Individualized Machine-learning-based Clinical Assessment Recommendation System

Devin Setiawan[1], Yumiko Wiranto[2], Jeffrey M. Girard[2], Amber Watts[2], Arian Ashourvan[2]

[1] *The University of Kansas, Department of Electrical Engineering and Computer Science, 1415 Jayhawk Blvd. Lawrence, KS 66045*

[2] *The University of Kansas, Department of Psychology, 1415 Jayhawk Blvd. Lawrence, KS 66045*

Abstract

Background: Traditional clinical assessments often lack individualization, relying on standardized procedures that may not accommodate the diverse needs of patients, especially in early stages where personalized diagnosis could offer significant benefits. We aim to provide a machine-learning framework that addresses the individualized feature addition problem and enhances diagnostic accuracy for clinical assessments.

Methods: Individualized Clinical Assessment Recommendation System (iCARE) employs locally weighted logistic regression and Shapley Additive Explanations (SHAP) value analysis to tailor feature selection to individual patient characteristics. Evaluations were conducted on synthetic and real-world datasets, including early-stage diabetes risk prediction and heart failure clinical records from the UCI Machine Learning Repository. We compared the performance of iCARE with a Global approach using statistical analysis on accuracy and area under the ROC curve (AUC) to select the best additional features.

Findings: The iCARE framework enhances predictive accuracy and AUC metrics when additional features exhibit distinct predictive capabilities, as evidenced by synthetic datasets 1-3 and the early diabetes dataset. Specifically, in synthetic dataset 1, iCARE achieved an accuracy of 0·999 and an AUC of 1·000, outperforming the Global approach with an accuracy of 0·689 and an AUC of 0·639. In the early diabetes dataset, iCARE shows improvements of 1·5-3·5% in accuracy and AUC across different numbers of initial features. Conversely, in synthetic datasets 4-5 and the heart failure dataset, where features lack discernible predictive distinctions, iCARE shows no significant advantage over global approaches on accuracy and AUC metrics.

Interpretation: iCARE provides personalized feature recommendations that enhance diagnostic accuracy in scenarios where individualized approaches are critical, improving the precision and effectiveness of medical diagnoses.

Funding: This work was supported by startup funding from the Department of Psychology at the University of Kansas provided to A.A., and the R01MH125740 award from NIH partially supported J.M.G.'s work.

1. Background

Clinical assessment is the ongoing process of gathering information about a patient and constructing an increasingly comprehensive conceptualization of their health and needs (e.g., for diagnosis, prognosis, or treatment planning). A critical task in clinical assessment is selecting the *next* piece of information to collect about the patient to maximize information gain. Given the unique nature of each patient's condition, it is essential to recognize that there are often no one-size-fits-all solutions. This need for personalization is especially high when symptom presentation and treatment effectiveness are heterogeneous across individuals; examples include oncology, psychiatry, and the treatment of chronic diseases such as diabetes, cardiovascular disease, and neurodegenerative disorders.[1–3] We can also find an example from the study of dementia where the informativeness of APOE ε4 as one of the best predictors of dementia varies by race.[4–6] Although useful, achieving personalization in clinical practice is challenging. Personalization requires massive data, raising privacy concerns and the potential misuse of sensitive information.[7] In this paper, we will discuss how the framework of *individualized feature selection* from machine learning (ML) can be used to efficiently guide the task of personalization in clinical assessment.

Feature selection is the process of identifying and prioritizing the most relevant and informative input variables (i.e., features) that will optimize model performance, interpretability, and generalization while minimizing model complexity and overfitting.[8] Overfitting occurs when a model becomes too complex, capturing noise in addition to the signal, which causes it to fail to generalize to unseen data (e.g., novel patients or new observations of known patients). It is important to reduce overfitting so that the model performs well in real-world scenarios.[9] This is usually achieved by using popular techniques like sequential forward selection (SFS) or backward elimination, which iteratively add or remove features to see their effect on model performance.[10–13] However, these traditional techniques lack individualization, resulting in every patient being given the same recommendation. Personalized feature selection, on the other hand, places patients at the center of the decision-making process, taking into account each individual patient's unique characteristics and recognizing that different patients may need different thresholds for diagnosis.[14] This problem definition aligns with the aim of personalized clinical assessment recommendations, where the goal is to tailor the choice of the next test based on the unique characteristics of each patient.

Recent studies in individualized feature selection have begun to address this gap by developing methods that personalize the selection of features based on individual patient data. For instance, a study on wearable electroencephalogram (EEG) monitoring platforms uses linear discriminant analysis (LDA) and the least absolute shrinkage and selection operator (LASSO) method to select discriminative features tailored to each subject's seizure patterns.[15] However, this approach does not focus on dynamic and iterative feature addition and is highly specific to EEG data.

Additionally, an unsupervised personalized feature selection framework tailors feature selection to each instance in high-dimensional data.[31] However, our objective is to tackle a supervised individualized feature selection problem. Additionally, a framework employing fixed prediction models, local feature explainers, and ensembles of imputed samples provides flexible risk estimation for samples with missing features.[16] This framework relies heavily on imputations and uses a single fixed prediction model. On the other hand, we want to create a framework that provides an individualized model directly without the need for imputations or a singular prediction model.

We propose a general framework that recommends which features to obtain next for each patient, promoting a more accurate diagnosis through personalization. Taking inspiration from *locally weighted learning*, iCARE leverages patient-specific data to tailor the selection of clinical assessments for individualized healthcare recommendations used in diagnosis.[17–19] Our approach utilizes a locally weighted model tailored to each patient, which was analyzed using a feature explainer, to dynamically adapt feature selection strategies based on each patient's unique characteristics. The iCARE framework relies on three main components: (1) a sample weight calculation module, (2) an ML model trained on weighted samples, and (3) a feature explainer for the generated models. We analyzed the framework using synthetic datasets to show its personalization capability and also compared it with a traditional approach on both synthetic and real-world datasets. We hypothesized that our framework would provide more accurate diagnoses than the traditional approaches.

2. Methods

2.1. Framework Architecture

Figure 1 provides an overview of the architecture of our iCARE framework. The architecture consists of an input processing module identifying missing features of incoming patients. A similarity calculation module is then used to calculate similarity scores between incoming patients and patients in the pool of known cases. This pool of known cases comprises labeled data, which includes values for predictive features such as age, sex, and test results, along with an outcome label indicating whether the individual is sick or not sick. It can be created from any data source representing known past cases with relevant features. Using these weights, a weighted logistic regression is trained using the pool of known cases. The weights assigned to each sample reflect its relevance to the novel patient's profile, allowing for personalized model training. The trained model is then analyzed using Shapley Additive Explanations (SHAP) to quantify the importance of individual features in the locally trained logistic regression model.[20] SHAP values are based on cooperative game theory in which a prediction is broken down to show how each feature influenced the outcome of a model. Finally, the feature recommendation module will take the explanations and produce a recommendation. It evaluates whether the

feature is present in the patient's initial feature set. If any significant feature is missing, the framework recommends its inclusion to further enhance predictive accuracy.

$$(1)\ sample\ weight\ = \frac{1}{distance}$$

$$(2)\ Feature\ Importance_i\ = \frac{1}{N}\sum_{n=1}^{N} |SHAP_i|$$

$$(3)\ Recommended\ Feature\ = max(Feature\ Importance)$$

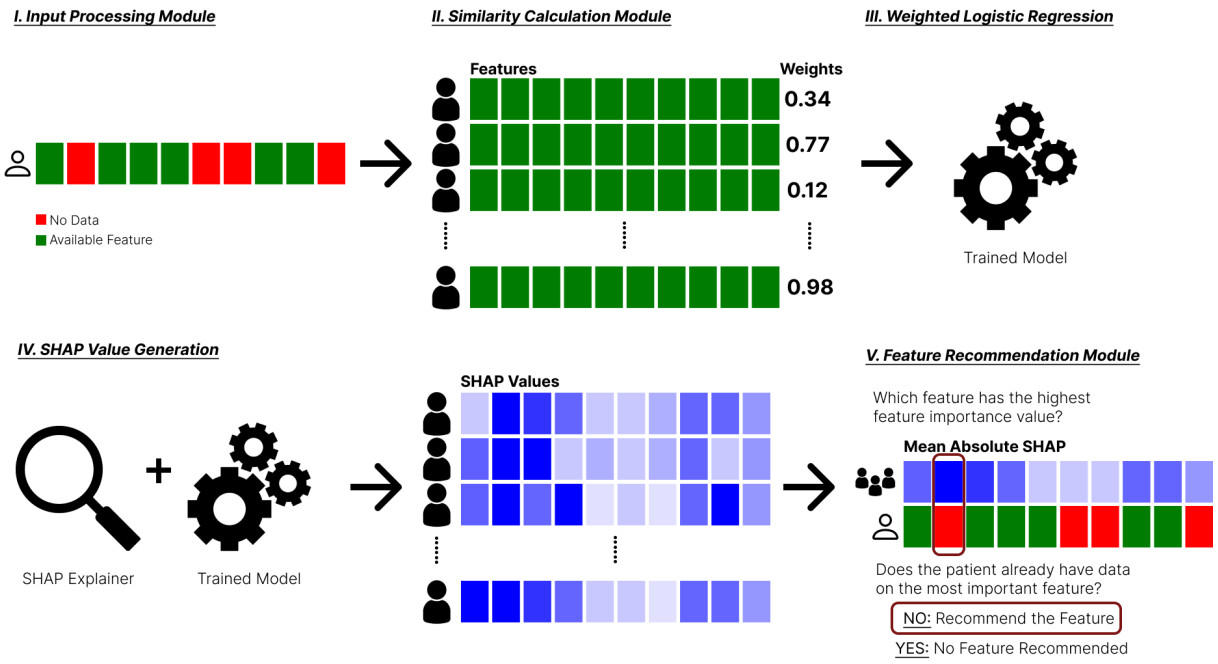

Figure 1: **Architecture of the iCARE framework.** Data were obtained from an incoming patient (I), and weights were generated for the pool of known cases in the Similarity Calculation Module (II). Using these sample weights, we generate a weighted logistic regression model for an incoming patient (III). SHAP values are then generated using a SHAP explainer for all the subjects in the pool of known cases (IV). The Feature Recommendation Module will then gather all the individual SHAP values and produce a recommendation if there is a missing feature that can be recommended to the patient (V).

2.2. Experimental Design

Figure 2 provides an overview of the experiment to compare the performance of the iCARE recommendation against a global feature recommendation (i.e., Global) strategy. Initially, we define a set of initial features using the *least important feature*. The dataset was split into a pool of known cases and test cases. With the procedure applied before, the test cases will have only the initial features, simulating conditions where patients don't have all the informative features. From here, we generated a global recommendation and an individualized recommendation. The

global recommendation is done by training a logistic regression on the pool of known cases and analyzing it with SHAP. The feature with the highest SHAP value is selected for recommendation. On the other hand, the individualized recommendation uses the iCARE framework. We then evaluate the recommendation and repeat this process 100 times.

To evaluate the recommendations, we append the pool of known cases and a single case with the recommended feature value from the initial dataset. We then train a logistic regression using the pool of known cases and predict the outcome for the single case using the model. In addition to this, we also define the locally weighted (LW) procedure, which just uses a weighted logistic regression on this step instead of a regular logistic regression. We repeat this process until all test cases receive the predicted outcome. We then collect this prediction and calculate the accuracy and AUC (Area Under the Receiver Operating Characteristic Curve) metrics. These metrics were then averaged over 100 iterations.

This experiment was repeated using a different number of initial features. We selected the least informative feature as it represents a realistic scenario where incoming patients will more likely have less informative features. This iterative approach allowed us to assess the impact of the model performance across the various frameworks on different initial available features.

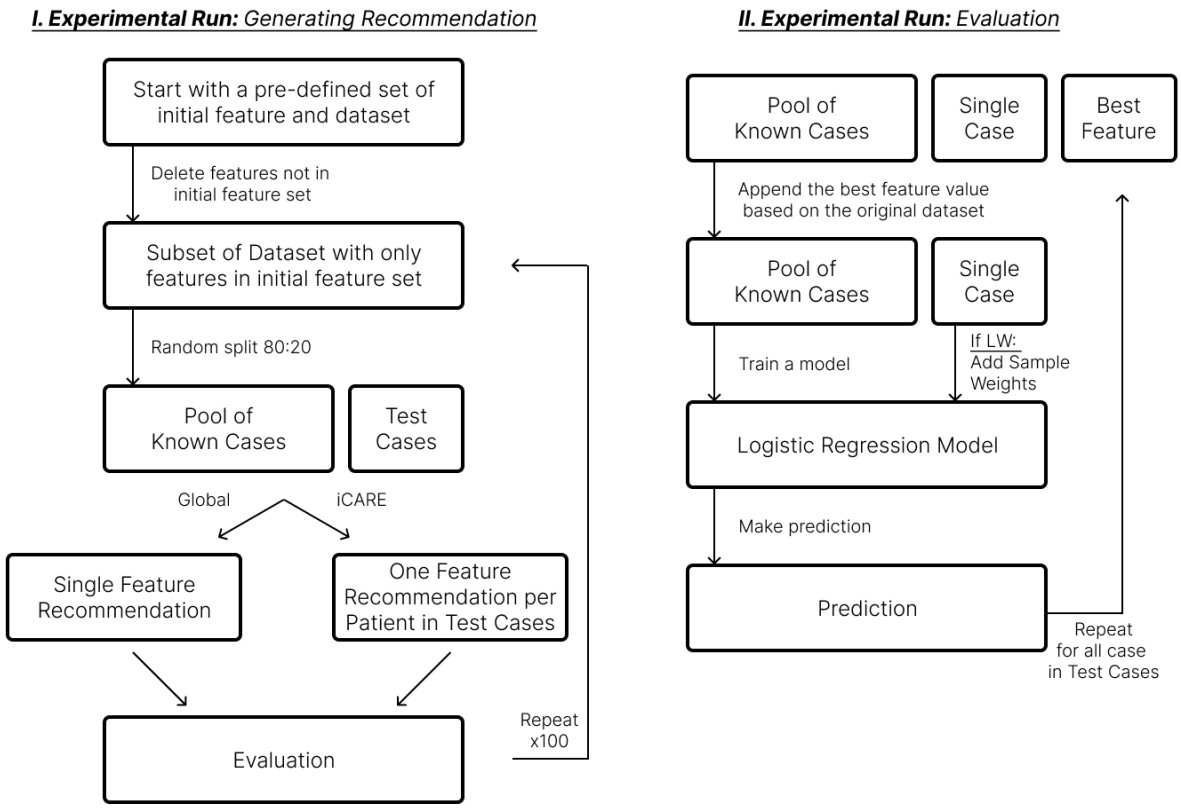

Figure 2: **Experimental workflow to evaluate the iCARE framework.** The figure above highlights the main experimental workflow to evaluate the iCARE framework against traditional global feature selection. This workflow produces two distinct approaches to generating recommendations, as shown by the Global (i.e., global feature selection) and iCARE (i.e., individualized feature selection) split in part I. In addition, there are two distinct approaches to training the inference model, as shown in part II, where the logistic regression model can be trained with or without sample weights (i.e., LW or no LW). This produces four approaches: Global, Global+LW, iCARE, and iCARE+LW.

2.3. Dataset

We evaluate our framework with both synthetic and real-world datasets. The synthetic datasets were created to simulate ideal and non-ideal scenarios. The real-world datasets utilized in this study were obtained from the UCI Machine Learning Repository, specifically the early-stage diabetes risk prediction, heart failure clinical records, and heart disease dataset.[21–23] We provide the code to generate the synthetic dataset, as well as the details on preprocessing steps for real-world datasets in the supplementary materials.

2.4. Statistical Analysis

We performed t-tests (α=0.05) on the accuracy and AUC to assess the statistical significance of the performance differences between the four frameworks. To account for familywise error and reduce the risk of Type I errors, we applied Holm-adjusted p-values to the results of these multiple comparisons. Using Holm-adjusted p-values provides a more conservative and reliable measure of statistical significance compared to the standard p-values obtained from the t-test.

3. Findings and Interpretation

3.1. Reasoning Process of the Framework

The iCARE framework is grounded in the principle of localized learning and feature importance analysis to generate personalized clinical recommendations. A locally weighted logistic regression model trained using weighted patient samples from the repository of known cases focuses on learning similar patients. Due to this, iCARE will excel in scenarios where patients with similar profiles benefit from similar recommendations. Given an incoming patient with available features and a selection of potential features to be recommended, iCARE will be able to recommend the best feature given that the available features are informative of the predictiveness of the added features. For example, if in the dataset, groups of people aged below 50 benefit from additional feature A, and those above 50 benefit from additional feature B, iCARE will be able to capture this information from age (i.e., available feature) and recommend the appropriate feature (i.e., feature A or B) to an incoming patient that will give the best information gain.

We created synthetic datasets 1-3 to simulate ideal scenarios and confirm our hypothesis on the reasoning process of iCARE. Synthetic dataset 1 represents the most ideal scenario, characterized by two additional features exhibiting predictive power over different regions of the initial features value space, as shown in Figure 3. Conversely, synthetic dataset 2 illuminates the necessity for sample-weighted inference (as indicated by LW) when confronted with non-linear predictive regions highlighted in Figure 4[24,25]. Furthermore, synthetic dataset 3 serves as a testament to the robustness of our framework, particularly in scenarios involving overlapping regions on the initial features value space that can be seen in Figure 5.

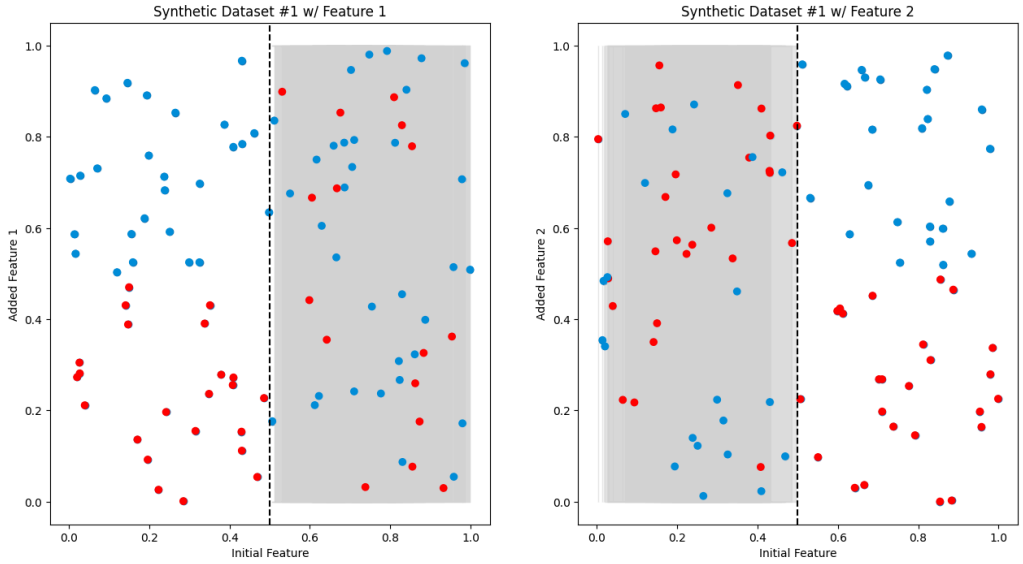

Figure 3: **Synthetic dataset 1.** Two 2D scatter plots displaying the relationship between the initial feature (x-axis) and the added feature (y-axis). The red dots represent negative samples (e.g., sick patients), while the blue dots represent positive samples (e.g., healthy patients). The left plot depicts added Feature 1, exhibiting predictive power for Initial Feature < 0·5, while random noise is observed in the shaded area above Initial Feature > 0·5. The right graph illustrates added Feature 2, demonstrating predictive power for Initial Feature > 0·5, with random noise observed in the shaded area below Initial Feature < 0·5.

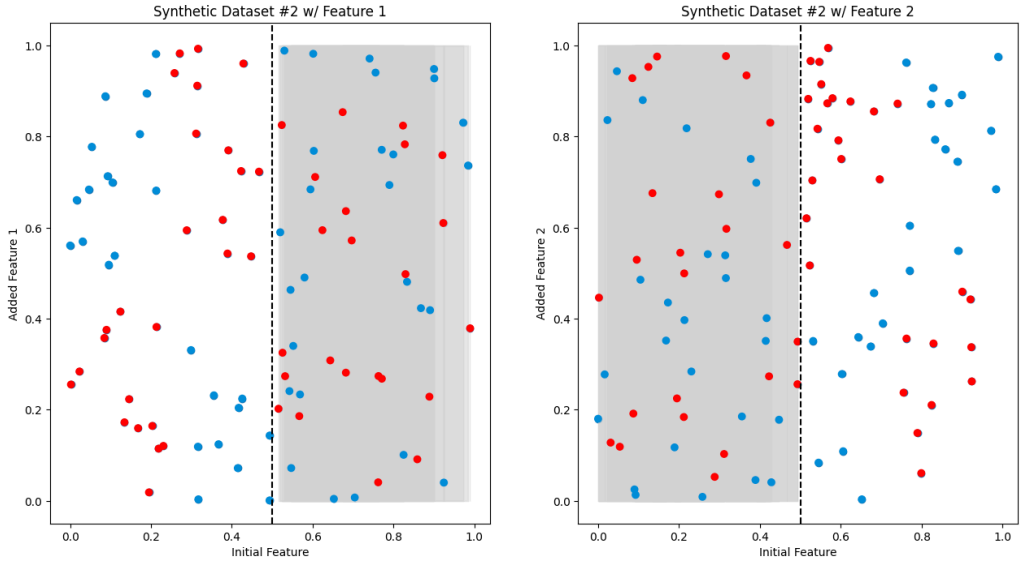

Figure 4: **Synthetic dataset 2.** Two 2D scatter plots, similar to Figure 3, showcase the relationship between the initial feature (x-axis) and the added feature (y-axis). The red dots represent negative samples (e.g., sick patients), while the blue dots represent positive samples (e.g., healthy patients). Notably, the predictive area in this dataset exhibits a non-linear pattern, suggesting a more complex relationship between the features.

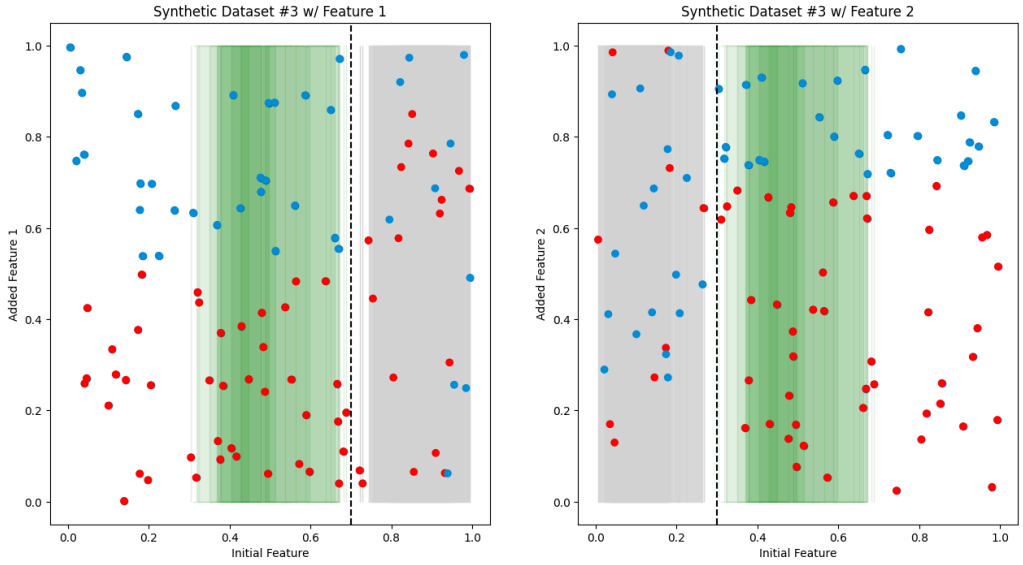

Figure 5: **Synthetic dataset 3.** 2D scatter plots resembling Figure 3, depicting the relationship between the initial feature (x-axis) and the added feature (y-axis). The red dots represent negative samples (e.g., sick patients), while the blue dots represent positive samples (e.g., healthy patients). Notably, the left graph demonstrates predictive power for $X < 0{\cdot}7$, while the

right graph showcases predictive power for X > 0·3. The green-shaded region highlights an overlapping area (0·3 < X < 0·7) where both features possess equal predictive power.

We created synthetic Datasets 4-5 to simulate hypothetical non-ideal scenarios. Synthetic Dataset 4, depicted in Figure 6, simulates a non-ideal scenario where both additional features are equally useful (i.e., the available feature does not give information about the predictiveness of the additional features). Notably, both the left and right graphs showcase identical predictive regions. This visualization emphasizes scenarios where both features share the same predictive power in the same region. In synthetic dataset 5, represented in Figure 7, we created a scenario where only one additional feature out of the rest is useful. This visualization emphasizes scenarios where one feature dominates others regarding predictive strength. The iCARE framework is expected to have similar performance to a global feature selection, highlighting no added benefit from personalization.

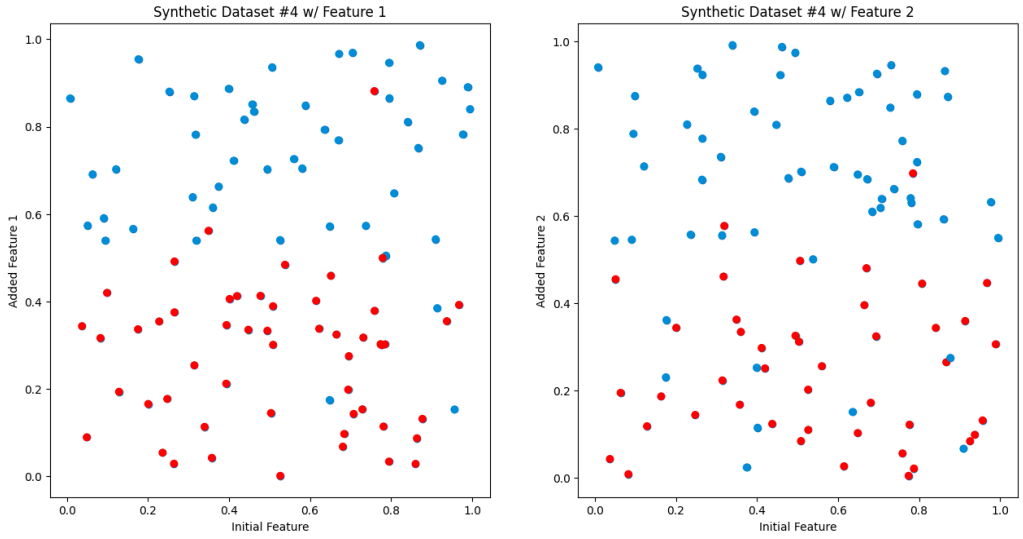

Figure 6: **Synthetic dataset 4.** Scatter plots depicting the relationship between the initial feature and the added feature, resembling the format of Figure 3. Notably, both the left and right graphs illustrate identical predictive regions.

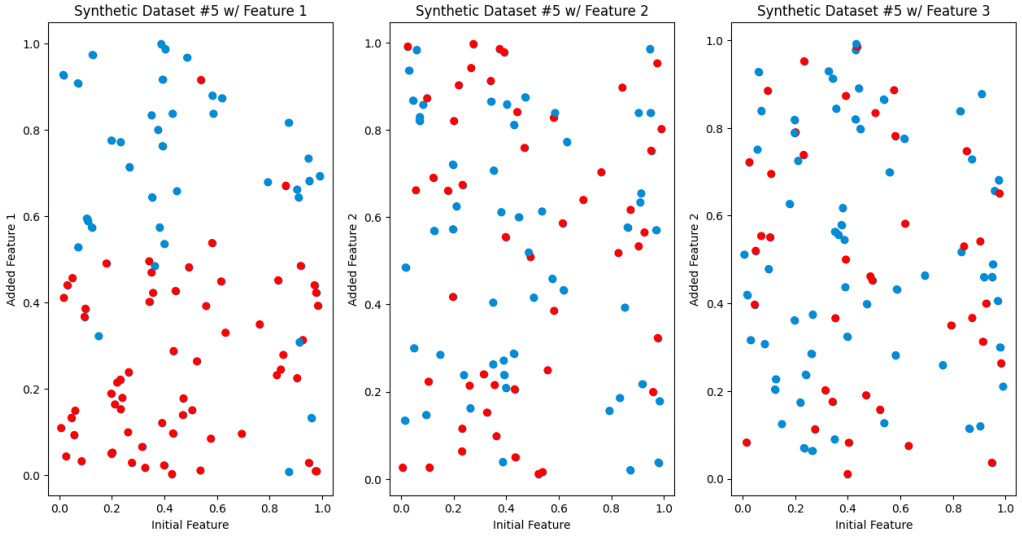

Figure 7: **Synthetic dataset 5.** Each scatter plot represents a different feature's predictive power. The first scatter plot demonstrates strong predictive capability, while the other two plots depict features with limited predictive utility. This visualization underscores the scenarios where one feature overpowers the other features.

3.2. Performance on Synthetic Dataset

In Figure 8, we provide the comparison between the different approaches on the synthetic datasets 1-3. In synthetic dataset 1, where two additional features exhibit predictive power over distinct regions, the iCARE frameworks are expected to perform significantly better than the Global frameworks. As expected, we obtain statistically significant (α=0.05) differences in iCARE versus Global metrics and iCARE+LW versus Global+LW metrics using t-test, with the iCARE performing better than its Global counterpart, confirming the framework's capability to provide the best recommendation when additional features' predictive capabilities are clearly distinguishable given the initial feature values. Similarly, in synthetic dataset 2, characterized by non-linear predictive regions, the iCARE frameworks, especially when incorporating locally weighted inference (LW), are expected to outperform their non-LW counterparts. Statistical significance (α=0.05) across all comparisons can be found on our t-test, notably for iCARE versus iCARE+LW and Global versus Global+LW metrics, unseen in synthetic datasets 1 and 3. Furthermore, in synthetic dataset 3, featuring overlapping regions with identical predictive power for both features, both iCARE frameworks are expected to perform slightly better than the Global framework. The actual results align with this expectation, demonstrating the framework's ability to make accurate recommendations even in cases where features exhibit similar predictive capabilities. Similar to synthetic dataset 1, statistical significance (α=0.05) can be observed on our t-test when comparing iCARE with the Global framework. These results confirm the hypothesis of our framework's ability to give the best recommendation in cases where the

additional features' predictive capabilities can be clearly distinguished, given the initial feature values.

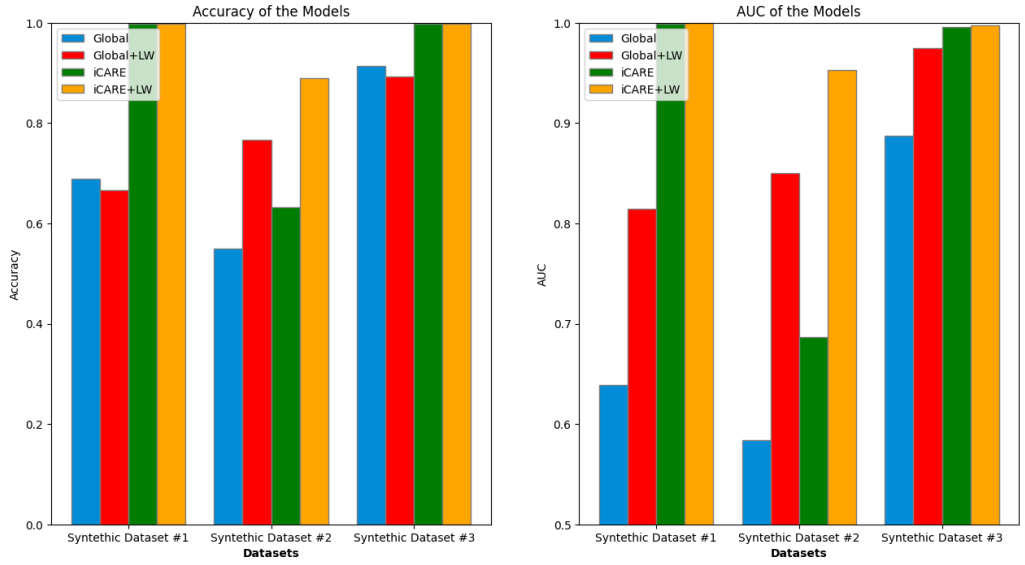

Figure 8: **Performance summary of Synthetic Dataset 1 - 3.** Comparison of accuracy (left) and area under the curve (AUC) (right) across three synthetic datasets. Each bar group represents a dataset, with values indicated for both global and local weighted metrics. For Dataset 1, the accuracy stands at 0·689, 0·667, 0·999, 0·999 with an AUC of 0·639, 0·814, 0·999, 1·0. In Dataset 2, the accuracy stands at 0·551, 0·767, 0·632, 0·891 with an AUC of 0·584, 0·850, 0·687, 0·953. Dataset 3 accuracy stands at 0·914, 0·894, 0·998, 0·998, along with an AUC of 0·888, 0·974, 0·996, 0·998. This comparison highlights variations in performance across the different synthetic datasets that represent ideal scenarios.

In Figure 9, we provide the comparison between the different approaches on the synthetic datasets 4-5. For synthetic dataset 4, characterized by features sharing the same predictive power in the space of the initial feature value, we expected little to no difference when comparing iCARE versus Global frameworks. The actual outcome confirms this expectation, as both iCARE and Global frameworks exhibit similar performance. Similarly, for synthetic dataset 5, where there is only one useful feature, we expected a similar outcome to synthetic dataset 4. As predicted, the actual results show little variation between iCARE and Global frameworks. We observed some variances in performance; however, this can primarily be attributed to the use of locally weighted inference (i.e., LW) rather than inherent differences in the iCARE framework itself.

Furthermore, synthetic datasets 4 and 5 revealed no statistical significance for comparisons between iCARE and iCARE+LW versus Global and Global+LW metrics, which aligned with our hypothesized outcomes. The complete result of the statistical test can be seen in Table 1. These

findings further confirm the hypothesis of our framework's ability to give the best recommendation in cases where the additional features' predictive capabilities can be clearly distinguished, given the initial feature values.

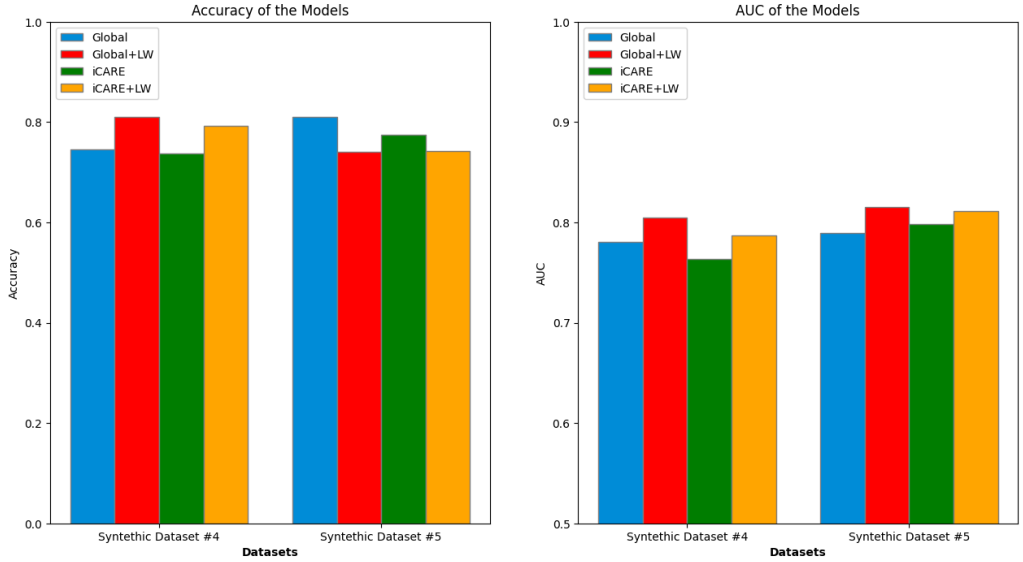

Figure 9: **Performance summary of synthetic dataset 4 - 5.** Comparison of accuracy (left) and area under the curve (AUC) (right) across Synthetic Datasets 4 and 5. Each bar group represents a dataset with performance metrics for both global and iCARE. For dataset 4, accuracy values obtained were 0·747, 0·810, 0·738, 0·792, and AUC values obtained were 0·781, 0·805, 0·764, 0·787. For dataset 5, accuracy values obtained were 0·811, 0·740, 0·774, 0·742, and AUC values obtained were 0·790, 0·815, 0·799, 0·811. These results reveal two distinct scenarios where iCARE learning fails to substantially improve global learning regarding feature addition and inference.

Table 1: **Statistical Test Results of Synthetic Dataset 1 - 5**

	Dataset 1		Dataset 2		Dataset 3		Dataset 4		Dataset 5	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
iCARE vs Global	0·310 ***	0·360 ***	0·082 ***	0·103 ***	0·084 ***	0·108 ***	-0·009	-0·017	-0·036 **	0·009
iCARE +LW vs Global+ LW	0·332 ***	0·186 ***	0·124 ***	0·102 ***	0·104 ***	0·023 ***	-0·018	-0·018	0·002	-0·004

iCARE vs iCARE +LW	0·000	-0·001	-0·259 ***	-0·266 ***	0·000	-0·002	-0·054 ***	-0·023	0·032 **	-0·013
Global vs Global+ LW	0·022	-0·176 ***	-0·217 ***	-0·267 ***	0·021 *	-0·087 ***	-0·064 ***	-0·025	0·071 ***	-0·025

The table shows the differences in accuracy (ACC) and area under the curve (AUC) metrics among different approaches. Specifically, it compares iCARE versus Global, iCARE+LW versus Global+LW, iCARE versus iCARE+LW, and Global versus Global+LW. Statistical significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. The p-values used for testing the statistical significance above are the Holm-adjusted p-values to correct for multiple comparisons.

3.3. Performance on Real-World Dataset

In extending our evaluation to real-world scenarios, we scrutinize the performance of our framework on datasets representative of clinical contexts. Specifically, we assess its effectiveness in predicting outcomes in early diabetes and heart failure datasets, leveraging a range of personalized recommendations of features to enhance predictive accuracy and AUC metrics. In the experiment on the early diabetes dataset using three initial features, we observe that personalization leads to increased Accuracy and AUC, as seen in Figure 10. The superiority of iCARE models is shown to be statistically significant, as shown in Table 2. The three initial features that were used in this experiment are age, gender, and obesity status. Using a global approach, the feature that is recommended the majority of the time is polydipsia (i.e., excessive thirst; 75/100 iterations). It suggests that, on average, polydipsia might be more informative across the entire population when combined with age, gender, and obesity status. However, when using iCARE, two features are recommended: Polyuria (Frequent Urination) and Polydipsia. On average, Polyuria is recommended for 68% of patients, and Polydipsia is recommended for 32% of patients. The prominence of the Polyuria recommendation suggests that Polyuria might provide more relevant or discriminative information for certain patients. Polydipsia and Polyuria are both classic symptoms of diabetes.[26,27] The framework's recommendation pattern suggests variability in symptom presentation and importance among different patients. The higher recommendation of Polyuria suggests that for many patients, this symptom may be an earlier or more pronounced indicator of diabetes than Polydipsia.

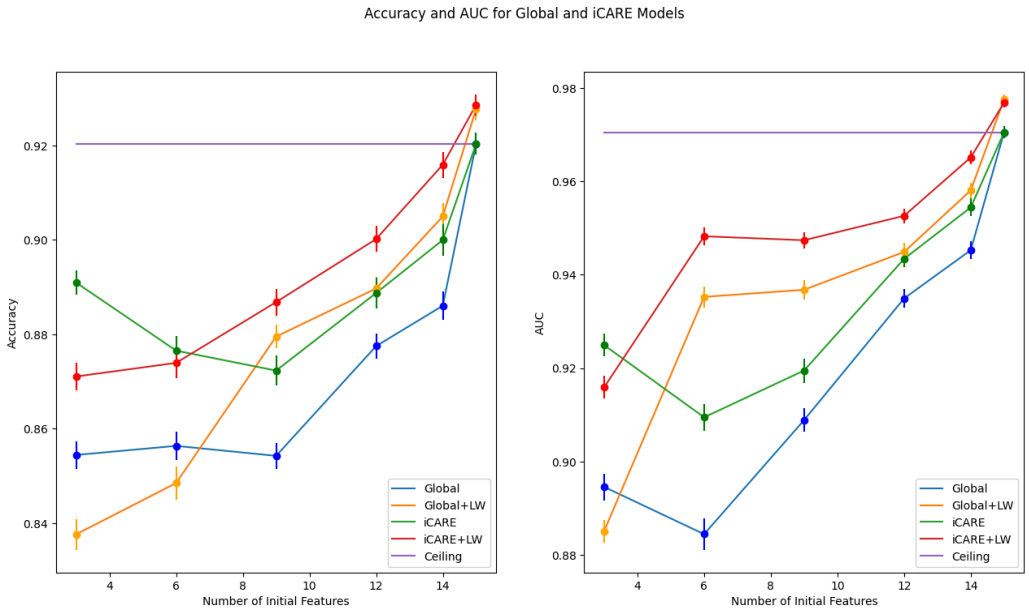

Figure 10: **Early Diabetes dataset performance summary.** This figure illustrates the mean performance of the early diabetes dataset on different feature spaces on accuracy and AUC metrics, with global and local perspectives represented by blue/orange and green/red lines, respectively. Error bars at each data point represent the standard deviation from the mean. The line graphs the maximum number of features towards the ceiling, represented by the purple line. The ceiling model represented an ML model trained on all features.

Table 2: **Early Diabetes Dataset Performance Statistical Test**

	3		6		9		12		14	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
iCARE vs Global	0·037 ***	0·030 ***	0·020 ***	0·025 ***	0·018 ***	0·011 **	0·011 **	0·008 **	0·014 **	0·009 **
iCARE +LW vs Global +LW	0·033 ***	0·031 ***	0·025 ***	0·013 ***	0·007	0·011 ***	0·010 *	0·008 **	0·011 **	0·007 **
iCARE vs iCARE +LW	0·020 ***	0·009 *	0·003	-0·039 ***	-0·015 **	-0·028 ***	-0·011 *	-0·009 ***	-0·016 ***	-0·011 ***

Global vs Global +LW	0·017 ***	0·009 *	0·008	-0·051 ***	-0·025 ***	-0·028 ***	-0·012 **	-0·010 **	-0·019 ***	-0·013 ***

The table shows the differences in accuracy (ACC) and area under the curve (AUC) metrics among different approaches applied to the early diabetes dataset, where the first row represents the number of initial features. Statistical significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. The p-values used for testing the statistical significance above are the Holm-adjusted p-values to correct for multiple comparisons.

In contrast, the iCARE framework does not yield substantial benefits on the heart failure dataset, as shown in Figure 11. We observed overlapping error bars in both accuracy and AUC metrics across different feature spaces in this dataset. In some instances (e.g., accuracy for the number of features = 4), iCARE models even underperform compared to their Global counterparts, highlighting the limitations of the approach in specific contexts. The statistical test in Table 3 shows that this difference in performance is statistically significant. This finding highlights a similar outcome to synthetic dataset 4, where it shows no added benefit when the additional features to be recommended have no distinct predictive capabilities, as well as synthetic dataset 5, where only one additional feature is useful as seen in synthetic dataset 5.

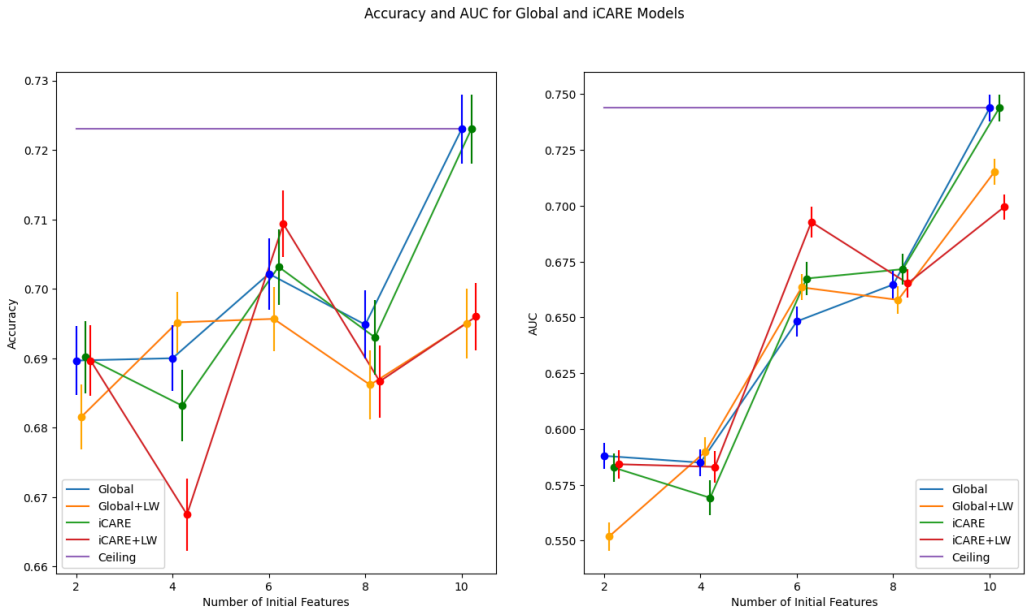

Figure 11: **Heart failure dataset performance summary.** This figure presents a comprehensive overview of mean accuracy and AUC metrics across various feature spaces on the heart failure dataset, offering insights into global and local perspectives depicted by blue/orange and green/red lines, respectively. Error bars show the standard deviation, while convergence towards the maximum features underscores notable trends.

Table 3: **Heart Failure Dataset Performance Statistical Test**

	2		4		6		8	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
iCARE vs Global	0·001	-0·005	-0·007	-0·016	0·001	0·019	-0·002	0·007
iCARE+ LW vs Global+ LW	0·008	0·032 ***	-0·028 ***	-0·007	0·014	0·029 **	0·001	0·007
iCARE vs iCARE+ LW	0·001	-0·002	0·016	-0·014	-0·006	-0·025 *	0·006	0·006
Global vs Global+ LW	0·008	0·036 ***	-0·005	-0·005	0·006	-0·015	0·009	0·007

The table shows the differences in accuracy (ACC) and area under the curve (AUC) metrics among different approaches applied to the heart failure dataset, where the first row represents the number of initial features. Statistical significance is denoted by * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. The p-values used for testing the statistical significance above are the Holm-adjusted p-values to correct for multiple comparisons.

3.4. Comparison with Other Frameworks

To further evaluate the effectiveness of iCARE, we compared its performance in feature selection against the imputation-based explanation-guided (Eguided) feature selection method.[16] As shown in Figure 12, iCARE achieved a +6% higher F1 score on average in the Early Diabetes Dataset and +10.6% higher F1 score on average in the Heart Disease Dataset over Eguided. The Heart Failure dataset highlighted before showed an example where personalized feature selection is not needed. The result showed consistency to our previous results where both iCARE and Eguided failed to perform better than a global feature selection. However, iCARE performed on average +2.2% higher in F1 score than Eguided on this dataset. These results suggest that iCARE generally provides superior performance on these datasets compared to Eguided, though a few important considerations must be noted. First, the Eguided framework was originally evaluated on a much larger dataset, comprising 100,000 samples and 252 features, while our datasets are considerably smaller. Additionally, in the original Eguided study, XGBoost was employed as the prediction model, whereas we used logistic regression for both, given that iCARE relies on

logistic regression. This difference in model selection may influence the relative performance of Eguided, as XGBoost could provide additional performance benefits in larger or more complex datasets. Overall, these findings reinforce the potential of iCARE as a robust feature selection framework for personalized and dynamic feature recommendations, particularly in clinical datasets of smaller scale.

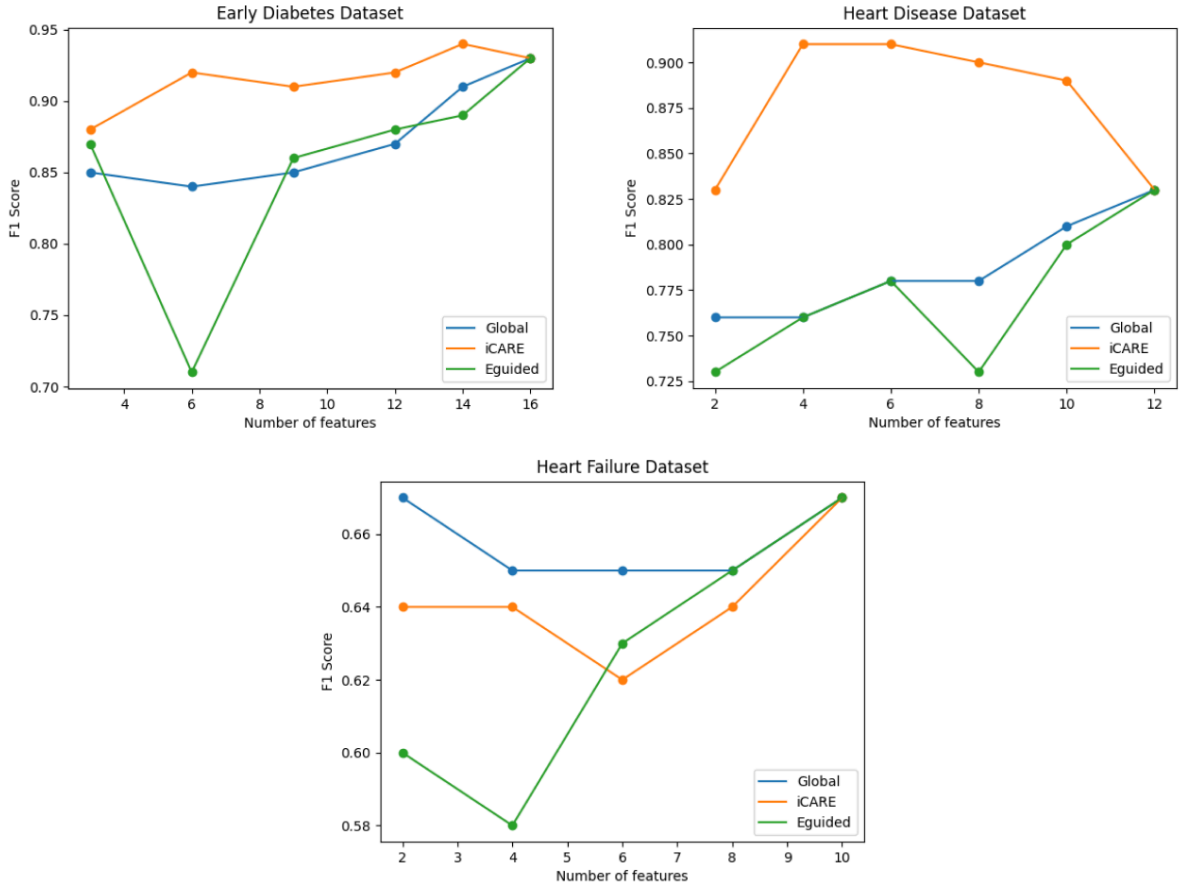

Figure 12: **iCARE vs Global vs Eguided feature selection performance summary.** This figure presents a comparative overview of the F1 scores across various feature spaces on three real-world datasets (Early Diabetes, Heart Disease, Heart Failure). The graphs depict the performance of the iCARE feature selection (orange), Global (blue), and Eguided imputation-based explanation-guided feature selection (green) approaches. The x-axis represents the number of initial features, while the y-axis shows the corresponding F1 scores, providing insight into each method's effectiveness in handling different feature subsets.

4. Discussion

4.1. Importance of Sample Weighing

Sample weighing was utilized in the sample calculation model to create a weighted logistic regression model. This weighted model emphasizes patients with characteristics similar to those of incoming patients. The weighing strategy allows SHAP to be locally sensitive to the context

of the current incoming patient, enabling feature importance rankings to be customized to the individual context rather than global trends. Sample weights have previously been used to address various challenges. For instance, a recent study proposed a weighted undersampling scheme for Support Vector Machines (SVM) to improve classification performance in dealing with imbalanced data sets.[28] This method assigned different weights to the majority of samples based on their distance to the hyperplane, akin to how iCARE assigns weights based on patient similarity. Another research study focused on personalized diagnosis for Alzheimer's Disease, utilizing subject-specific classifiers iteratively refined through reweighting of training data.[29] Although not aimed at addressing the feature recommendation problem, the rationale for employing sample weighting remains relevant, as it serves to prioritize key subjects. Overall, incorporating sample weights in iCARE enables personalized feature rankings that can navigate diverse patient populations and complex clinical scenarios.

4.2. SHAP as Feature Importance Measure

Within the iCARE framework, SHAP values play a pivotal role in selecting the most important features for personalized feature addition. We use SHAP to quantify the importance of individual features within the locally trained logistic regression model. By assigning importance values to each feature for a specific prediction, SHAP facilitates understanding the factors influencing the model's output. In the context of iCARE, SHAP integration with a weighted classifier presents a novel approach to personalized feature recommendation. This combination allows for the prioritization of features based on their impact on the current patient's prediction. While SHAP has been previously employed to measure feature importance, its integration within the framework of a weighted classifier for personalized recommendation distinguishes iCARE as a novel and impactful approach to healthcare decision support systems.[30]

4.3. Preliminary Examination of Dataset

As shown in the experiment, not every dataset requires personalization. The heart failure dataset in our experiment does not benefit from our iCARE framework. We used two procedures to determine whether a dataset is suitable for personalization. The fastest dataset analysis method is to use SHAP value analysis on a ceiling model. If the analysis reveals multiple important features that contribute to the model's predictions, it suggests that the dataset may benefit from personalization. While the presence of multiple important features increases the likelihood of benefiting from personalization, it does not guarantee it. Another approach involves leveraging a pool of known cases to cross-validate the performance of the personalized model, similar to how we test our framework. While this method is slower compared to SHAP value analysis, it directly assesses the performance of personalized models through statistical testing on performance metrics accuracy and AUC values. This method confirms whether personalization is beneficial and allows us to predict how much performance gain can be expected from personalization.

4.4. Limitations and Future Directions

The iCARE framework has several limitations. First, it currently lacks a mechanism to determine whether a dataset warrants personalized feature recommendation automatically. This reliance on a naive dataset evaluation approach necessitates multiple experimental iterations, which may not be feasible in all scenarios. Future work could focus on developing robust criteria or indicators to assess the need for personalization more efficiently. Second, iCARE involves training a locally weighted model for every incoming patient, which may not be suitable for machine learning models requiring extensive training time or scenarios necessitating numerous rapid inferences. iCARE also requires a pool of known cases with a complete set of features to provide individualized feature recommendations. Lastly, iCARE assumes that the initial features available are informative of the predictive space for potential additional features, an assumption that may not always hold true. Future research should explore methods to comprehensively assess the informativeness of initial features to enhance the framework's effectiveness. Addressing these limitations and advancing research in these directions could further enhance the capabilities and applicability of the iCARE framework, ultimately contributing to improved personalized clinical assessments and decision-making in healthcare settings.

5. Conclusion

The iCARE system addresses the challenge of personalized feature selection in clinical assessments by dynamically tailoring the selection of clinical tests based on each patient's unique characteristics. The framework excels over a global feature selection framework in predictive accuracy, especially in cases where the initial features are informative of the predictiveness of the added features. Although personalization might not be needed in all cases, iCARE provides a flexible framework that can be applied using other machine learning algorithms. We believe that with further testing, this general framework can be applied in various fields, extending its utility beyond clinical assessments.

Declaration of generative AI use:

During the preparation of this work the author used ChatGPT in order to improve the writing for clarity and proofreading. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content in the publication.

References

1 Krzyszczyk P, Acevedo A, Davidoff EJ, *et al.* The growing role of precision and personalized medicine for cancer treatment. *Technology* 2018; **06**: 79–100.
2 N P, MB D, T P. BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. In: Adam MP, Feldman J, Mirzaa GM, *et al.*, eds. GeneReviews®. Seattle (WA): University of Washington, Seattle, 1993. http://www.ncbi.nlm.nih.gov/books/NBK1116/ (accessed July 13, 2024).
3 Fernandes JB, Teixeira F, Godinho C. Personalized Care and Treatment Compliance in

Chronic Conditions. *JPM* 2022; **12**: 737.
4 Beydoun MA, Weiss J, Beydoun HA, *et al.* Race, APOE genotypes, and cognitive decline among middle-aged urban adults. *Alz Res Therapy* 2021; **13**: 120.
5 Rajan KB, McAninch EA, Wilson RS, Weuve J, Barnes LL, Evans DA. Race, APOE ε4, and Long-Term Cognitive Trajectories in a Biracial Population Sample. *JAD* 2019; **72**: 45–53.
6 Powell DS, Kuo P-L, Qureshi R, *et al.* The Relationship of APOE ε4, Race, and Sex on the Age of Onset and Risk of Dementia. *Front Neurol* 2021; **12**: 735036.
7 Goetz LH, Schork NJ. Personalized medicine: motivation, challenges, and progress. *Fertil Steril* 2018; **109**: 952–63.
8 Miao J, Niu L. A Survey on Feature Selection. *Procedia Computer Science* 2016; **91**: 919–26.
9 Ying X. An Overview of Overfitting and its Solutions. *J Phys: Conf Ser* 2019; **1168**: 022022.
10 Muni DP, Pal NR, Das J. Genetic programming for simultaneous feature selection and classifier design. *IEEE Trans Syst, Man, Cybern B* 2006; **36**: 106–17.
11 Kudo M, Sklansky J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition* 2000; **33**: 25–41.
12 Saurabh Pal RA. Elimination and Backward Selection of Features (P-Value Technique) In Prediction of Heart Disease by Using Machine Learning Algorithms. *TURCOMAT* 2021; **12**: 2650–65.
13 Maulidina F, Rustam Z, Hartini S, Wibowo VVP, Wirasati I, Sadewo W. Feature optimization using Backward Elimination and Support Vector Machines (SVM) algorithm for diabetes classification. *J Phys: Conf Ser* 2021; **1821**: 012006.
14 Drescher CW, Shah C, Thorpe J, *et al.* Longitudinal Screening Algorithm That Incorporates Change Over Time in CA125 Levels Identifies Ovarian Cancer Earlier Than a Single-Threshold Rule. *JCO* 2013; **31**: 387–92.
15 Peng G, Nourani M, Harvey J, Dave H. Personalized Feature Selection for Wearable EEG Monitoring Platform. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). Cincinnati, OH, USA: IEEE, 2020: 380–6.
16 Beebe-Wang N, Qiu W, Lee S-I. Explanation-guided dynamic feature selection for medical risk prediction. 2023. https://openreview.net/forum?id=1itfhff53V.
17 Atkeson CG, Moore AW, Schaal S. Locally Weighted Learning. *Artificial Intelligence Review* 1997; **11**: 11–73.
18 Chi C-L, Street WN, Katz DA. A decision support system for cost-effective diagnosis. *Artif Intell Med* 2010; **50**: 149–61.
19 Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.
20 Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017; published online Nov 24. http://arxiv.org/abs/1705.07874 (accessed July 13, 2024).
21 Early Stage Diabetes Risk Prediction. 2020. DOI:10.24432/C5VG8H.
22 Heart Failure Clinical Records. 2020. DOI:10.24432/C5Z89R.
23 Lapp, David. Heart Disease Dataset. 2019. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?resource=download.
24 Adnan N, Najnin T, Ruan J. A Robust Personalized Classification Method for Breast Cancer Metastasis Prediction. *Cancers* 2022; **14**: 5327.
25 Escudero J, Ifeachor E, Zajicek JP, Green C, Shearer J, Pearson S. Machine Learning-Based Method for Personalized and Cost-Effective Detection of Alzheimer's Disease. *IEEE Trans Biomed Eng* 2013; **60**: 164–8.
26 Fournier A. Diagnosing diabetes: A practitioner's plea: Keep it simple. *J Gen Intern Med* 2000; **15**: 603–4.
27 Gubbi S, Hannah-Shmouni F, Koch CA, Verbalis JG. Diagnostic Testing for Diabetes Insipidus. In: Feingold KR, Anawalt B, Blackman MR, *et al.*, eds. Endotext. South Dartmouth (MA): MDText.com, Inc., 2000. http://www.ncbi.nlm.nih.gov/books/NBK537591/ (accessed

July 13, 2024).

28 Kang Q, Shi L, Zhou M, Wang X, Wu Q, Wei Z. A Distance-Based Weighted Undersampling Scheme for Support Vector Machines and its Application to Imbalanced Classification. *IEEE Trans Neural Netw Learning Syst* 2018; **29**: 4152–65.

29 Y Z, M K, X Z, J Y, D K, G W. Personalized Diagnosis for Alzheimer's Disease. In: Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017. Springer International Publishing, 2017: 205–13.

30 Fryer D, Strümke I, Nguyen H. Shapley values for feature selection: The good, the bad, and the axioms. 2021; published online Feb 22. http://arxiv.org/abs/2102.10936 (accessed July 13, 2024).

31 Li J, Wu L, Dani H, Liu H. Unsupervised Personalized Feature Selection. *AAAI* 2018; **32**. DOI:10.1609/aaai.v32i1.11628.