Journal of Healthcare Informatics Research (2026) 10:70–94
https://doi.org/10.1007/s41666-025-00222-8

RESEARCH ARTICLE

Explanation Beyond Individual Features: Instance-wise Feature Grouping for EHR Predictive Analytics

Chin Wang Cheong[1] · Kejing Yin[1] · William K. Cheung[1] · Ivor Tsang[2]

Received: 15 July 2024 / Revised: 17 September 2025 / Accepted: 20 October 2025 /
Published online: 6 November 2025
© The Author(s) 2025

Abstract

Identifying relevant input features which contribute to the output of a clinical prediction model can enhance the model explainability. To allow the explainability to be more personalized, instance-wise feature selection (IWFS) methods can be adopted where features are selected specifically for each input instance. Existing IWFS methods often grapple with feature selection instability, and thus precarious interpretation. As relevant features among the instances in a dataset do overlap, feature grouping tricks have been proposed to regularize the selection, but often at the expense of sacrificing the downstream prediction accuracy. To this end, we propose a novel instance-wise feature grouping method called FlexGPC to achieve robust and stable selection by learning i) flexible representation for feature groups, and ii) flexible combination of feature groups implemented using neural networks. To evaluate the effectiveness of FlexGPC, we explore various feature group combination schemes and conduct extensive experiments for performance comparison using real-world electronic health records (EHR) data. Our experimental results show that FlexGPC outperforms all the SOTA baselines in terms of accuracy and feature selection stability for both downstream mortality and next-admission diagnosis prediction tasks. We also illustrate that computational phenotyping can be achieved at the same time, with the identified feature groups being the potential phenotypes.

Keywords Explanability · Feature selection · Deep learning · Electronic health records · Predictive analytics

1 Introduction

Machine learning (ML) methods have been found promising for clinical prediction tasks. Other than achieving high accuracy, their explanability is always emphasized due to the safety-critical nature of healthcare application and the increasing legal

Extended author information available on the last page of the article

and ethical concerns of AI models [1–3]. Identifying relevant input features which contribute to the prediction outcome can allow clinicians to discover the key factors which lead to the outcome based on the ML model. This is essentially a *feature selection* problem which has been well studied in the literature [4, 5], and widely used for explaining relevant risk factors for different diseases [6–10]. For example, [9] adopted a bagging-based feature selection framework and identified smoking habits, lack of exercise, and unbalanced diet of both mothers and children to be the key risk factors of childhood obesity.

Conventional feature selection methods can only identify a common feature subset for all data instances. Yet there are many cases where the relevant feature subset varies over the instances in the dataset. For example, the relevant subset of clinical features for medical diagnosis should depend on the specific health condition of an individual patient. To support more personalized explanability, *instance-wise feature selection* (IWFS) methods can be adopted to identify the relevant feature subset per data instance [11, 12]. As the selected feature subset is considered relevant to the specific output of the downstream task, the IWFS model can be seen as an "explainer" that explains the prediction results through the selected features.

Despite the promising results, existing IWFS methods often grapple with feature selection instability, and thus precarious interpretation. It is always desirable for a feature selection method to select a consistent subset of features given reasonable variation of an input [13, 14]. Achieving that for IWFS, however, is challenging since a lot more selection variables are to be estimated. Neural network models (e.g., MLP) have been proposed to represent the selection mapping [11, 15–17]. Yet, learning the selection network often suffers from overfitting.

With the observation that relevant features for individual data instances in fact possess different degrees of overlapping, feature grouping tricks can be explored to regularize the feature selection to improve the selection stability. *Instance-wise feature grouping* (IWFG) methods have recently been proposed [16, 18] to carry out the instance-wise selection over a set of feature groups instead of individual features. In addition, the feature groups obtained via learning also form interpretable feature grouping patterns (e.g., phenotypes from EHR data), which have also been understood to be salient for the prediction task. Introducing the grouping for the gain in explanability and stability, however, often results in degradation of the downstream prediction accuracy due to the constraints imposed for the regularization.

To this end, we propose Flexible Group-wise Combination (FlexGPC) which aims to explain the prediction output by learning i) a set of more flexibly represented feature groups and ii) a selection network to achieve flexible combination (selection) of the feature groups. Compared with the existing IWFG methods, FlexGPC tries to "soften" the feature group representation can allow the relative importance of each feature within a feature group to be captured, and to "soften" their combination to open up more options for combining the feature groups to form the feature mask. The learning objective of FlexGPC is also designed to further emphasize the model interpretabilty. In this paper, two particular feature group combination schemes, namely *convex combination* and *restricted affine combination* are adopted and integrated into the proposed FlexGPC. The former is a natural choice while the latter enables a more flexible combination by allowing some features to be "de-selected". Figure 1 illus-

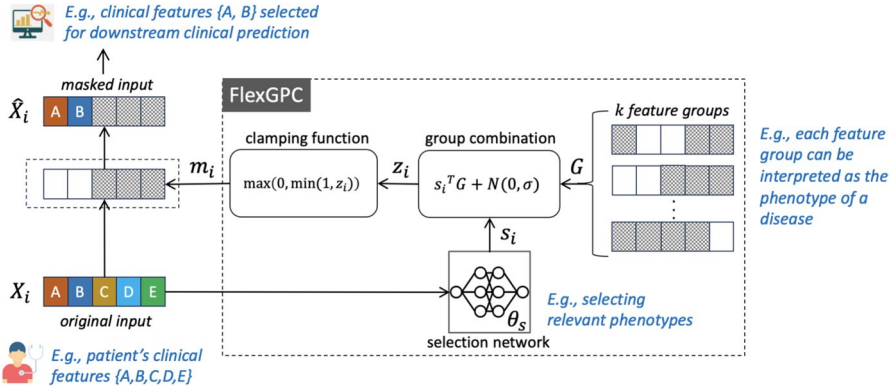


Fig. 1 FlexGPC achieves instance-wise feature selection via adaptive feature group combination for enhancing clinical prediction model explanability. G, are the feature groups, FlexGPC combine them for each data point by the mixture weight s

trates the overall architecture of FlexGPC which can be incorporated into different model architectures for various clinical prediction tasks.

We conduct extensive experiments to evaluate the effectiveness of the proposed FlexGPC for mortality and next-admission diagnosis prediction tasks using the real-world EHR datasets. By properly relaxing the assumptions commonly used in existing IWFG methods, we demonstrate that FlexGPC can achieve substantial improvement in both prediction accuracy and selection stability over the SOTA baselines we tested. The improvement becomes more obvious when the data missingness rate is high which is common for EHR data. Among the two feature group combination schemes, restricted affine composition gives substantially better performance. We provide interpretation to the feature groups identified from the real-world EHR datasets. To illustrate its applicability to other problem domains, we also apply FlexGPC to image data and gene expression data to demonstrate its effectiveness. The key contributions of this paper can be summarized as follows:

1. We propose a novel IWFG model called FlexGPC that enables more flexible feature representation and grouping to achieve robust, stable and more explanable instant-wise feature selection.
2. We design the selection network to implement convex and restricted affine combinations for feature grouping in FlexGPC, and the corresponding learning algorithm to balance the objectives of flexibility of feature grouping and model interpretability.
3. We show that FlexGPC can be incorporated into different clinical analytics models for mortality and next-admission diagnosis prediction, with significant performance improvement over the SOTA IWFS methods in terms of both prediction accuracy and feature selection stability.

2 Related Work

This section provides a brief review of methods proposed for instance-wise feature selection and grouping. With the objective of enhancing model explanability, a number of them were evaluated based on the EHR data analytics tasks.

2.1 Instance-wise Feature Selection

Conventional feature selection aims at identifying a (*global*) relevant subset of features for the whole dataset. The problem has been well studied under different settings for the selection task, including supervised, semi-supervised, and unsupervised [19, 20]. *IWFS* tries to identify a distinct feature subset for each data instance. The selector-predictor approach is commonly adopted, in which a selector network is learned to map each input to a specific feature selection mask and the masked input is fed to a predictor network. For instance, L2X [11] performs instance-wise feature selection for explaining black-box models by maximizing the mutual information between the selected features and the response variable. INVASE [12] extends L2X so that the size of the feature subset selected can also be inferred based on the input instance. As the use of mutual information cannot capture causal influence [15], relative entropy distance together with sparse and class-discriminative features was adopted in [17]. Also, LLSPIN [21] was proposed for selecting features in low-sample-size data.

2.2 Instance-wise Feature Grouping

Grouping highly correlated features is effective in reducing the feature selection solution space [22]. Existing methods for global feature grouping [23, 24] identify feature groups using clustering algorithms, and representative features per cluster can then be selected [25, 26]. This feature grouping idea has also been extended to the instance-wise setting. Instance-Wise Feature Grouping (IWFG) aims to identify feature groups from the data that can be specifically combined according to the input instance to form the feature mask. gI is an IWFG method that formulates the problem using two notions of feature redundancies based on information theory [16]. Additionally, GroupFS is a group-wise feature selection method proposed for supervised tasks [24], which first clusters the data instances and then identifies the cluster-specific feature subset.

2.3 Feature Acquisition

Feature acquisition involves sequentially selecting a subset of features to achieve optimal prediction performance. [27] and [28] utilize Q-learning to sequentially select features. [29] propose a generative surrogate model that captures dependencies among input features to evaluate the potential information gained from the acquisitions. [30] propose an amortized optimization approach as an alternative to the reinforcement learning method, which is notoriously difficult to train. [31] select features in batches rather than individually to reduce query costs. In contrast to feature

acquisition, instance-wise feature selection assumes all features are available upfront, allowing for improved accuracy.

To better situate FlexGPC within the literature, we provide Table 1 that summarize therepresentative instance-wise feature selection (IWFS) and instance-wise feature grouping (IWFG) approaches.

3 Problem Formulation

3.1 Key Challenges

While instance-wise feature selection (IWFS) enables more personalized and interpretable predictions, several fundamental challenges remain:

1. **Stability under perturbations.** IWFS models are highly sensitive to small variations in the input, which can lead to inconsistent feature subsets being selected for nearly identical instances. This instability reduces the reliability of the explanations in safety-critical domains such as healthcare.
2. **Accuracy–interpretability trade-off.** Introducing grouping or sparsity constraints improves interpretability but often comes at the cost of predictive accuracy. Balancing these objectives remains a central challenge in designing explainable selection models.
3. **Flexibility limits of convex mixing.** Existing IWFG methods that rely on convex combination of groups can only span a limited subset of possible feature masks. This lack of expressiveness prevents them from capturing more nuanced or counter-pattern structures that are needed in clinical settings.
4. **Robustness to missingness.** Real-world EHR datasets are plagued by high rates of missing values. A practical IWFS method must generate stable and meaningful masks even when large portions of the feature space are absent.

We denote $X \in \mathbb{R}^{n \times d}$ as the input data (e.g., clinical features) and $Y \in \mathbb{R}^n$ as the prediction labels (e.g., clinical outcomes), where n is the number of data instances (e.g., patient records) and d is the number of features. IWFS aims at learning a specific feature selection mask $m_i \in \mathbb{R}^d$ for each i^{th} data instance $X_i \in \mathbb{R}^d$ where $0 \leq m_{ij} \leq 1$. The masked input $\hat{X}_i$ is obtained by taking the element-wise product $X_i \odot m_i$. The expected outcome of IWFS is to obtain m_i for each X_i so that the masked input $\hat{X}_i$

Table 1 Comparison of representative IWFS/IWFG methods

Method	Selector type	Grouping	Interpretability	Weaknesses
L2X	MLP	No	Salient features	Unstable, no grouping
INVASE	Actor–critic policy net	No	Adaptive subset size	High variance, unstable
LSPIN	Locally sparse NN	No	Sparse local masks	Misses global structure
gI	Info-theoretic redundancy	Yes	Learns groups	High complexity
GroupFS	MoE with discrete gating	Yes	Cluster-level subsets	Rigid, low flexibility

FlexGPC is designed to address their limitations by enhancing expressiveness (via RAC), stability (via regularization and clamping), and robustness under missingness

will correspond to the *relevant* features which in principle can give better prediction result than the case when the original input X_i is used.

In this paper, we propose a novel IWFG method called FlexGPC which comprises a feature groups matrix and a selection network for estimating the values of the feature selection variables, as shown in Fig. 1. In the context of clinical prediction, the feature groups can be interpreted as phenotypes of different disorders. The role of the selection network is to select and combine the relevant phenotypes, and the masked input is the resulting set of relevant clinical features for the downstream prediction. Compared to the existing IWFG methods, FlexGPC allows soft feature groups and relax the assumptions typically imposed on selection variables (e.g., m-hot representation) for the feature grouping. In particular, we study two combination schemes, namely convex combination and restricted affine combination. A selection network can be designed to estimate the selection variables accordingly.

The proposed FlexGPC module can be integrated with a prediction model, and the overall model can be learned end-to-end (to be detailed in the sequel). We design our learning objective so that the masked input contains both informative and relevant features. Also, regularization terms are adopted to encourage sparsity for both the feature mask and the feature groups to enhance the interpretability of FlexGPC.

3.2 Overall Framework of FlexGPC

We first describe two key components of FlexGPC: i) feature groups and ii) feature group selection network for implementing different feature group combination schemes.

3.2.1 Feature Groups

Let $G \in \mathbb{R}^{k \times d}$ represent a set of feature groups, where k is the number of groups and d is the number of features. The elements of G can take values in the range of $0 \leq G_{ij} \leq 1$. $G_j \in \mathbb{R}^d$ is a vector representing the j^{th} feature group with its elements indicating the feature importance within the group.

3.2.2 Feature Group Combination

Let $s_i \in \mathbb{R}^k$ be a vector of selection variables indicating the extent that the k different feature groups should be selected for the data instance X_i. The value of s_i is computed by feeding X_i into a selection network which is to be learned. Instead of assuming m-hot representation for s_i as in [16], we allow s_{ij} to take continuous values. The "selection" step essentially becomes an input-specific combination process. In the sequel, we will call the selection variables as *selection weights*.

In particular, we investigate two combination schemes for representing a feature mask, namely *convex* combination, and *restricted affine* combination.

Definition 1 Convex Combination (CC) aggregates a set of feature groups $G_1, G_2, \ldots, G_k$ for an input X_i with the corresponding selection weights $s_{i1}, s_{i2}, \ldots, s_{ik}$ where $\sum_j s_{ij} = 1$ and $0 \leq s_{ij} \leq 1$ for all $j = 1, ...k$.

The value of s_i is computed by feeding X_i to a selection network designed accordingly. Instead of using multi-hot representation, we allow $0 \leq s_{ij} \leq 1$ for $j = 1, ...k$. Thus, the selection network is essentially performing an input-specific *convex* combination process. We can also explore other ways of combination.

Definition 2 Restricted Affine Combination (RAC) aggregates a set of feature groups $G_1, G_2, \ldots, G_k$ for an input X_i via a weighted sum with the corresponding weights $s_{i1}, s_{i2}, \ldots, s_{ik}$, where $\sum_j |s_{ij}| = 1$ and $-1 \leq s_{ij} \leq 1$ for all $j = 1, ...k$.

RAC deliberately allows the selection weights computed by the selection network to take negative values. This means that we allow one feature group to *de-select* its associated features which are selected due to other feature groups.

Figure 2 illustrates the potential benefit of introducing feature de-selection in the combination process. If only positive values are allowed for the selection weights, we need four feature groups to represent the five hypothetical masked inputs. If negative selection weights are allowed, only three feature groups are needed. In general, given a fixed number of feature groups, RAC can allow more feature masks to be represented compared to CC. It is not difficult to theoretically show that RAC can span a larger subspace of feature masks than CC.

Theorem 1 (Expressiveness of Restricted Affine Composition). Suppose $d > 1$, $k_1 < d$, for any space spanned by a matrix $G_1 \in R^{k_1 \times d}$ with convex composition,

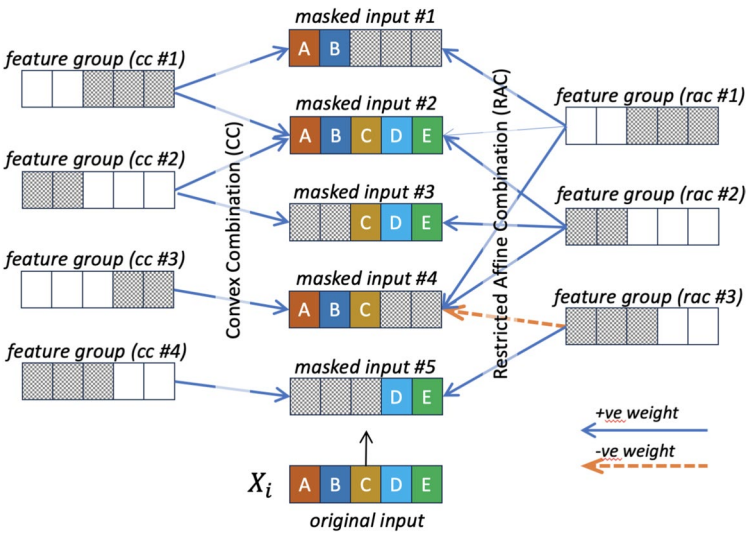


Fig. 2 Illustration of masked inputs which require 4 feature groups based on CC (left) but only 3 groups using RAC by allowing negative selection weight (right)

that exist $G_2 \in R^{k_2 \times d}$ span the larger or equal space with restricted affine composition, and $k_2 < k_1$, if the following hold true:

Condition: There is a feature group separation of G_1 to $G_{1,1:k_2}$ and $G_{1,k_2:k_1}$ such that $G_{1,k_2:k_1}$ are pairwise differences of $G_{1,1:k_2}$.

Proof The space spanned by G_1 is a subspace of $\mathbb{R}^d$, we will denote it by S_{G_1}. By the feature group separation condition, we can write any vector v in S_{G_1} as a convex combination of the vectors in $G_{1,1:k_2}$ and the pairwise differences in $G_{1,k_2:k_1}$. We also denote $G_2 = G_{1,1:k_2}$, and we denote the space spanned by G_2 by S_{G_2}.

We can express any vector $\mathbf{m_i}$ in the space spanned by $G1, k_2 : k_1$, as a restricted affine combination of the vectors in G_2. This is due to the condition that $G_{1,k_2:k_1}$ are pairwise differences of $G_{1,1:k_2}$, and we know that restricted affine composition allows for negative weights, which can express differences between vectors.

Furthermore, any vector in $G_{1,1:k_2}$ can also be expressed as a restricted affine combination of vectors in G_2 as they are the same, by choosing the weights such that they sum up to one and no weight is less than zero, effectively creating a convex combination.

So, any vector in the space spanned by G_1 can be expressed as a restricted affine combination of the vectors in G_2. Therefore, $S_{G_1} \subseteq S_{G_2}$. □

Feature Mask *m* For each visit *i*, we compute the feature mask using restricted affine composition $m_i = U(s_1 \mathrm{g}_1 + s_2 \mathrm{g}_2 + \ldots + s_n \mathrm{g}_n)$. Where g_i is the ith feature group of *G* and $0 \leq m_i \leq 1$. And *U*(*w*) is a clamping function that map *w* from 0 to 1:$U(w) = max(0, min(1, w))$.

We could define the *clamped restricted affine composition*(CRAC) as *U*(*s*) where *w* is defined in the definition 2. Similarly, we define *clamped convex composition* as *U*(*s*) where *s* is defined in the definition 1.

Clamped restricted affine composition is more expressive than restricted affine Composition when $0 \leq G \leq 1$. As shown below:

Theorem 2 (Expressiveness of Clamped Restricted Affine Composition). we denote the space spanned by *G* with restricted affine composition as S_G, and the space spanned by *G* with clamped restricted affine composition as $\hat{S_G}$. And $G >= 0$. $S_G \subset \hat{S_G}$ if some feature groups overlap.

Where we define two feature groups g_1 and g_2 are overlap If $\exists j \in \mathbb{D}, g_{1j}, g_{2j} > 0$ and $\exists j \in \mathbb{D}, g_{1j} \neq g_{2j}$

Proof For any vector *v* in S_G, we can construct it using clamped restricted affine composition with exactly the same set of $s_1, s_2, \ldots, s_n$. as for $\forall j \in \mathbb{D}, 0 \leq v_j \leq 1$. So the clamp function *U*(.) have no effect on it. Therefore, $S_G \subseteq \hat{S_G}$.

Then, we show how to construct v in $\hat{S_G}$ but not in S_G. Suppose g_1 and g_2 are two feature groups that are overlapped; we calculate the difference between them as $g_3 = g_1 - g_2$. g_3 cannot be constructed by restricted affine composition as $\exists j \in \mathbb{D}, g_{3j} < 0$, which violate the positivity of G. With $U(.)$, we can map the negative value to 0, so $g_3 \in \hat{S_G}$. Thus, $S_G \subset \hat{S_G}$

Apart from expressiveness, stability of learned masks is a important criteria that make sure the differences of masks learned during different runs are small. In the following, we will show some definition and properties of the stability and showing stability of the proposed method.

Definition: Uniform Stability A method is said to be β-uniformly stable if, for any two datasets S and S' that differ by one point, and for any input x_i, the computed feature masks m_i and m_i' from S and S' respectively differ by no more than β.

Mathematically, this can be expressed as:

$$\|m_i - m_i'\|_2 \leq \beta$$

for all x_i in the input space, where $m_i = f(S, x_i)$, $m_i' = f(S', x_i)$, f is the method being used (in this case, the feature mask computation using composed feature grouping), and $\|\cdot\|_2$ is the Euclidean norm. The constant β represents the maximum allowed change in the feature mask due to a single change in the dataset.

Assumption 1 Feature Group Learning Stability

Suppose we have two datasets S and S' that differ by one point. Let G and G' be the feature groups learned from S and S' respectively. We assume that the maximum difference between corresponding feature groups in G and G' is bounded by a constant γ.

$$\max_{i \in \{1,\ldots,k\}} \|G_i - G_i'\|_2 \leq \gamma$$

Here, G_i and G_i' represent the i-th feature group in G and G' respectively, $\|\cdot\|_2$ is the Euclidean norm, and γ is a constant representing the maximum change in the feature groups due to a single change in the data. This assumption essentially states that the feature group learning method is stable, in the sense that a small change in the data results in at most a γ-sized change in the feature groups.

Property: Lipschitz Continuity of the Restricted Affine Composition The restricted affine composition of a set of vectors $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n$ with corresponding scalar weights $s_1, s_2, \ldots, s_n$ is a Lipschitz continuous function due to its linearity.

This is mathematically expressed as follows:

$$\|w(\mathrm{v}) - w(\mathrm{v}')\|_2 \leq L\|\mathrm{v} - \mathrm{v}'\|_2$$

for all vectors v and v′, where $w(\mathrm{v}) = s_1\mathrm{v}_1 + s_2\mathrm{v}_2 + \ldots + s_n\mathrm{v}_n$, L is the Lipschitz constant, and $\|\cdot\|_2$ is the Euclidean norm.

Given the constraints on the weights ($|s_1| + |s_2| + \ldots + |s_n| = 1$ and $-1 \leq s_1, s_2, \ldots, s_n \leq 1$), the Lipschitz constant for this operation is 1, meaning that this operation is 1-Lipschitz continuous. Therefore, the distances between points in the input space are not increased by this operation.

Theorem 2 (Uniform Stability of the Method)

Suppose we have two datasets S and S' that differ by one point, and corresponding feature groups G and G' such that $\|G - G'\|_2 \leq \gamma$. Let $f(S, x_i)$ and $f(S', x_i)$ be the feature masks computed for any input x_i under datasets S and S' respectively, where f represents our method. Let the Lipschitz constant of the restricted affine composition used in the method be L.

Then, the method is $L\gamma$-uniformly stable, i.e., for any input x_i, the difference between the computed feature masks is bounded by $L\gamma$:

$$\|f(S, x_i) - f(S', x_i)\|_2 \leq L\gamma$$

The above inequality means that a small change in the input data (at most one data point) leads to a bounded change in the computed feature masks, thus demonstrating the stability of the method.

Proof We start with two datasets S and S', which differ by at most one point. From assumption 1, we have the stability of feature groups learning, which implies that $|G - G'|_2 \leq \gamma$ for corresponding feature groups G and G'.

Now, let's consider an input x_i for which we compute the feature masks $m_i = f(S, x_i)$ and $m_i' = f(S', x_i)$ using datasets S and S', respectively.

Using the Lipschitz continuity of the restricted affine composition (property), the difference in the computed feature masks is bounded by the Lipschitz constant multiplied by the difference in the feature groups, i.e., $|m_i - m_i'|_2 \leq L|G - G'|_2$.

Substituting the upper bound of $|G - G'|_2$ from assumption 1 into this inequality, we have $|m_i - m_i'|_2 \leq L\gamma$.

Thus, we have shown that for any input x_i, the difference between the computed feature masks m_i and m_i' is bounded by $L\gamma$, which completes the proof of $L\gamma$-uniform stability of the method.

3.2.3 Selection Network Implementation

We implement the selection network using MLP. For CC, the selection network, denoted as $\theta_s^{cc}(X_i)$, computes the output $s_i \in \mathbb{R}^k$ using the function:

$$s_i \;=\; \theta_s^{cc}(X_i) \;=\; \text{softmax}\left(W_2\text{ReLu}(W_1X_i)\right) \tag{1}$$

where $W_1 \in \mathbb{R}^{l\times d}$ and $W_2 \in \mathbb{R}^{k\times l}$ are the parameters of the selection network. The softmax function is used to guarantee $0 \le s_{ij} \le 1$ for all j and $\sum_j s_{ij} = 1$.

For RAC, the selection network, denoted as $\theta_s^{rac}(X_i)$, computes the output $s_i \in \mathbb{R}^k$ using the function:

$$s_i \;=\; \theta_s^{rac}(X_i) \;=\; \frac{W_2'\text{ReLu}(W_1'X_i)}{||W_2'\text{ReLu}(W_1'X_i)||_1} \tag{2}$$

where $W_1' \in \mathbb{R}^{l\times d}$ and $W_2' \in \mathbb{R}^{k\times l}$ are the parameters of the selection network $\theta_s^{rac}()$, and $|||\cdot|||_1$ denotes the L1-norm. The use of the formula can guarantee $-1 \le s_{ij} \le 1$ for all j, and $\sum_j |s_{ij}| = 1$.

3.2.4 Feature Mask via Groups Selection

Finally, the feature mask m_i for X_i can be computed as $m_i = s_i^T G$. To achieve a more robust mask, we adopt the LSPIN mapping [21] during the model training:

$$m_i = \max(0, \min(1, s_i^T G + \epsilon)) \tag{3}$$

where ϵ is drawn from $N(0, \sigma)$ during model training, and σ is a hyperparameter for setting the noise level. The mapping has the benefit of pushing the elements in m_i to take values closer to either 0 or 1. Adding noises helps the model explore more combinations of feature groups during training, which can in turn improve the model robustness. After training, the mask can be computed using (3) by setting ϵ to zero.

3.3 Prediction Models with FlexGPC Incorporated

FlexGPC can be incorporated into different prediction models and train them end-to-end to achieve robust and stable instance-wise feature grouping and selection, as well as high prediction accuracy. In particular, we apply FlexGPC to two prediction tasks: mortality prediction and next-admission diagnosis prediction. Depending on whether the sequential relationship of the features are exploited or not, the two tasks can readily be formulated as sequence-to-sequence classification or just standard classification.

For standard classification, as shown in Fig. 3, we can first feed the input feature vector to FlexGPC and then the masked input to an MLP to form *FlexGPC-MLP*. For sequence-to-sequence classification, we can integrate FlexGPC with sequential models like Transformer. Specifically, given an input sequence $\text{X}_i = (X_i^1, X_i^2, ..., X_i^{Li})$ where $X_i^j \in \mathbb{R}^d$ and the corresponding output label sequence $\text{Y}_i = (Y_i^1, Y_i^2, ..., Y_i^{Li})$ where $Y_i^j \in \mathbb{R}$, we can feed each element X_i^j to FlexGPC to obtain its masked version $\hat{\text{X}}_i^j$. Then, we feed the masked input sequence $\hat{\text{X}}_i$ to a Transformer, and then to an MLP to form *FlexGPC-Trans-MLP*.

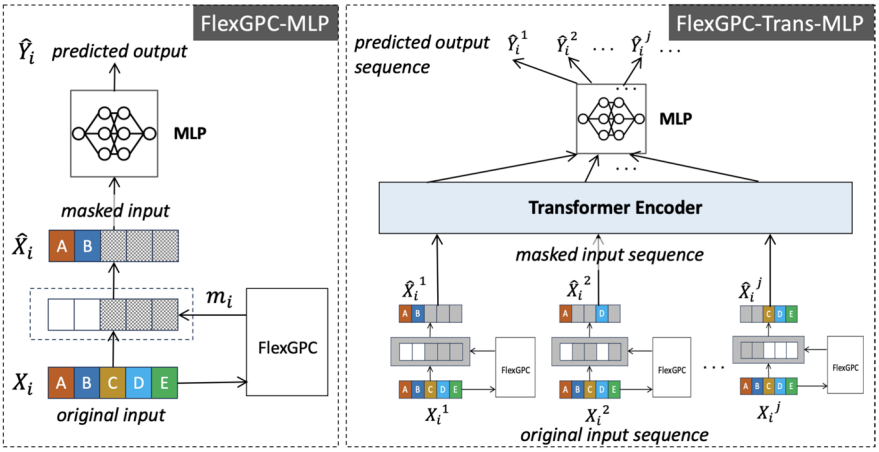


Fig. 3 FlexGPC incorporated into classification and sequence-to-sequence prediction models

3.3.1 Model Training

Both FlexGPC-MLP and FlexGPC-Trans-MLP can be trained end-to-end to achieve good feature selection and prediction performance at the same time. For the training, we adopt the following objective function with three loss terms:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{recons}} + \mathcal{L}_{\text{reg}}. \tag{4}$$

a) Prediction Loss is measured by cross-entropy:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{Y}_i^{\intercal}\log(\hat{\mathrm{Y}}_i) + (1-\mathrm{Y}_i)^{\intercal}\log(1-\hat{\mathrm{Y}}_i)\right]. \tag{5}$$

to guide the model learning to give high prediction accuracy given the *relevant* features selected.

b) Reconstruction Loss measures the discrepancy between X_i and its reconstructed version based on the masked input $\hat{\mathrm{X}}_\mathrm{i}$:

$$\mathcal{L}_{\text{recons}} = \sum_{i=1}^{n} ||X_i - \phi(\hat{X}_i)|| \tag{6}$$

where $\phi()$ is implemented using a two-layer MLP. It encourages the *informative* features that could recover the redundant features to be captured and retained.

c) Regularization Term comprises two parts:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}^{m}_{\text{reg}} + \mathcal{L}^{G}_{\text{reg}} = \sum_{i,j} m_{ij} + I(G) \tag{7}$$

$$I(G) = -\frac{1}{d}\sum_{i=1}^{\text{d}}\sum_{j=1}^{\text{k}} G^{T}_{ij}\log(G^{T}_{ij}) - \frac{1}{k}\sum_{i=1}^{\text{k}}\sum_{j=1}^{\text{d}} G_{ij}\log(G_{ij})$$

where $\mathcal{L}^{m}_{\text{reg}}$ encourages sparse mask, and $\mathcal{L}^{G}_{\text{reg}}$ encourages discriminative and sparse feature groups. Both are introduced for enhancing FlexGPC's interpretability.

Note that before training directly on $\mathcal{L}$, we pre-train the feature masks m based on the following *pre-training loss*:

$$\mathcal{L}_{\text{pretrain}} = \sum_{i,j}(m_{ij} - \text{X}_{i,j})^2. \tag{8}$$

We find that using this loss for pre-training can force the feature masks m to be similar to the input X. It allows the parameters of the selection network to be initialized less randomly before the subsequent model training, thereby further improving the selection stability.

4 Experiment Setup

4.1 Datasets

To evaluate the effectiveness of FlexGPC for clinical prediction tasks, two real-world EHR datasets, namely MIMIC-III (Medical Information Mart for Intensive Care) [32] and eICU [33], are used.

MIMIC-III is a public dataset containing data of over 46,000 patients admitted to intensive care units (ICU). eICU is a multi-center database, containing over 200,000 ICU admissions. We filter out patients with less than 2 hospital admissions to allow next-diagnosis prediction to be carried out. As a result, we extract 6, 453 patients with 2.7 admissions per patient on average for MIMIC-III, and 12, 293 patients with 2.2 admissions on average for eICU. The average numbers of diagnoses and medications per admission are 12.0 and 38.9 respectively for MIMIC-III, while 14.6 and 21.4 for eICU.

To further demonstrate that FlexGPC is also applicable to other problem domains, we conduct additional experiments using a gene expression dataset and the MNIST image dataset. For the gene expression dataset, we use one that contains cells collected from the human pancreas, with 1,937 cells, 20,125 genes, and 14 cell types.[1] We follow the pre-processing adopted in [34], where genes expressed in less than three cells are excluded from further analysis and the gene expression counts per cell are normalized. MNIST is a database of handwritten digits, which contains 60,000

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2230757

training images. Each image is a 28x28 pixel grayscale image of a handwritten digit (0 through 9). And there are 10,000 test images.

The statistics of the datasets are summarized in Table 2.

4.2 Parameter Setting

For implementing $\theta_s^{cc}()$, $\theta_s^{rac}()$ and $\phi()$, the size of the hidden layer l is 400. For the prediction MLP in FlexGPC-MLP and FlexGPC-Trans-MLP, the size of the hidden layer is 200. The Transformer used in FlexGPC-Trans-MLP has one layer, a single head and a dimension of 200. We tested different numbers of feature groups k from 50 to 300. For model training, σ in (3) is set to 1. The batch size is 100. 80% of the data is used for training, 10% for validation, and 10% for testing. We run our experiments on a server with four NVidia Tesla V100-PCIE-32GB GPU, 250 GB memory, and Intel(R) Silver 4114 CPU. Adam optimizer is used for the training with five repetitions.

4.3 Performance Evaluation

We test the performance of FlexGPC first on two clinical prediction tasks: i) next-admission diagnosis prediction [35], and ii) mortality prediction [36]. They are formulated as sequence-to-sequence classification problems with FlexGPC-Trans-MLP adopted. We further test the effectiveness of FlexFPC on two other problems iii) cell type identification [34], and iv) handwritten digit recognition. They are formulated as standard classification problems where FlexGPC-MLP is adopted.

4.3.1 Metrics for Prediction Accuracy

For the next-admission diagnosis prediction [37, 38], we first derive the ground-truth labels by grouping the diagnoses in the next admissions into 793 groups based on the first three digits of their ICD-9 codes, and then carry out multi-label classification accordingly. The prediction accuracy is measured by:

$$Accuracy@k = \frac{\#\text{ of true positives in the top } k \text{ predictions}}{\#\text{ of positives}}.$$

For mortality prediction, we use *area under the ROC curve (AUC)*. For cell type identification and handwritten digita recognition, we use *Accuracy@1*.

4.3.2 Metric for Feature Selection Stability

We evaluate the stability of feature selection by:

Table 2 Statistics of datasets

Data set	MIMIC-III	eICU	Gene	MNIST
# of Samples	6,453	12,293	1,937	60,000
# of Features	6,054	3,353	20,125	784
Average visits per patients	2.7	2.2	/	/

$$Stability(i) = \frac{\sum_{p,q \in B} RankSim(m_i^p, m_i^q)}{B^2}$$

It measures the similarity of the feature masks learned in B runs (5 in our experiments), where

$$RankSim(m_i^p, m_i^q) = \frac{\sum_{i=1}^{d} \mathbb{1}\{\text{index}_i(m_i^p) = \text{index}_i(m_i^q)\}}{d}$$

gives the ranking similarity of the two masks m_i^p and m_i^q obtained in p^{th} and q^{th} runs respectively. The higher the value, the better the stability is.

4.4 Baselines

We compare the performance of FlexGPC-Trans-MLP and FlexGPC-MLP with a number of baselines.

- *INVASE* [12] consists of a selector network, a predictor network, and a baseline network and uses the actor-critic methodology for training.
- *LSPIN* [21] is a locally sparse neural network where the local sparsity is learned to identify the relevant feature subset for each data instance.
- *gI* [16] learns and combine feature groups to form the feature mask by minimizing the loss of representation and relevant redundancies.
- *GroupFS* [24] is a group-wise feature selection method that uses the Mixture of Experts (MoE) model with discrete gating to select feature masks.
- FlexGPC$_{CC}$ replaces the restricted affine combination (RAC) component of FlexGPC with the convex combination (CC).

For fair comparison, we modify all the baselines so that they adopt the same prediction network as in FlexGPC-Trans-MLP or FlexGPC-MLP, depending on the prediction task.

For the two sequence-to-sequence prediction tasks, we report also the performance of *Hi-BEHRT* [39] which adopts a more sophisticated hierarchical Transformer to capture information of long sequences in the EHR.

5 Results

This section presents the results of performance comparison with the baselines based on the four prediction tasks.

5.1 Performance on EHR Data

The performance comparison results on next-admission diagnosis prediction and mortality prediction are summarized in Table 3 (MIMIC-III) and Table 4 (eICU). We observe that the proposed FlexGPC generally outperforms baselines under different

Table 3 Performance comparison based on MIMIC-III

missing rate			
Model	**0%**	**20%**	**40%**
Next-admission diagnosis prediction (Accuracy@20)			
INVASE	0.571 ± 0.015	0.609 ± 0.019	0.589 ± 0.017
LSPIN	0.615± 0.022	0.600 ± 0.023	0.598 ± 0.025
gI	0.602 ± 0.014	0.600 ± 0.017	0.591 ± 0.018
GroupFS	0.592 ± 0.012	0.589 ± 0.012	0.586 ± 0.014
FlexGPC$_{CC}$	0.610 ± 0.019	0.592 ± 0.019	0.586 ± 0.018
FlexGPC$_{Avg}$	0.589 ± 0.011	0.576 ± 0.016	0.571 ± 0.012
FlexGPC$_{MLP}$	0.575 ± 0.09	0.572 ± 0.011	0.561 ± 0.015
FlexGPC	**0.646± 0.016**	**0.621 ± 0.015**	**0.614 ± 0.017**
Hi-BEHRT	0.631 ± 0.013	0.618± 0.012	0.610 ± 0.011
Mortality prediction (AUC)			
INVASE	0.872 ± 0.021	0.839 ± 0.019	0.812 ± 0.020
LSPIN	0.944 ± 0.010	**0.880 ± 0.015**	0.877 ± 0.016
gI	0.917 ± 0.014	0.835 ± 0.014	0.810 ± 0.017
GroupFS	0.845 ± 0.011	0.837 ± 0.013	0.823 ± 0.013
FlexGPC$_{CC}$	0.859 ± 0.008	0.848 ± 0.010	0.836 ± 0.012
FlexGPC$_{Avg}$	0.896 ± 0.015	0.892 ± 0.017	0.881 ± 0.014
FlexGPC$_{MLP}$	0.862 ± 0.017	0.855 ± 0.022	0.853 ± 0.013
FlexGPC	**0.945 ± 0.010**	0.879 ± 0.012	**0.878 ± 0.011**
Hi-BEHRT	0.865 ± 0.021	0.832 ± 0.025	0.824 ± 0.023

FlexGPC-Trans-MLP is abbreviated as FlexGPC

degrees of data missingness. Using RAC for combining the feature groups in most cases outperforms the use of CC. This shows the benefit of RAC to allow the possibility of learning feature groups for "de-selection". Compared with Hi-BEHRT which adopts a more sophisticated Transformer, FlexGPC together with only a simple Transformer obtains better performance. Also, Fig. 4 shows that for next-admission diagnosis prediction, FlexGPC consistently outperforms FlexGPC$_{CC}$ for different number of feature groups being tested. The same results are observed for mortality prediction. In general, FlexGPC which uses the RAC for the feature groups combination requires a smaller number of feature groups to achieve higher accuracy as compared to FlexGPC$_{CC}$.

To further confirm the effectiveness of CC and RAC for feature grouping, we test two variants of FlexGPC for selecting features and combining feature groups. FlexGPC$_{Avg}$ aggregates the feature groups via simple averaging, while FlexGPC$_{MLP}$ learns the feature mask using a one-layer neural network without feature grouping. As shown in Tables 3 and 4, FlexGPC can outperform both by a large margin. For example, for the next-admission diagnosis prediction, the Accuracy@20 on MIMIC-III drops from 0.646 for FlexGPC to 0.589 for FlexGPC$_{Avg}$ and 0.575 for FlexGPC$_{MLP}$.

Regarding feature selection stability, Table 5 shows the stability scores of FlexGPC and FlexGPC$_{CC}$ for next-admission diagnosis prediction. To test also their robustness, we conduct the test with different levels of Gaussian noise added to the input. Both outperform all the baselines. gI is the second best where feature

Table 4 Performance comparison based on eICU

missing rate

Model	**0%**	**20%**	**40%**
Next-admission diagnosis prediction (Accuracy@20)			
INVASE	**0.900± 0.011**	0.878 ± 0.012	0.856 ± 0.015
LSPIN	0.885 ± 0.009	0.876 ± 0.010	0.854 ± 0.010
gI	0.882 ± 0.007	0.879 ± 0.07	0.856 ± 0.09
GroupFS	0.870 ± 0.010	0.862 ± 0.012	0.835 ± 0.011
FlexGPC_{CC}	0.873± 0.007	0.863 ± 0.008	0.845 ± 0.100
FlexGPC_{Avg}	0.869± 0.006	0.872 ± 0.009	0.849 ± 0.8
FlexGPC_{MLP}	0.865± 0.008	0.860 ± 0.007	0.841 ± 0.13
FlexGPC	0.896± 0.009	**0.882± 0.010**	**0.863± 0.012**
Hi-BEHRT	0.872± 0.019	0.867 ± 0.017	0.842 ± 0.018
Mortality prediction (AUC)			
INVASE	0.731 ± 0.015	0.710 ± 0.014	0.703 ± 0.016
LSPIN	0.719 ± 0.012	0.701 ± 0.013	0.695 ± 0.013
gI	0.712 ± 0.015	0.710 ± 0.017	0.687± 0.018
GroupFS	0.638 ± 0.012	0.617 ± 0.012	0.552 ± 0.015
FlexGPC_{CC}	0.736± 0.007	0.731 ± 0.008	0.687± 0.011
FlexGPC_{Avg}	0.738± 0.012	0.725 ± 0.009	0.682± 0.013
FlexGPC_{MLP}	0.722± 0.005	0.724 ± 0.007	0.667± 0.09
FlexGPC	**0.740± 0.006**	**0.740± 0.010**	**0.729± 0.011**
Hi-BEHRT	0.716 ± 0.018	0.695 ± 0.021	0.695 ± 0.019

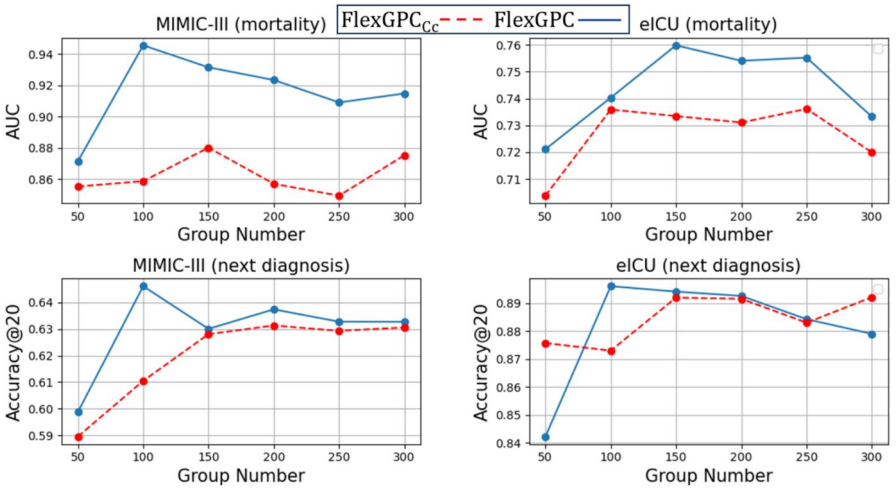


Fig. 4 Effect of the number of feature groups on the accuracy of FlexGPC_{CC} (red dashed line) and FlexGPC (blue line). FlexGPC outperforms FlexGPC_{CC} in most cases

Table 5 Stability comparison for MIMIC-III for next-admission diagnosis prediction

Model/Noise level	0	0.1	0.5
INVASE	0.815	0.801	0.785
LSPIN	0.809	0.792	0.773
gI	0.835	0.817	0.799
GroupFS	0.826	0.811	0.794
FlexGPC$_{CC}$	0.878	0.843	**0.806**
FlexGPC	**0.882**	**0.852**	0.795

grouping is also adopted. Similar improvement is also observed on eICU data, and for mortality prediction.

5.2 Phenotypes Extracted from MIMIC-III

Figure 5 shows three feature groups (phenotypes) (G_1, G_2, G_3) extracted from MIMIC-III using FlexGPC. With reference to a specific patient, G_1 and G_2 are identified by FlexGPC as positive (selected) feature groups, and G_3 as a negative (deselected) group. Figure 5a illustrates how the three feature groups can be combined using RAC estimated by the selection network to give a feature mask shown in Fig. 5b).

The feature mask suggests that the patient is dealing with a serious, potentially metastatic cancer, accompanied by psychological (depression) and metabolic (hyperlipidemia) comorbidities and a high-risk vascular condition (aneurysm). G_1 is a group consisting of mental health issues that are related to the patients (*311*, *V667*) and some neoplasms-related diseases. However, not all diseases in G_1 are in the patient's record. E.g., the patient did not develop *1578* (Malignant neoplasm of other specified sites of pancreas) and *1970* (Secondary malignant neoplasm of lung). The model can de-select them in G_1 by subtracting from it feature group G_3 which contains those diseases. Similarly, G_2 is related to bleeding issues which contain diseases related to the patient (*2724*, *4414*, *311*). Those unrelated diseases (*2851*, *5781*) again can be de-selected by subtracting from it G_3. The capability of identifying the feature groups which can be combined via group selection and de-selection to represent the whole dataset is a key feature enabled by FlexGPC.

5.3 Performance on Gene Expression and MNIST Datasets

To demonstrate the applicability of the proposed FlexGPC to other problem domains, we apply FlexGPC-MLP to the human pancreas gene expression data for cell type identification and to the MNIST dataset for the handwritten digit recognition task. The results are summarized in Table 6.

FlexGPC obtains the highest accuracy for the cell type identification task. INVASE and LSPIN achieve slightly worse accuracy of about 0.94, while gI is significantly worse, obtaining an accuracy of 0.74. We also report the two stability scores in the same table. FlexGPC gives the most stable feature selection performance. Similar conclusions can be drawn from the handwritten digit recognition task, where FlexGPC obtains the highest accuracy and stability.

ICD	Descriptions	Weights
	Feature Group G_1	
311	Depressive disorder, not elsewhere classified	1.0000
V667	Encounter for palliative care	1.0000
1970	Secondary malignant neoplasm of lung	1.0000
1578	Malignant neoplasm of other specified sites of pancreas	0.8937
1977	Malignant neoplasm of liver, secondary	0.6985
2724	Other and unspecified hyperlipidemia	0.5971
	Feature Group G_2	
2851	Acute posthemorrhagic anemia	1.0000
2724	Other and unspecified hyperlipidemia	1.0000
1970	Secondary malignant neoplasm of lung	1.0000
4414	Abdominal aneurysm without mention of rupture	0.9775
311	Depressive disorder, not elsewhere classified	0.8087
5781	Blood in stool	0.4404
	Feature Group G_3	
1977	Malignant neoplasm of liver, secondary	1.0000
2851	Acute posthemorrhagic anemia	1.0000
5781	Blood in stool	0.8074
1970	Secondary malignant neoplasm of lung	0.4372
1578	Malignant neoplasm of other specified sites of pancreas	0.3777

(a) Three phenotypes identified as relevant to a patient by FlexGPC.

$$\underbrace{\begin{bmatrix} 311 & 0.7467 \\ 2724 & 0.5971 \\ V667 & 0.5369 \\ 1977 & 0.3952 \\ 2851 & 0.3489 \\ 4414 & 0.3432 \end{bmatrix}}_{m_1} = 0.5011 * \underbrace{\begin{bmatrix} 311 & 1.000 \\ V667 & 1.000 \\ 1970 & 1.000 \\ 1578 & 0.8937 \\ 1977 & 0.6985 \\ 2724 & 0.5971 \end{bmatrix}}_{G_1} + 0.2676 * \underbrace{\begin{bmatrix} 2851 & 1.000 \\ 2724 & 1.000 \\ 1970 & 1.000 \\ 4414 & 0.9775 \\ 311 & 0.8087 \\ 5781 & 0.4404 \end{bmatrix}}_{G_2} - 0.2011 * \underbrace{\begin{bmatrix} 1977 & 1.000 \\ 2851 & 1.000 \\ 5781 & 0.8074 \\ 1970 & 0.4372 \\ 1578 & 0.3777 \end{bmatrix}}_{G_3}$$

(b) Restricted affine combination (RAC) of G_1, G_2 and G_3 by FlexGPC, where the coefficients $\{+0.5011, +0.2676$ and $-0.2011\}$ are computed based on the patient's clinical features.

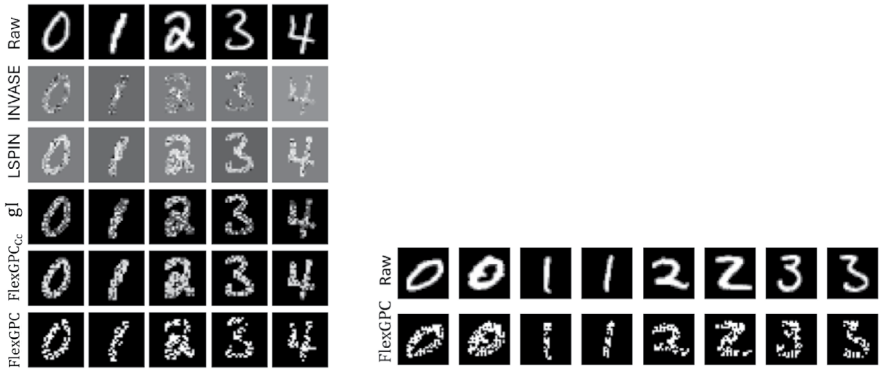
ICD	Descriptions	Weights
311	Depressive disorder, not elsewhere classified	0.7467
2724	Other and unspecified hyperlipidemia	0.5971
V667	Encounter for palliative care	0.5369
1977	Malignant neoplasm of liver, secondary	0.3952
2851	Acute posthemorrhagic anemia	0.3489
4414	Abdominal aneurysm without mention of rupture	0.3432

(c) The resulting feature mask m_1 for the patient.

Fig. 5 Illustration of a feature mask obtained by FlexGPC for a patient record in MIMIC-III

Table 6 Accuracy and stability comparison on human pancreas gene expression and MNIST datasets

	Cell Type Identification		Handwritten Digit Recognition	
Model	Accuracy@1	Stability	Accuracy@1	Stability
INVASE	0.948 ± 0.012	0.831	0.977 ± 0.090	0.810
LSPIN	0.943 ± 0.010	0.829	0.963 ± 0.011	0.833
gI	0.747 ± 0.013	0.861	0.965 ± 0.012	0.857
GroupFS	0.769 ± 0.013	0.881	0.958 ± 0.015	0.823
FlexGPC$_{Cc}$	0.943 ± 0.011	0.872	0.963 ± 0.011	0.866
FlexGPC	**0.959 ± 0.012**	**0.899**	**0.979 ± 0.011**	**0.852**



(a) Feature masks by all methods. (b) Feature masks for digits of different styles.

Fig. 6 Comparison of learned feature masks for digits in MNIST

Figure 6a shows the learned feature masks based on different approaches where we observe that RAC can identify more salient regions in the image for the prediction as compared to the others. Figure 6b shows that the feature masks obtained by FlexGPC can effectively select the important pixels corresponding to the same digits written in different styles.

5.4 Visualization of Groupings of Feature Masks

Figure 7 shows the visualization of the feature masks learned by FlexGPC for the MIMIC-III dataset. To facilitate the visualization, we first apply K-means clustering to the dataset based on the hamming distance of the ground truth next-admission diagnoses to obtain the disease group label for each patient visit. The feature masks for the patient visits under the same disease group label are then presented together row-wise. In the Fig. 7, we see that the masks under the same group learned by FlexGPC share similar clinical features (histories of diagnoses and medications). For the cell type identification, we group together the cells with the same ground truth cell type label. Figure 8 shows a snapshot of feature masks learned from the gene expression data. Again, each row shows a feature mask, and the feature masks are grouped according to their ground truth cell type. We can observe that prominent and

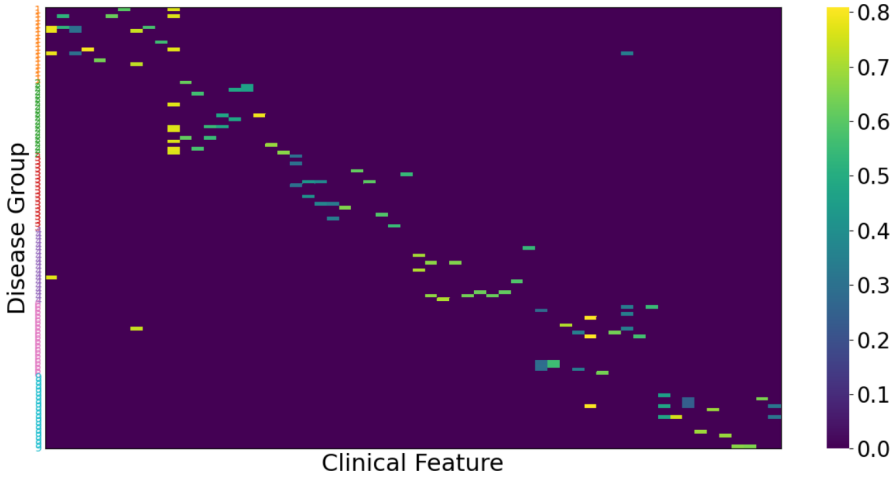


Fig. 7 Feature masks learned from the MIMIC-III data by FlexGPC

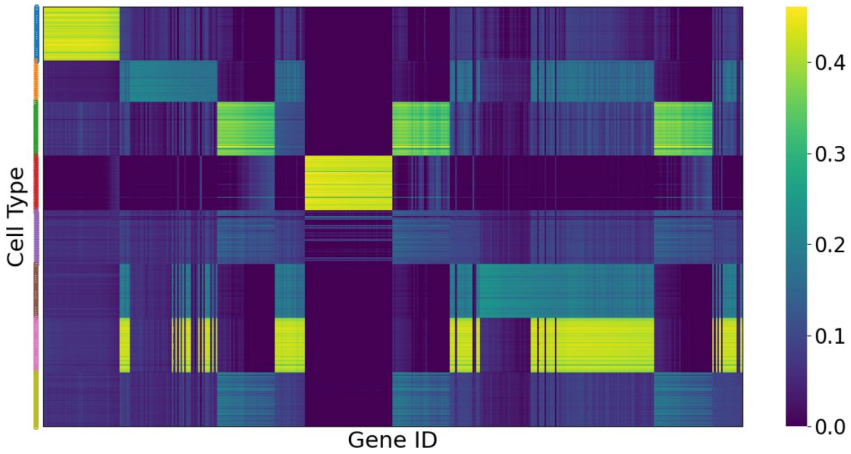


Fig. 8 Feature masks learned from the human pancreas data by FlexGPC

distinct gene blocks can be discovered for each cell type. This implies that FlexGPC can effectively discover the biomarkers for cells of the same type.

6 Conclusion

In this paper, we propose a novel instance-wise feature grouping model called FlexGPC. We incorporated FlexGPC into different prediction models and trained them end-to-end for various analytics tasks based on electronic health records (EHR), gene expression, and image data. Across all tested tasks, FlexGPC enhanced both downstream prediction accuracy and feature selection stability. Additionally, FlexGPC

offers fine-grained interpretation through the use of feature groups and increased expressiveness due to the flexible combination of feature groups. We demonstrated how FlexGPC, with a selection network implementing restricted affine combination, supports the selection and de-selection of feature groups, thereby enhancing the robustness and stability of Instance-Wise Feature Grouping (IWFG), as confirmed by extensive experiments. For future work, we plan to consider temporal information from sequential data to infer the feature mask, allowing for the exploration of dynamic interactions among input features. Additionally, developing more explicit methods to handle missing data is another direction towards achieving more robust instance-wise feature selection.

Author Contributions William K. Cheung and Ivor Tsang supervised the project. Chin Wang Cheong designed the model, implemented the code and conducted the experiments. William K. Cheung, Chin Wang Cheong, Kejing Yin contributed to the drafting of the manuscript. Ivor Tsang comments and reviewed the manuscript. All authors discussed the results and approved the final version before submission.

Funding Open access funding provided by Hong Kong Baptist University Library. This research is partially supported by the Research Matching Grant Scheme RMGS2021_8_06 from the Hong Kong Government, the National Natural Science Foundation of China (NSFC) under Grant 62302413 and the Health and Medical Research Fund (HMRF) under Grant 23220312.

Data Availability The MIMIC-III data set and eICU data set can be downloaded from https://physionet.org/content/mimiciii-demo/1.4/ and https://eicu-crd.mit.edu/gettingstarted/access/ respectively. The gene expression data set is available at the following URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM223075. And the MNIST data set is available at https://www.kaggle.com/datasets/hojjatk/mnist-dataset.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Higgins D, Madai VI (2020) From bit to bedside: A practical framework for artificial intelligence product development in healthcare. Adv Intell Syst 2(10):2000052. https://doi.org/10.1002/aisy.202000052, https://arxiv.org/abs/onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202000052
2. Amann J, Blasimme A, Vayena E, Frey D, Madai V (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Making 20. https://doi.org/10.1186/s12911-020-01332-6

3. Yang CC (2022) Explainable artificial intelligence for predictive modeling in healthcare. J Healthcare Inf Res 6(2):228–239. https://doi.org/10.1007/s41666-022-00114-1
4. Scheurwegs E, Cule B, Luyckx K, Luyten L, Daelemans W (2017) Selecting relevant features from the electronic health record for clinical code prediction. J Biomed Inf 74:92–103. https://doi.org/10.1016/j.jbi.2017.09.004
5. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
6. Pathan MS, Nag A, Pathan MM, Dev S (2022) Analyzing the impact of feature selection on the accuracy of heart disease prediction. Healthcare Anal 2:100060
7. Ebrahimi A, Wiil UK, Naemi A, Mansourvar M, Andersen K, Nielsen AS (2022) Identification of clinical factors related to prediction of alcohol use disorder from electronic health records using feature selection methods. BMC Med Inf Decis Making 22
8. Chen Y, Zhang J, Qin X (2022) Interpretable instance disease prediction based on causal feature selection and effect analysis. BMC Med Inform Decis Mak 22(1):51. https://doi.org/10.1186/s12911-022-01788-8
9. Shi X, Nikolic G, Epelde G, Arrúe M, Bidaurrazaga Van-Dierdonck J, Bilbao R, De Moor B (2021) An ensemble-based feature selection framework to select risk factors of childhood obesity for policy decision making. BMC Med Inf Decis Making 21
10. Bosschieter TM, Xu Z, Lan H, al. (2024) Interpretable predictive models to understand risk factors for maternal and fetal outcomes. J Healthcare Inf Res 8(1):65–87. https://doi.org/10.1007/s41666-023-00151-4
11. Chen J, Song L, Wainwright M, Jordan M (2018) Learning to Explain: An information-theoretic perspective on model interpretation. In: Proceedings of the 35th international conference on machine Learning, pp 883–892
12. Yoon J, Jordon J, Schaar M (2019) Invase: instance-wise variable selection using neural networks. In: Proceedings of international conference on learning representations
13. Xin B, Hu L, Wang Y, Gao W (2015) Stable feature selection from brain smri. In: Proceedings of the AAAI conference on artificial intelligence, pp 1910–1916
14. He Z, Yu W (2010) Stable feature selection for biomarker discovery. Comput Biol Chem 34(4):215–225. https://doi.org/10.1016/j.compbiolchem.2010.07.002
15. Chang S, Zhang Y, Yu M, Jaakkola T (2020) Invariant rationalization. In: Proceedings of the international conference on machine learning, pp 1448–1458
16. Masoomi A, Wu C, Zhao T, Wang Z, Castaldi P, Dy JG (2020) Instance-wise feature grouping. In: Proceedings of neural information processing systems, pp 13374–13386
17. Panda P, Kancheti SS, Balasubramanian V (2021) Instance-wise causal feature selection for model interpretation. In: Proceedings of IEEE conference on computer vision and pattern recognition workshops, pp 1756–1759
18. Chormunge S, Jena S (2018) Correlation based feature selection with clustering for high dimensional data. J Electr Syst Inf Technol 5(3):542–549. https://doi.org/10.1016/j.jesit.2017.06.004
19. Witten DM, Tibshirani R (2010) A framework for feature selection in clustering. J Am Stat Assoc 105(490):713–726
20. Akhiat Y, Asnaoui Y, Chahhou M, Zinedine A (2020) A new graph feature selection approach. In: Proceedings of the 6th IEEE congress on information science and technology, pp 156–161. https://api.semanticscholar.org/CorpusID:233136414
21. Yang J, Lindenbaum O, Kluger Y (2021) Locally sparse neural networks for tabular biomedical data. In: Proceedings of international conference on machine learning, pp 25123–25153
22. Kuzudisli C, Bakir-Gungor B, Bulut N, Qaqish B, Yousef M (2023) Review of feature selection approaches based on grouping of features. PeerJ 11
23. Dai Y, Gao Z, Zhu Y, Zhang W, Li H, Wang Y, Li Z (2022) Feature grouping for no-reference image quality assessment. In: Proceedings of 7th international conference on automation, control and robotics engineering, pp 204–208
24. Xiao Q, Li H, Tian J, Wang Z (2022) Group-wise feature selection for supervised learning. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 3149–3153
25. Sahu B, Dehuri S, Jagadev AK (2017) Feature selection model based on clustering and ranking in pipeline for microarray data. Inf Med Unlocked 9:107–122

26. Alimoussa, M, Porebski, A, Vandenbroucke, N, Thami, ROH, El Fkihi, S (2021) Clustering-based sequential feature selection approach for high dimensional data classification. In: Proceedings of the 16th international joint conference on computer vision, imaging and computer graphics theory and applications, pp 122–132
27. Shim H, Hwang SJ, Yang E (2018) Joint active feature acquisition and classification with variable-size set encoding. In: Proceedings of advances in neural information processing systems, vol. 31
28. Janisch J, Pevný, T, Lisý V (2017) Classification with costly features using deep reinforcement learning. Proceedings of the AAAI conference on artificial intelligence 33. https://doi.org/10.1609/aaai.v33i01.33013959
29. Li Y, Oliva JB (2020) Active feature acquisition with generative surrogate models. In: Proceedings of international conference on machine learning. https://api.semanticscholar.org/CorpusID:222142179
30. Covert I, Qiu W, Lu M, Kim N, White N, Lee S-I (2023) Learning to maximize mutual information for dynamic feature selection. In: Proceedings of the 40th international conference on machine learning
31. Asgaonkar V, Jain A, De A (2024) Generator assisted mixture of experts for feature acquisition in batch. In: Proceedings of the 38th international conference on artificial intelligence
32. Johnson AEW, Pollard TJ, Shen L, Lehman L-w H, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Scientific Data 3:160035
33. Pollard T, Johnson A, Raffa J, Celi L, Mark R, Badawi O (2018) The eICU collaborative research database, a freely available multi-center database for critical care research. Scientific Data 5:180178. https://doi.org/10.1038/sdata.2018.178
34. Xu K, Cheong C, Veldsman W, Lyu A, Cheung W, Zhang L (2023) Accurate and interpretable gene expression imputation on scrna-seq data using IGSimpute. Briefings in Bioinf 24
35. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S (2016) Deepr: a convolutional net for medical records. IEEE J Biomed Health Inf 22–30
36. Sha Y, Wang MD (2017) Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics, pp 233–240
37. Choi E, Bahadori MT, Song L, Stewart WF, Sun J (2017) GRAM: Graph-based attention model for healthcare representation learning. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 787–795
38. Song L, Cheong CW, Yin K, Cheung WK, Fung BCM, Poon J (2019) Medical concept embedding with multiple ontological representations. In: Proceedings of the 28th international joint conference on artificial intelligence, pp 4613–4619
39. Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, Lukasiewicz T, Rahimi K (2023) Hi-BEHRT: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE J Biomed Health Inform 27:1106–1117

Authors and Affiliations

Chin Wang Cheong[1] · Kejing Yin[1] · William K. Cheung[1] · Ivor Tsang[2]

✉ Chin Wang Cheong
cwcheong@comp.hkbu.edu.hk

Kejing Yin
cskjyin@comp.hkbu.edu.hk

William K. Cheung
william@comp.hkbu.edu.hk

Ivor Tsang
ivor.tsang@uts.edu.au

[1] Department of Computer Science, Hong Kong Baptist University, 224 Waterloo Road, Kowloon Tong, Hong Kong, China

[2] School of Computer Science, University of Technology Sydney, 15 Broadway Ultimo, Sydney, New South Wales 2007, Australia