < L$), we slide a window of length l over the time series T to extract all subsequences $\{h_1, h_2, \ldots h_{L-l+1}\}$ of length l (Figure 4a). Then, the distance is computed as

$$dist(s, T) = \min_{\forall i \in \{1,2,\ldots,L-l+1\}} dist(s, h_i). \qquad (1)$$



(a) (b)

Figure 4: (a) Best matching distance $dist(s,T)$ ***between the shapelet*** s ***and the time series*** T***. (b) The distance between the shapelet and each blue/red time series is represented as a blue/red point on the order line, respectively. The distance threshold*** δ ***is computed such that the shapelet discriminates between the classes.***

The distance between the shapelet S and each time series is represented as a point on the order line (Figure 4b). Then, the distance threshold δ is computed such that the shapelet discriminates between the classes[2].

Then, the EDSC method iterates over all time series in D to extract all subsequences of length l, where l is the length of the potential shapelets. The EDSC method varies l between $minL$ and $maxL$ which are user parameters.

3.2 Ranking the shapelets

The EDSC method extracts all possible shapelets and computes the distance threshold for each shapelet. The number of the extracted shapelets from the dataset is large. Therefore, to find the discriminative shapelets, EDSC assigns a score to each shapelet that incorporates both the earliness and the accuracy.

The earliness defines how early, on average, the shapelet matches the target time series (*the shapelet* S ***matches*** *the time series* T *if* $dist(s,T) \le \delta$). Technically, the earliness between the shapelet $S = (s, l, \delta, c)$ and the time series T of length L is defined as

$$EML(S, T) = \min_{\forall i \in \{1,2,\ldots,L-l+1\}} dist(s, h_i) \le \delta, \qquad (2)$$

where h_i are all subsequences of the time series T of length l. Then, using the earliness, the weighted recall of the shapelet is computed as

$$WRecall(S) = \frac{1}{\|T_{\overline{c}}\|} \sum_{T \in D} \frac{1}{\sqrt[\alpha]{EML(S,T)}}, \qquad (3)$$

where α is a user defined parameter that determines the importance of the earliness, and $\|T_{\overline{c}}\|$ is the number of non-target time series. Finally, the utility score of the shapelet is defined as

$$Utility(S) = \frac{2 \times Precision(S) \times WRecall(s)}{Precision(S) + WRecall(S)}, \qquad (4)$$

where $Precision$ is the fraction of the matched time series that are relevant (target time series) and is computed as

$$Precision(S) = \frac{\|\{d_i \le \delta \bigwedge Class(T_i) = c\}\|}{\|\{d_i \le \delta\}\|}, \qquad (5)$$

where $d_i = dist(s, T_i)$ and $Class(T_i)$ is the class of the i^{th} time series T_i.

3.3 Pruning the shapelets

The EDSC method sorts the shapelets descending based on their utility scores (Equation 4). It starts with the highest ranked shapelet and removes all time series from the dataset that are covered by the shapelet. The shapelet $S = (s, l, \delta, c)$ ***covers*** the time series T if the shapelet matches the time series ($dist(s,T) \le \delta$) and has the same class as the shapelet. Then, EDSC stores the shapelets and the next highest shapelet is considered. If the shapelet covers any of the remaining time series, the shapelet will be added to the list and all covered time series will be removed. The method iteratively does so until all time series in the dataset are covered. In this manner, the EDSC method ends up with a small list of shapelets that is used in the classification phase.

3.4 Classification phase

The EDSC method initially reads a portion of length $minL$ from the test time series (where $minL$ is the length of the shortest extracted shapelet). The highest-ranked shapelet is considered. If the shapelet matches the current segment of the time series then the time series is classified as the class of the shapelet and the prediction is made. Otherwise, the next shapelet from the ranked list is considered and the process of checking each shapelet is repeated. If none of the shapelets match the current portion of the test time series then the method reads one more time stamp and continues classifying the time series (Figure 1). If the method reaches the end of the time series and none of the shapelets match it, then EDSC marks the time series as a not-classified example.

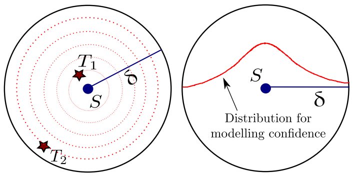
[2] The details of the distance threshold computation are explained in [15].

4. THE PROPOSED METHOD FOR UNCERTAINTY ESTIMATION

4.1 Motivating example

We start by proposing a method to provide an uncertainty estimate to the interpretable EDSC method. Assume that we have a shapelet $S = (s, l, \delta, c)$ and a time series T. If the distance $dist(s,T)$ between T and S is less than or equal to δ, then T is classified as class c. In this scenario, we did not measure how confident the classification is.

Imagine that the shapelet S is represented as a point as in Figure 5a. The radius of the black colored circle around the shapelet represents the shapelet's distance threshold δ. Assume that we have two time series T_1 and T_2. If the distance between S and each time series T_1 and T_2 is less than δ then T_1 and T_2 are represented as points inside the circle and both are then classified as the shapelet's class. However, the distance between T_1 and S is less than the distance between T_2 and S which reflects the fact that the shapelet is more certain about the classification of T_1 than the classification of T_2.



(a) (b)

Figure 5: (a) The shapelet is represented as a point and the radius of the circle around S represents the shapelet's distance threshold δ. T_1 and T_2 are two time series that are less than δ-distant apart from the shapelet. (b) The confidence of the shapelet: the shapelet is more certain about closer time series than time series that are δ-distant apart.

In Figure 5a, different red dotted circles represent different levels of confidence. Intuitively, the shapelet is more certain about the time series in the most inner red circle than the time series in the bigger red circles. Therefore, the uncertainty reaches the highest level when the time series lies on the boundary of the black circle.

4.2 Uncertainty estimation

Instead of modeling the uncertainty directly, we model the confidence $C(c)$ of classifying a time series as class c. The uncertainty $U(c)$ of classifying a time series as class c can be computed as

$$U(c) = 1 - C(c). \tag{6}$$

4.2.1 Confidence of a single shapelet

We assume that we have a shapelet $S = (s, l, \delta, c)$ and a time series T. We also assume that the distance $dist(s,T)$ between T and S is less than or equal to δ, so we say that S matches T and then T is classified as class c.

Due to imperfections in sensors architecture, observed time series are often affected by measurement noise. Therefore, distance between time series T and shapelet S contains uncertainty in itself. To account for uncertainty in measurements, we define distance between T and S as a random variable d

$$d = dist(s,T) + \epsilon, \tag{7}$$

where ϵ is some random variable with mean equal to 0 and standard deviation equal to σ.

Knowing that shapelet S matches time series T, confidence $C_S(c)$ of classifying T as class c based on a shapelet S has two components: 1) confidence in the fact that d is less than a threshold δ and 2) confidence in the ability of shapelet S to accurately classify time series T. To take these two components into account, we define $C_S(c)$ as

$$\begin{aligned} C_S(c) =& C_S(d < \delta, class(T) = c | S \text{ matches } T) \\ =& C_S(d < \delta | S \text{ matches } T) C_s(class(T) = c | S \text{ matches } T), \end{aligned} \tag{8}$$

where the assumption of independence is considered. To calculate the first component $C_S(d < \delta | S \text{ matches } T)$ we use Equation 7

$$\begin{aligned} C_S(d < \delta | S \text{ matches } T) =& P(d < \delta | dist(s,T) < \delta) \\ =& P(dist(s,T) + \epsilon < \delta | dist(s,T) < \delta) \\ =& P(\epsilon < \delta - dist(s,T) | dist(s,T) < \delta). \end{aligned} \tag{9}$$

If we assume that ϵ follows some distribution with 0 mean and σ standard deviation (we do not assume any parametric form of the distribution), then we can calculate a lower bound for confidence using Cantelli's inequality (sharpened version of Chebyshev's inequality which is commonly used in medical domain [7])

$$\begin{aligned} C_S(d < \delta | S \text{ matches } T) =& P(\epsilon < \delta - dist(s,T) | dist(s,T) < \delta) \\ =& 1 - P(\epsilon > \delta - dist(s,T)) \\ \geq& 1 - \frac{\sigma^2}{\sigma^2 + (\delta - dist(s,T))^2} \\ =& \frac{(\delta - dist(s,T))^2}{\sigma^2 + (\delta - dist(s,T))^2}. \end{aligned} \tag{10}$$

The closer $dist(s,T)$ is to δ the lower the confidence is. Also, larger σ means lower confidence. Figure 5b demonstrates graphically this concept.

We calculate the second component of Equation 8 as

$$\begin{aligned} & C_S(class(T) = c | S \text{ matches } T) \\ =& P(class(T) = c | dist(s,T) < \delta) \\ =& \frac{P(class(T) = c, dist(s,T) < \delta)}{P(dist(s,T) < \delta)} \quad = Precision(S), \end{aligned} \tag{11}$$

where we used $Precision$ from Equation 5. Then, the lower bound for confidence is calculated as

$$C_S(c) \geq \frac{(\delta - dist(s,T))^2}{\sigma^2 + (\delta - dist(s,T))^2} * Precision(S). \tag{12}$$

Since both terms in this product take value between 0 and 1, the highest value of the $C_S(c)$ is 1. Equation 12 incorporates two measures: how far is the time series from the shapelet and the performance of the shapelet in the training data.

Note that *each shapelet has its own threshold and hence its own confidence distribution. Therefore, for any time series, we compute the distance $dist(s,T)$ between the time series and the shapelet. If the distance is less than or equal to the threshold, then the confidence $C_S(c)$ is computed using Equation 12. If the distance is greater than the threshold, the confidence is not computed because the time series lies outside the region (circle) of the shapelet. Hence, the confidence is computed only when the shapelet matches the time series.*

4.2.2 Aggregated class confidence

If there is only one shapelet from the class c that matches the time series, then the confidence estimate for predicting the time series as class c is the same as the shapelet's confidence and is computed using Equation 12.

Now, we consider the case of computing the class confidence when multiple shapelets match the time series. Let us start with a simple case.

Simple Case.

Assume we have two shapelets S_1 and S_2 from class c that match the time series, then the class confidence $C(c)$ of classifying the time series as c is computed as

$$\begin{aligned} C(c) &= C_{S_1 \cup S_2}(c) \\ &= C_{S_1}(c) + C_{S_2}(c) - C_{S_1 \cap S_2}(c) \\ &= C_{S_1}(c) + C_{S_2}(c) - C_{S_1}(c) * C_{S_2}(c), \qquad (13) \end{aligned}$$

where the assumption of the shapelets' independence is considered. The value of the class confidence $C(c)$ is greater than the confidence of any of the individual shapelets S_1 and S_2, which does make sense because our confidence for classifying the time series as class c is increased by having two matched shapelets.

General Case.

Assume that $S^c = \{S_1, S_2, \ldots, S_N\}$ is the set of all shapelets from class c that match the current time series. Then the classification confidence of the time series for class c is computed as

$$\begin{aligned} C(c) &= C_{S^c}(c) \\ &= C_{S_1 \cup S_2 \cup \ldots \cup S_N}(c) \\ &= \sum_{k=1}^{N} (-1)^{k+1} \sum_{\substack{I \subset \{1,2,\ldots,N\} \\ |I|=k}} C_{S_I}(c), \qquad (14) \end{aligned}$$

where the last sum runs over all subsets I of the indices $\{1, \ldots, N\}$ which contain exactly k elements, and $S_I = \cap_{i \in I} S_i$. The value of the class confidence $C(c)$ is greater than the confidence of any of the individual shapelet $S_i; i = \{1, 2, \ldots, N\}$ because the confidence for classifying the time series as class c is increased by having multiple matched shapelets.

Now, the class confidence $C(c)$ satisfies all properties of the confidence measure, i.e. takes values on the range $[0, 1]$ and has value higher than any individual shapelet.

Moreover, the uncertainty estimate for classifying a time series should decrease as time evolves and more information about the time series becomes available. The uncertainty measure defined in Equation 6 satisfies this property. Assume that at time t there are k shapelets that match the time series and at time $t + 1$ two more shapelets match the time series. Since the k shapelets (at time t) are a subset of the $k + 2$ shapelets (at time $t + 1$) then the confidence of the $k + 2$ shapelets has to be greater than the confidence of the k shapelets (as in Equations 14), which means that the confidence at time $t + 1$ is greater than the confidence at time t. In other words, the uncertainty propagates over time and decreases as time evolves.

4.3 Modified EDSC with Uncertainty estimates (MEDSC-U)

Typically, the uncertainty generated by having only one matched shapelet is higher than the uncertainty generated by having multiple matched equal-performance shapelets. Therefore, in order to better capture a reliable uncertainty estimate, we modify the EDSC method to obtain more discriminative shapelets. In particular, the pruning and classification phases of EDSC are modified. We call our method MEDSC-U.

4.3.1 Pruning phase

The MEDSC-U method sorts the shapelets descending based on their utility scores and starts with the highest ranked shapelet S. The method removes all time series from the dataset that are covered by the shapelet and stores the shapelet S and ***all other shapelets that have the same utility score (Equation 4) as*** S***, we call these shapelets as equal-performance shapelets***.

Then, the next ranked shapelet is considered. If the shapelet covers any of the remaining time series, the shapelet and all other equal-performance shapelets are added to the extracted list and all covered time series are removed. The method iteratively does so until all time series in the dataset are covered. In this manner, the MEDSC-U method ends up with a longer list of equal-performance shapelets that is used for classification purposes. The list of the shapelets extracted from the MEDSC-U method is longer than the extracted list from the EDSC method. However, the longer list does not contain shapelet with lower utility score (Equation 4) than the shorter list. The intuition is that *with the richer model, the uncertainty would be better estimated.*

4.3.2 Classification phase

To better use the uncertainty estimates generated by MEDSC-U, we change the classification process. When we have a time series with an unknown label, we compute the distance between the current stream of the time series and ***all*** discriminative shapelets extracted by MEDSC-U (we do not start with the highest one until a match is found as EDSC). Then we compute the uncertainty for each class based on ***all matched*** shapelets from that class. If we do not have any matched shapelet for a class then we do not have uncertainty associated with that class.

At each time point we compute the uncertainty for each class. The time series is classified based on the class that has minimum uncertainty. If the produced uncertainty is not satisfactory for the user, the method continues classifying the time series until a predefined level of uncertainty is obtained.

Note that if a shapelet matches the time series at time t, then, just by chance, there is a high probability that the same shapelet matches the time series at time $t+1$ and at consecutive time points. That will increase the confidence estimation while it happens just by chance. To prevent that, MEDSC-U does not allow the same shapelet to match the time series at consecutive time points. In other words, if the shapelet S has length l and matches the time series at time t, then MEDSC-U does allow S to match the time series at time points $t + 1$ up to $t + l/2$.

4.3.3 Recommending an uncertainty threshold

At each time point MEDSC-U classifies the time series as the class that has the minimum uncertainty at that time point. MEDSC-U continues classifying the time series as long as the produced uncertainty is not satisfactory for the user. However, the domain expert might have no clue about what the recommended uncertainty threshold would be. In addition, the uncertainty threshold may be different not only from one dataset to another but even from one class to another. Therefore, we provide a simple way to find a good uncertainty threshold to be used for each class in order to get confident results.

We apply the MEDSC-U method on a validation dataset which is different from the training dataset. Then, we compute the precision of each class at different values of uncertainty thresholds. Based on a desired value of precision, the user can choose the corresponding uncertainty threshold (we illustrate this on a case study in Section 5.2.1).

5. EXPERIMENTAL RESULTS

We evaluated our method on 20 time series datasets (Table 2) from the UCR time series archive [9]. To compare the proposed method to the original EDSC method, we use the same set of parameters as recommended in [15] ($\alpha = 3, minL = 5, maxL = L/2$, *and Chebyshev's inequality for computing distance threshold*). The code and all details about our experiments can be found at our website[3].

Table 2: Datasets description. The 20 datasets are sorted descending by the number of classes and time series length.

Dataset	Training Size	Test Size	Time Series Length	No. of Classes
SwedishLeaf	500	625	128	15
FacesUCR	200	2050	131	14
FaceAll	560	1690	131	14
MedicalImages	381	760	99	10
Fish	175	175	463	7
Lightning7	70	73	319	7
OSULeaf	200	242	427	6
SynCon	300	300	60	6
OliveOil	30	30	570	4
DiatomSizeReduction	16	306	345	4
TwoPatterns	1000	4000	128	4
CBF	30	900	128	3
GunPoint	50	150	150	2
ECGFiveDays	23	861	136	2
ECG200	100	100	96	2
MoteStrain	20	1252	84	2
TwoLeadECG	23	1139	82	2
SonyAIBORobotSurface	20	601	70	2
SonyAIBORobotSurfaceII	27	953	65	2
ItalyPowerDemand	67	1029	24	2

We used 4 evaluation measures:

1. **Coverage**: the percentage of time series out of the test time series dataset that are classified. For example, if 8 out of 10 time series are classified then the coverage is 80%.

2. **Relative Accuracy**: the average between sensitivity and specificity relative to the covered time series. For example, if the coverage is 80% and all covered time series are classified correctly then the relative accuracy is 100%.

3. **Accuracy**: the average between sensitivity and specificity relative to the total number of test time series. if one method has better accuracy and less coverage and the second method has better coverage and less accuracy, so it is not clear which one is better. Therefore, in order to obtain a fair comparison, the method classifies the not-covered examples according to the majority class and then the accuracy is computed.

4. **Earliness**: the fraction of the time points used for classification.

5.1 MEDSC-U versus EDSC

MEDSC-U allows more equal-performance shapelets to be included in the final list in order to have reliable uncertainty estimates. We first evaluate the effect of including more shapelets on the accuracy and earliness performance of the MEDSC-U method without using the advantages of uncertainty estimates. Therefore, the MEDSC-U method classifies the test time series based on the closest matching shapelet even with high uncertainty estimates 1.

The accuracy (blue circles) and earliness (red squares) of both the MEDSC-U and EDSC methods on each dataset are shown in Figure 6. In the area under the diagonal, the EDSC method is better than the MEDSC-U method, while in the area above the diagonal the MEDSC-U method is better than the EDSC method. We plot 100-earliness instead of earliness to preserve "*the higher the better*" property.

As shown in Figure 6, most of the circles lie very close to the diagonal line which means that MEDSC-U method has comparable accuracy with the EDSC method. Therefore, including equal-performance shapelets into the model does not negatively affect the accuracy of the model. The reason for this is that: since the additionally included shapelets have the same accuracy performance on the training data, it would be better to include them to allow for variability among time series' patterns and that would slightly improve the earliness of the classification decision, as evident in the right panel of Figure 6.

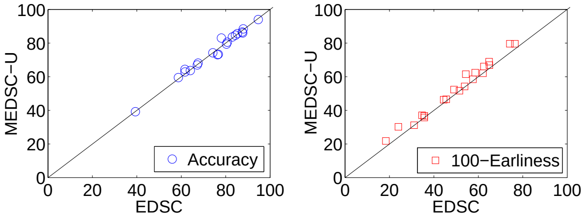


Figure 6: Comparison between MEDSC-U and ESDC with respect to the accuracy (left) and the earliness (right). The left (right) panel has 20 circles (squares) for the 20 datasets where each circle (square) represents the accuracy (100-earliness) of the two methods on exactly one dataset, respectively. The EDSC method is better on the lower triangle while the MEDSC-U method is better on the upper triangle. Points (datasets) on the diagonal indicate that the two methods have similar accuracy/earliness performance. For more details (per each dataset) about these experiments check the website.

5.2 Case studies

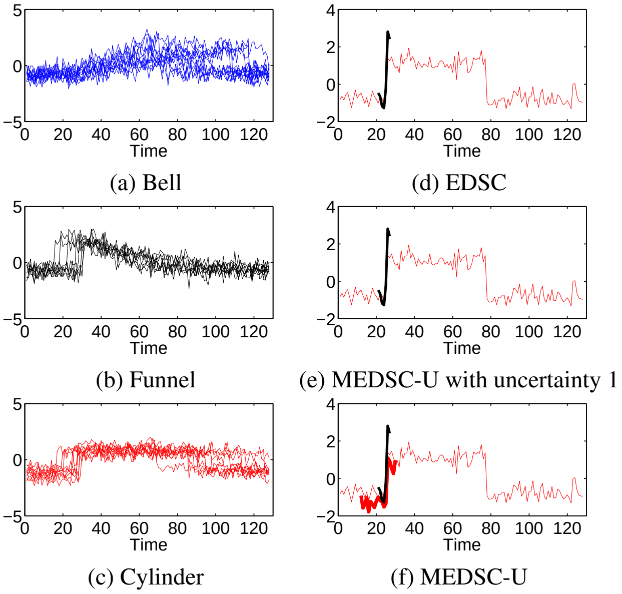
We show the effectiveness of our uncertainty method on real examples from different datasets. In particular, we show how the method provides *more confident class prediction by either having a shapelet that confidently matches the time series (Section 5.2.1) or having multiple shapelets match the time series (Section 5.2.2) as in Equation 14*. In addition, we explain how to appropriately choose an uncertainty threshold using CBF dataset as a case study.

5.2.1 *Classification based on a confident shapelet: CBF case study*

The CBF dataset has three classes, as shown in the first column of Figure 7. The EDSC method has extracted one shapelet from each class and has achieved 88% accuracy. However, the EDSC method incorrectly classified the cylinder (red) example shown in Figure 7d as funnel (black) example. The same case happened if we just used the MEDSC-U with uncertainty 1 as shown in panel (e).

[3] http://www.dabi.temple.edu/∼mohamed/uncertainty/.

This is because the shapelets extracted by EDSC or MEDSC-U from the funnel and the cylinder classes are similar to each other. Therefore, it might happen, as shown in the figure, that a cylinder example is classified incorrectly as funnel example or vice versa. To overcome this problem we need to measure the uncertainty for the classification decision, especially from these two classes, and therefore we use lower value of uncertainty in order to obtain more confident classification by the MEDSC-U method.



(a) Bell

(d) EDSC

(b) Funnel

(e) MEDSC-U with uncertainty 1

(c) Cylinder

(f) MEDSC-U

***Figure 7:** (a,b,c) Time series from the CBF datasets from the bell, funnel, and cylinder classes, respectively. (d,e) A cylinder time series which is incorrectly classified by the EDSC method (d) and the MEDSC-U method with uncertainty* 1 *(e), respectively. (f) The cylinder time series is correctly classified by MEDSC-U with lower uncertainty* 0.13*.*

The uncertainty associated with the classification using the black shapelet at time point 27 is 0.49. If this value is not satisfactory (because we know that these two classes are similar to each other and as we show later that this uncertainty is not enough to classify a time series as black), we might wait and not provide classification at this point in the hope that the uncertainty reduces under a certain threshold. At time point 30 a cylinder shapelet matches the time series with uncertainty 0.13, so the method is more confident to '*correctly*' classify the time series as a cylinder class than to classify it as a funnel class.

Therefore, the user can decide if he/she is satisfied about the classification results using the uncertainty estimates provided by the method. That uncertainty threshold could be varied from one class to another class. For example, since we know that the shapelets from the funnel and cylinder classes are similar, then if the test time series is classified as one of these classes, the user may lower the value of the uncertainty threshold and delays the decision in order to obtain accurate results. On the other hand, if the test time series is classified as the bell class (completely different from the other two classes), then higher value of uncertainty threshold might be sufficient in order not to delay the results.

Recommending an uncertainty threshold.

As we just mentioned, the uncertainty threshold may be different from one class to another class. The domain expert might have no notion about the recommended threshold for each class. To find the recommended uncertainty threshold, MEDSC-U is applied on a validation dataset and the precision of each class is computed for each uncertainty threshold as shown in Figure 8.

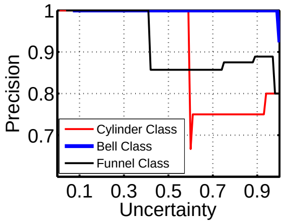


***Figure 8:** The precision of the MEDSC-U method on the CBF dataset for the three classes at different values of uncertainty threshold. The blue class (the blue curve along with the upper x-axis) is easy to classify by MEDSC-U even with high uncertainty, while the red and black classes are similar to each other which requires lower uncertainty threshold for each of them.*

As shown in Figure 8, the blue class has 100% precision at each level of uncertainty and then the precision drops at high value of uncertainty (very close to uncertainty estimate 1), while the precision of the other two classes drop earlier. This illustrates that the blue class is dissimilar to the other classes and can be recognized by our method even with high value of uncertainty such as 0.9 (or even higher). However, the precision for the red (Cylinder) and black (Funnel) classes dropped at approximately 0.6 and 0.4, respectively. Therefore, 0.6 (0.4) would be a good estimate for the uncertainty thresholds for the red (black) class, respectively. Also, if the domain expert has a desired value of precision, say for example 0.9, then we draw a horizontal line at the precision 0.9 in Figure 8 and find the uncertainty thresholds corresponding to the intersection of that line and the three (classes) precision curves.

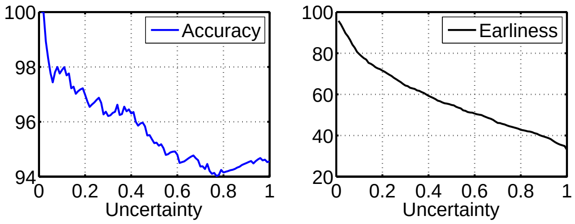
This is very consistent with the results reported in Figure 7. Since the test time series is classified initially as black class with uncertainty 0.49 and we know from Figure 8 that the recommended threshold for the black class is 0.4, we wait and do not classify the time series at that time point. The next matched shapelet classifies the time series as red class with uncertainty 0.13 and the recommended threshold for the red class is 0.6, therefore the classification is done at that time point which is the correct class.

To simplify the presentation of the paper, the remaining results are shown using a single threshold for all classes instead of showing results using class-specific and dataset-specific threshold.

Uncertainty versus performance measures.

This begs the question "*How does the uncertainty threshold affect the accuracy and earliness performance?*". We have shown that if we reduce the uncertainty threshold from 0.49 to 0.13 we get more accurate results, but it delays the decision for 3 time points. The results for varying different threshold versus the accuracy is shown in Figure 9a. *Note that we use the same uncertainty threshold for all classes.* It is clear from the figure that the accuracy increases when lowering the value of the uncertainty threshold. However, as shown in Figure 9b, the classification decisions is delayed as expected. For example, using the uncertainty threshold 0.9 delays the results until, on average, 40% of the time series length (at time point ~ 51) while the accuracy would be $\sim 95\%$ (the original EDSC method has achieved accuracy 88% and earliness 35%).

Now we have conveyed the message that if the uncertainty associated with the classification decision is high, then we wait until another shapelet with lower uncertainty to match the time series in order to provide more accurate results. However, this is not the only



(a) (b)

***Figure 9:** The accuracy (a) and the earliness (b) of MEDSC-U for different uncertainty thresholds on CBF datasset. Lower value of uncertainty threshold gives more accurate results but delays the decision.*

way to obtain more confident results. More confident results may be achieved when multiple shapelets from the same class match the time series. We illustrate this using the ECGFiveDays dataset in the next section.

5.2.2 Classification based on multiple shapelets: ECGFiveDays case study

The ECGFiveDays dataset has two classes (the red and the blue classes). A time series from the blue class is shown in Figure 10. The time series matches the first (red) shapelet at time 50 with uncertainty 0.98 which means that the uncertainty associated for the red class is 0.98 and no uncertainty associated for the blue class because no blue shapelet matches the time series up to that point. Then, at time 79, the time series matches a blue shapelet with uncertainty 0.97. Therefore, at that point the method classifies the time series as a red class with uncertainty 0.98 and as a blue class with uncertainty 0.97. These uncertainties propagate until time 83, where a new (blue) shapelet matches the time series. The uncertainties from the two matched blue shapelets are aggregated using Equation 14 to give total uncertainty 0.52 for the blue class, which reveals the fact that the method is now more certain to classify the time series as a blue class (correct classification). In other words, more confident results are obtained by having multiple shapelets matching the time series.

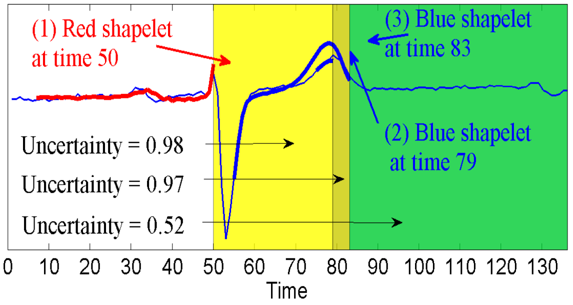


***Figure 10:** A time series from the blue class of the ECGFiveDays dataset is classified at the time point* 83 *with uncertainty* 0.52*. The yellow region denotes the region where the model is uncertain about the clasification. The green region represents the region where the model is confident about the classification.*

Therefore, the uncertainty of the classification decision is reduced by having multiple events (such as multiple shapelets appearing in the portion of the time series seen so far), and that uncertainty decreases overtime. The average uncertainty over all time series at each time point is shown in Figure 11. As shown in the figure, the uncertainty decreases over time, which emphasizes that the model becomes more and more certain about the classification as time evolves.

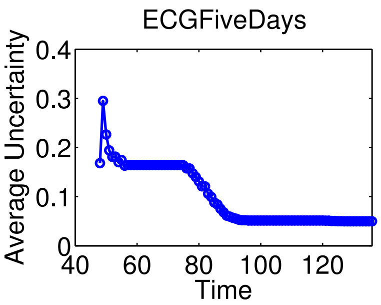
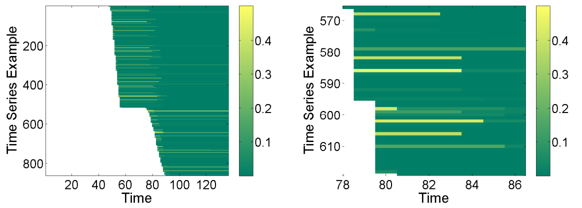


***Figure 11:** The average uncertainty over all patients at each time point for the ECGFiveDays dataset. The method starts with high uncertainty and then becomes certain about the classification as time evolves.*

For Figure 11, it has to be pointed out that the average uncertainty increases sometimes and then decreases (for example, at time 48 is 0.17 and at time 49 is 0.29 and then it decreases over time). The reason behind the increase in the average of uncertainties is that the model covered more examples at time 49 than at time 48, therefore, the uncertainties of the new covered examples increase the value of the average uncertainty from time 48 to time 49. However, the uncertainty for each time series example does not increase over time.

In Figure 12a, we show the uncertainties for each time series over time. The MEDSC-U method classifies many time series using only the first 60-80 time points but with high uncertainty (yellow bars). These uncertainties decrease over time (yellow-to-green bars). For instance, in Figure 12b, some of the time series are shown where it is clear that the method becomes more certain about the classification between time 82 - 86.



(a) (b)

***Figure 12:** The values of uncertainty over time for each time series from the ECG2FiveDays dataset. The white bar indicates that there is no classification at that point and hence there is no uncertainty.The MEDSC-U method provides more certain results as time evolves. The temporal uncertainty for each time series never increases.*

Now we have shown that the proposed interpretable early classification MEDSC-U method that provides uncertainty estimates gives more accurate results over time than the EDSC method but that affects the earliness of the decision. Therefore, the uncertainty threshold controls the trade off between the accuracy and the earliness of the method. Moreover, that uncertainty threshold may differ from one dataset to another, and even from one class to another as we explained previously in Section 5.2.1.

In order to make the flow of the paper easy to follow and due to the lack of the page space, we have shown these aspects on two datasets: CBF and ECGFiveDays. However, the results for the other datasets are very consistent with what we have shown and therefore the remaining results are presented in the appendix. In addition, all details of these results are on our website. In the next section we compare our method to other methods that provide uncertainty estimates.

5.3 Comparison to localQDA

Two existing methods we have mentioned in the related work section that provide uncertainty for time series classification that could be used in the context of early classification. The PR method represents the patient risk as a time series and estimates the uncertainty as the distance between the evolving approximate daily risk of a patient and the hyperplane learned by SVM [12]. So, the uncertainty is measured as the distance between the measurement (value of the time series at a particular time point) and the hyperplane. The proposed uncertainty estimates could be used for the early classification context, i.e. the model proceeds over the time series until a confident results (furthest to the hyperplane) obtained. However, the PR model requires labels for each time point in the time series, not just a label for the entire time series, and therefore we can not compare MEDSC-U to the PR method.

The state-of-the-art method for early classification of signals using a quadratic discriminant analysis (localQDA) classifier was developed [11]. The localQDA method does not provide uncertainty but instead it provides a reliability bound on the classifier's decision for every time point. As noted by [11], "*With probability at least τ (reliability), will the classification decision from incomplete data be the same as that which would be made from the complete data?*" The reliability τ measures the probability that the early classification will be the same as the classification at the end of the time series. The uncertainty measure provided by our method is not directly comparable to the reliability measure provided by localQDA method, but they are correlated. Thereby, we assume that the $Uncertainty = 1 - \tau$.

For fair comparison between MEDSC-U and localQDA, we use the same set of parameters as they had recommended in [11]. We show the comparison between the two methods on the ECG200 datasets as a case study in Table 3. The remaining comparisons are in the appendix.

Table 3: Comparison of our proposed MEDSC-U method to the state-of-the-art localQDA method on the ECG200 dataset at different values of uncertainty ($1 - \tau$). localQDA is accurate than MEDSC-U but MEDSC-U prvides classification decisions much ealier than localQDA.

	τ	0.1	0.25	0.5	0.75	0.9
	Uncertainty	0.9	0.75	0.5	0.25	0.1
Accuracy	localQDA	**89**	**89**	**88**	**88**	**87**
	MEDSC-U	82	81	81	82	82
Earliness	localQDA	60.19	62.18	64.02	66.27	68.34
	MEDSC-U	**23.24**	**24.23**	**28.14**	**29.45**	**35.14**

As shown in Table 3, localQDA is more accurate than MEDSC-U but the classification decision is provided much later than the decision from the MEDSC-U method. We also compute F_β score for different values of $\beta = \{0.1, 1, 2\}$. F_β score is the weighted average of the accuracy and 100−earliness where $\beta = 2$ weights earliness higher than accuracy, $\beta = 0.1$ puts more emphasis on accuracy than earliness, and $\beta = 1$ is the balanced harmonic mean.

As shown in Figure 13, our method has comparable $F_{0.1}$ score with the state-of-the-art localQDA method, which weights accuracy more than earliness. For the other two measures F_1 and F_2, MEDSC-U clearly outperforms localQDA at every level of uncertainty.

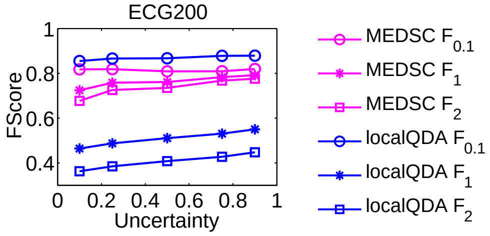


Figure 13: Comparison between the MEDSC-U method and localQDA on the ECG200 dataset for different values of uncertainty (1-tau).

To make a conclusion from all comparisons between MEDSC-U and localQDA, we plot the number of datasets where MEDSC-U has higher F_β score than the localQDA method (or vice versa) at each uncertainty threshold. The results are shown in Figure 14.

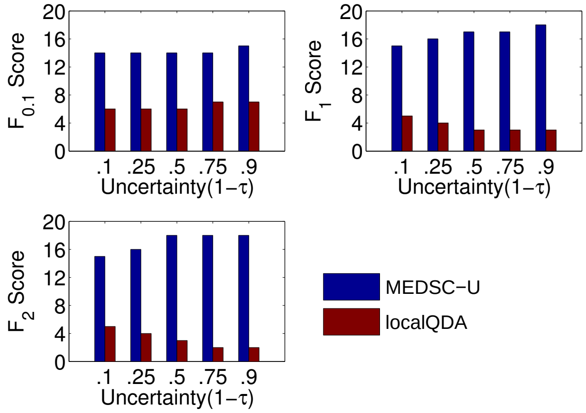


Figure 14: Number of datasest where MEDSC-U (blue bar) has better F_β score than the localQDA (red bar), or vice versa, at each uncertainty threshold. Clearly, MESDC-U ourperforms localQDA in most cases especially in F_1 and F_2 settings where the earliness is important.

As shown in Figure 14, for $F_{0.1}$, MEDSC-U outperformed the localQDA method in 13 (or little bit more) datasets at each uncertainty estimate. For F_1 and F_2 scores, MEDSC-U clearly outperformed localQDA in most of the datasets at each level of uncertainty. These results show that our proposed MEDSC-U method is comparable to or even better than the state-of-the-art localQDA method in every F score, as shown by our experiments. In addition to that, MEDSC-U is very simple to implement and provides interpretable results (shapelets) convincing to the practitioners, which are not addressed by the state-of-the-art localQDA method.

6. CONCLUSION AND FUTURE WORK

Providing classification of time series as early as possible is vital in many domains including the medical domain, where early diagnosis can save patients' lives by providing early treatment. However, applications often require the method to be interpretable and have uncertainty estimates. We extended the interpretable early classification method (EDSC) and proposed the MEDSC-U method to measure the temporal uncertainty with the classification. The proposed uncertainty estimates meets the requirements of uncertainty where it has range $[0, 1]$ and propagates over time. The

MEDSC-U method is very simple to implement and provides interpretability for the classification results. In addition, it is more effective than the state-of-the-art method, as shown in our experiments on twenty datasets. The temporal uncertainty estimates provided by MEDSC-U can be extended to the multivariate case [2, 4] where uncertainties from shapelets from different variables could be integrated as in Equation 14.

Acknowledgment

This work was funded, in part, by DARPA grant [DARPAN66001-11-1-4183] negotiated by SSC Pacific grant.

7. REFERENCES

[1] K. M. Borgwardt, S. Vishwanathan, and H.-P. Kriegel. Class prediction from time series gene expression profiles using dynamical systems kernels. In *Pacific Symposium on Biocomputing*, volume 11, pages 547–558, 2006.

[2] M. F. Ghalwash and Z. Obradovic. Early classification of multivariate temporal observations by extraction of interpretable shapelets. *BMC Bioinformatics*, 13(195), August 2012.

[3] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Early diagnosis and its benefits in sepsis blood purification treatment. In *International Workshop on Data Mining for Healthcare*, Philadelphia, PA, Sep 2013.

[4] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *IEEE International Conference on Data Mining (ICDM)*, Dallas, TX, Dec 2013.

[5] M. Grbovic and S. Vucetic. Tracking concept change with incremental boosting by minimization of the evolving exponential loss. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases, Part I, ECML/PKDD'11*, pages 516–532, 2011.

[6] M. P. Griffin and J. R. Moorman. Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *PEDIATRICS*, 107(3):97–104, 2001.

[7] F. Ieva, R. Longhi, A. M. Paganoni, and M. P. Protti. Estimating point and interval frequency of antigen-specific CD4+ T cells based on short in vitro expansion and improved poisson distribution analysis. *PLoS ONE*, 7(8), 2012.

[8] E. Keogh and T. Rakthanmanon. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of the 13th SIAM International Conference on Data Mining*, pages 668–676, 2013.

[9] E. Keogh, Q. Zhu, B. Hu, Y. Hao, X. Xi, L. Wei, and C. A. Ratanamahatana. The UCR time series classification and clustering homepage, 2011.

[10] T. Lin, N. Kaminski, and Z. Bar-Joseph. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, 24:i147–i155, July 2008.

[11] N. Parrish, H. S. Anderson, M. R. Gupta, and D. Y. Hsiao. Classifying with confidence from incomplete information. *Journal of Machine Learning Research*, 14:3561–3589, 2014.

[12] J. Wiens, J. Guttag, and E. Horvitz. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Neural Information Processing System (NIPS)*, 2012.

[13] Z. Xing, J. Pei, and P. S. Yu. Early prediction on time series: A nearest neighbor approach. In *Proceedings of the 21st international joint conference on Artifical intelligence*, pages 1297–1302, 2009.

[14] Z. Xing, J. Pei, and P. S. Yu. Early classification on time series. *Knowl Inf Syst*, 31:105–107, 2011.

[15] Z. Xing, J. Pei, P. S. Yu, and K. Wang. Extracting interpretable features for early classification on time series. In *Proceedings of 11th SIAM International Conference on Data Mining*, pages 439–451, 2011.

[16] L. Ye and E. Keogh. Time series shapelets: A new primitive for data mining. In *Proceeings of the 15th ACM SIGKDD Conference on Knolwedge Discovery and Data Mining (KDD)*, 2009.

APPENDIX

A. REMAINING RESULTS

The accuracy, earliness, and coverage performance for some of the datasets are shown in Figure 15, and the results for comparing the proposed MEDSC-U method with the localQDA method are shown in Figure 16, while, for the lack of page space, the remaining results are on our website.

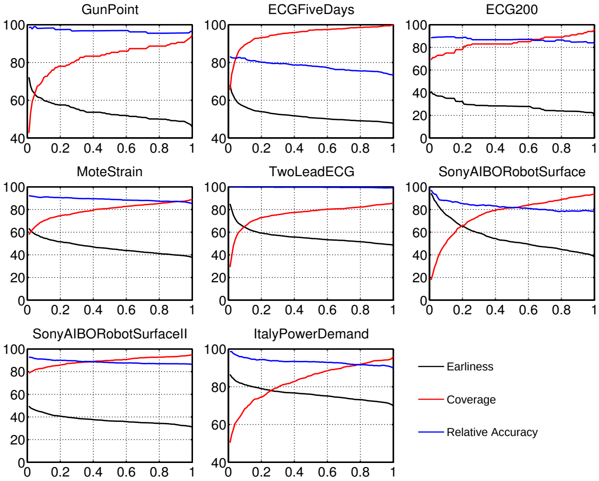


Figure 15: Uncertainty (x-axis) versus relative accuracy (blue), earliness (black), and coverage(red) for 5 datasets.

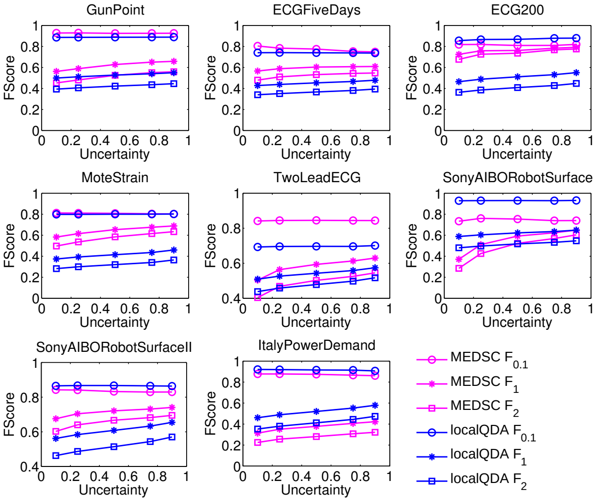


Figure 16: Comparison between MEDSC-U and localQDA for different values of uncertainty. X-axis is the uncertainty threshold $(1 - \tau)$ and y-axis is the F_β score.