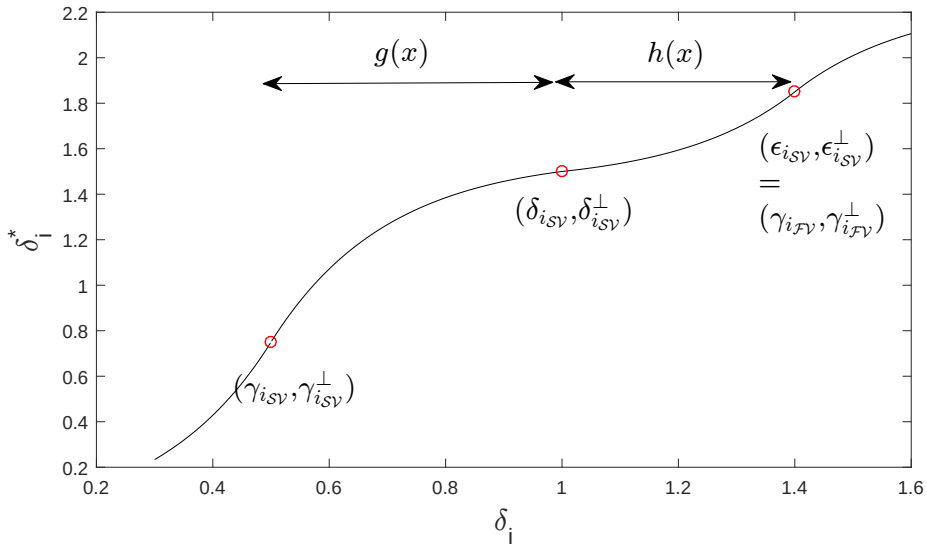

Figure 4.3: Optimized Piece-wise Interpolated Function p.

vector of new ECG sample in the spherical coordinate system is calculated and the mapping function is applied on its hyper-spherical coordinates to yield the transformed feature vector. After this step, we calculate the transformed data samples in the Cartesian coordinate system, which further fed into the personalized classification stage defined by Eq. 2.8 to generate the corresponding type of yellow alarm.

4.6 Experimental Results

In this section, the performance of the proposed method is evaluated in terms of two aspects. We first analyze the classification performance of the system and then present the comparative results with respect to other representative ECG classifiers. Furthermore, the classification results are partitioned into two sets: red alarms generated by the global classifier and the final labels by combining the yellow and red alarms. In this way, the impact of personalized classifier on the final labels can be revealed. Finally, the prediction power

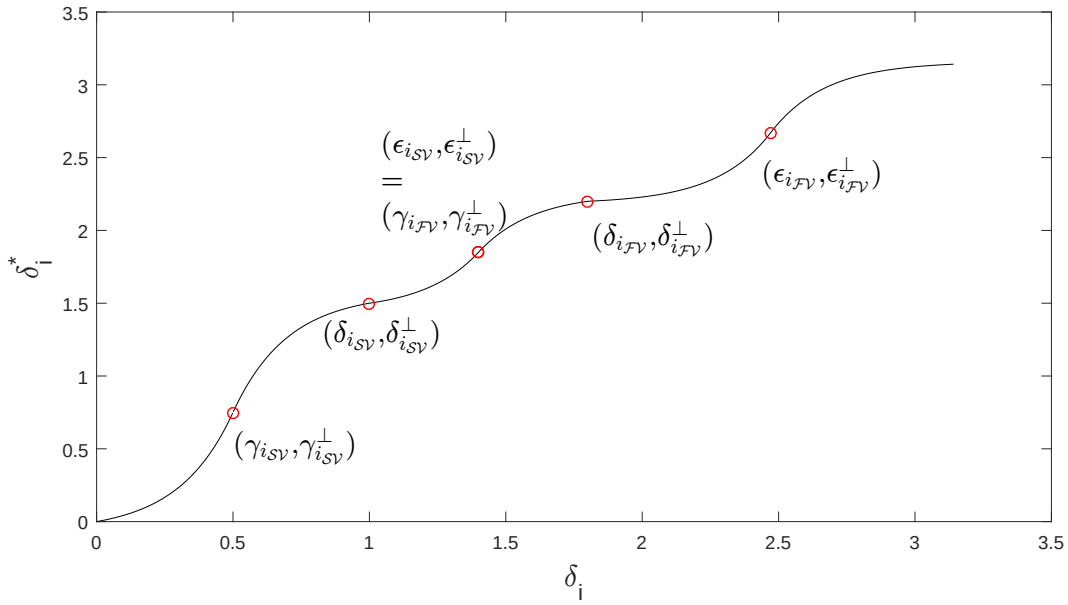

Figure 4.4: Optimized Mapping Function f.

of the proposed method in terms of providing precaution hints for upcoming red alarms is evaluated.

4.6.1 Classification Performance

The experimental results are assessed in terms of classification performance of 4 AAMI ECG classes using the test subset of MITBIH Arrhythmia DS2. Originally, DS2 contains 15357 samples after feature extraction. While training the personalized classifier, the first 20% of the normal samples serve as initialization set for the personalized dynamic normal cluster. Therefore, we exclude the first 20% of normal samples from each record since they're included in the training process. Consequently, the actual test set contains 12414 samples in total consisting of 10105 type-N, 1702 type-V, 508 type-S and 99 type-F samples.

To present the result, we select the weighted k-Nearest Neighbors method with $k = 10$ as our choice of global classifier because it is one of the low-complexity methods with a relatively

Table 4.1: Cumulated Confusion Matrix for All Records in DS2. The numbers are *final label*(*primary label by global classifier*)

		Ground Truth			
		N	V	S	F
Result	N	9255(10076)	21(38)	72(90)	1(5)
	V	657(22)	1678(1663)	8(2)	9(7)
	S	71(6)	3(1)	417(416)	0(0)
	F	122(1)	0(0)	11(0)	89(87)

good classification accuracy [66].The parameter α, used in Eq. 4.8 in the deviation detection module is set to 1 for test purpose.

Table 4.1 summarizes the cumulated confusion matrix for all records in the test set. In order to compare the result of global classifier (red alarms) and combined results (final labels including both red and yellow alarms), the sample numbers are presented in the following format: *final label*(*primary label by global classifier*). In order to measure the classification performance, we adopt three metrics: accuracy(Ac), sensitivity(Se), specificity(Sp), as proposed in [12, 13, 16]. All three metrics are calculated based on the true positive TP, false positive FP, false negative FN and true negative TN in a binary confusion matrix, where we combined all abnormality classes into a one abnormal class/ or where one class is the specific abnormality class and all other abnormality and normal classes combined into one class, whichever is correct.. Therefore all four metrics are calculated for each class by converting the 4x4 matrix to a 2x2 matrix.

While cumulative classification results are demonstrated in Table 4.1, the robustness of the proposed method should be evaluated based on the performance variation over 22 test records in DS2. Further, medians and IQRs (interquartile range) for each metric and each class are included in Table 4.2 to represent the robustness of proposed methods. The robustness of the system is asses in terms of the variation between the performance of different types. The system is more robust if this variation is lower. In Table 4.2, we observe that among all

abnormality classes, the proposed method demonstrates a stable performance on class V but a lower stability for classes S and F.

As MITDB is widely used to verify ECG classifier performance, we compared the proposed system with five significant methods proposed in the literature. According to AAMI standards, the performance of ECG classification should be evaluated over the binary classifiers applied to *Ventricular (V)* versus *non-V* types and *Supraventricular (S)* versus *non-S* types. For methods proposed in the literature, the same evaluation metrics are commonly used applied to records from MITDB. To standardize the metrics, we select 11 ECG records which are common among all 5 methods and compare the median of each classification metrics over these 11 records. The comparison results are presented in Table 4.3. Generally speaking, the proposed method shows a higher sensitivity for both types V and S. Especially for type S, the proposed method shows an advantage over all three metrics compared to the five reference methods.

4.6.2 Prediction Performance

As an important feature of the proposed method, yellow alarms triggered by the personalized classifier indicate a higher probability of observing subsequent abnormalities. In order to verify this functionality, all beats following a yellow alarm of a specific type is investigated to asses the chance of upcoming red alarms of different types. This process is repeated for yellow alarms of all types. We only account for the first abnormality type which occurs after

Table 4.2: Classification Performance and Within-Set Variation of Proposed System

statistics	N			V			S			F		
	Ac	*Se*	*Sp*	*Ac*	*Se*	*Sp*	*Ac*	*Se*	*Sp*	*Ac*	*Se*	*Sp*
cumulated	92.4	91.59	95.93	94.38	98.59	93.71	98.67	82.09	99.38	98.85	89.9	98.92
median	94.45	92.21	95.42	96.17	99.55	95.71	99.38	80.65	99.84	99.11	90.91	99.11
IQR	6.33	10.08	11.91	5.17	1.64	8.62	1.76	19.35	0.61	1.58	23.33	1.49

Table 4.3: V and S classification performance compared with five algorithms in literature using 11 common records in MITDB

Methods	VEB			SVEB		
	Ac	Se	Sp	Ac	Se	Sp
Proposed	96.6	98.2	92.4	98.63	88.89	99.41
Hu *et al.* [12]	94.8	78.9	96.8	N/A	N/A	N/A
de Chazal *et al.* [9]	96.4	77.5	N/A	N/A	N/A	N/A
Jiang and Kong [15]	98.8	78.9	96.8	97.5	74.9	98.8
Ince *et al.* [16]	97.9	90.3	98.8	96.1	81.8	98.5
Kiranyaz *et al.* [17]	98.9	95.9	99.4	96.4	68.8	99.5

the yellow alarm. As we used confusion matrix to evaluate the classification accuracy, the performance of prediction can be summarized by a confusion matrix with the 3 abnormal types. Probabilities of observing a certain type of abnormal beat after a yellow alarm is calculated using the prediction confusion matrix and compared to the prior probability of observing the abnormality of the same type. This process is formulated in the following two equations:

$$\begin{aligned} P(\hat{y}_{k+i} = X_r | \hat{y}_k = X_y) &= \frac{\#\text{ of } y_{k+i} = X \text{ after } \hat{y}_k = X_y}{\#\text{ of true alarms after } \hat{y}_k = X_y} \\ P(\hat{y}_{k+i} = X_r) &= \frac{\#\text{ of true alarm of type } X \ (y_k = X)}{\#\text{ of all true alarms}} \end{aligned} \tag{4.10}$$

The prediction power of each abnormality type is evaluated by comparing $P(\hat{y}_{k+i} = X_r | \hat{y}_k = X_y)$ and $P(\hat{y}_{k+i} = X_r)$. A shown in Table 4.4, the probability of observing a certain type of abnormalities after a yellow alarm is higher than its prior probability and this fact is consistent for abnormality types. For example, without knowing the type of a yellow alarm, the probability of observing a type V sample is 71.54%, while the probability of observing a type V sample after observing a yellow alarm of a type V is 77.45% (XXX% higher than the prior probability). The improvement are consistent among all three types of abnormalities

Table 4.4: predictive probability versus prior probability without windowing

		# of predicted ground truth			% of predicted ground truth		
		V	S	F	V	S	F
yellow alarm	V	467	122	14	**77.45**	20.23	2.32
	S	36	15	0	70.59	**28.41**	0
	F	40	60	5	38.10	57.14	**4.76**
total		543	197	19	**71.54**	**25.96**	**2.50**

Table 4.5: predictive probability versus prior probability within 10 beats' window

		# of predicted ground truth			% of predicted ground truth		
		V	S	F	V	S	F
yellow alarm	V	290	85	12	**74.94**	21.96	3.10
	S	22	13	0	62.86	**37.14**	0
	F	29	37	6	40.28	51.39	**8.33**
total		341	135	18	**69.03**	**27.32**	**3.64**

but the system shows a stronger prediction power for type S.

In the above analysis we consider the first subsequent red alarm regardless of the time passes since the preceding yellow alarm. In order to study the impact of the timing window (the time between the yellow alarm and the subsequent red alarm), we also studied a window of 10 consecutive samples following a yellow alarm. Similarly, the prior and posterior probabilities are compared to evaluate the performance of the prediction capacity as shown in Table.4.5.

Compared with the result without windowing, the prediction performance within a 10-beat window shows that the proposed algorithm can better predict the occurrence of abnormalities if a certain timing window is used. Especially for type S, the probability of observing a sample of type S within 10 beats after a yellow alarm of type S is 27.32%, i.e. the posterior probability rises to 37.14%. With almost 10% increase, it is proved that the yellow alarm types are informative. The results shows that same improvements are made within the 10-sample window as well. In general, the predicting performance are promising, indicating the efficiency of personalized classifier and deviation analysis. We believe that this concept is