DISC: Dynamic Feature Selection for Cost-Sensitive Medical Diagnosis

Yu-sheng Li[12*], Xincen Duan[3*], Beili Wang[3], Wei Guo[3†], Han-jia Ye[12†]

[1] School of Artificial Intelligence, Nanjing University, Nanjing, China
[2] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[3]Zhongshan Hospital, Fudan University, Shanghai, China
liys@lamda.nju.edu.cn, duan.xincen1@zs-hospital.sh.cn, wang.beili1@zs-hospital.sh.cn,
guo.wei@zs-hospital.sh.cn, yehj@lamda.nju.edu.cn

Abstract

Accurate medical diagnosis often relies on both textual self-reported symptoms and structured medical examination results of patients. However, these examinations vary significantly in cost—measured in time, money, or patient discomfort—creating a challenging trade-off between diagnostic accuracy and resource efficiency. To address this issue, we propose a dynamic diagnostic framework that incrementally selects medical examinations based on individual characteristics of each patient. Starting with textual self-reported symptoms and basic demographic, the system determines follow-up examinations step-by-step, improving accuracy while minimizing additional costs. Specifically, we introduce Dynamic feature selection with Instance-Specific Cost sensitivity (DISC). DISC treats each examination as a feature and learns to acquire them sequentially to optimize predictive performance under personalized cost constraints. To support richer clinical understanding, we further develop a multimodal framework that integrates unstructured self-reported symptom text with structured medical examination data. We conduct experiments on 680,000 patients with 43 million medical examination records, demonstrating that DISC high diagnostic accuracy even when accounting for examination costs. Our work provides substantial momentum for the advancement of AI in healthcare, offering both methodological and practical foundations that can significantly accelerate the deployment of intelligent, cost-aware diagnostic systems in real-world clinical settings.

Code — https://github.com/LAMDA-Tabular/DISC

Introduction

In real-world medical consultations, each medical examination (e.g., hemoglobin concentration, blood glucose level, liver enzyme panels) comes with a cost—whether financial, temporal, or physical (Bernardino et al. 2022; Foster, Koprowski, and Skufca 2014). Most tests require payment, may involve long waiting periods for results, and can sometimes cause discomfort or risk to patients. Without careful planning, excessive examinations can impose significant burdens on both individuals and the broader healthcare system. As a result, medical diagnoses are often conducted over multiple interactions: physicians initially assess available information (symptoms and examination results), then selectively request additional examinations to support decision-making.

In practice, these decisions must also account for patient-specific constraints. For instance, a patient may be unable to afford expensive tests, or may only have access to basic examinations due to time or logistical limitations. To address this challenge, we propose to model the diagnostic decision-making process using machine learning, with the goal of automatically and adaptively selecting a personalized set of medical examinations. Our objective is to *make high diagnostic accuracy while accounting for each patient's varying sensitivity to acquisition costs*.

More specifically, we consider the set of all possible medical examinations is treated as *structured features*. The diagnosis then becomes the target prediction. This formulation allows us to view diagnosis as a *dynamic feature selection* (Janisch, Pevnỳ, and Lisỳ 2020) problem: the model begins with basic information and *dynamically* selects which examination to perform next, based on the current acquired results. However, existing methods for dynamic feature selection face several limitations. Many existing approaches (Ma et al. 2019; Ghosh and Lan 2023; Covert et al. 2023) focus on improving accuracy under a feature-count budget but overlook that different examinations carry heterogeneous costs. Although a number of cost-aware methods (Dulac-Arnold et al. 2011; Shim, Hwang, and Yang 2018; Valancius, Lennon, and Oliva 2024) explicitly model feature costs, they typically assume a uniform sensitivity to cost across all instances, thus failing to capture patient-specific preferences or constraints.

To address these limitations, we propose DISC: **D**ynamic feature selection with **I**nstance-**S**pecific **C**ost sensitivity. DISC is a general framework that incorporates instance-dependent cost sensitivity into the dynamic feature selection process while supporting multimodal inputs. Starting from the patient's self-reported symptoms, DISC sequentially selects the most informative and cost-effective examinations to perform, guided by both feature costs and patient-specific tolerance to cost. At each step, DISC updates its predictions based on accumulated features and continues acquisition until a stopping criterion is met.

In particular, DISC consists of three core components. (1)

[*]These authors contributed equally to this work.
[†]The corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

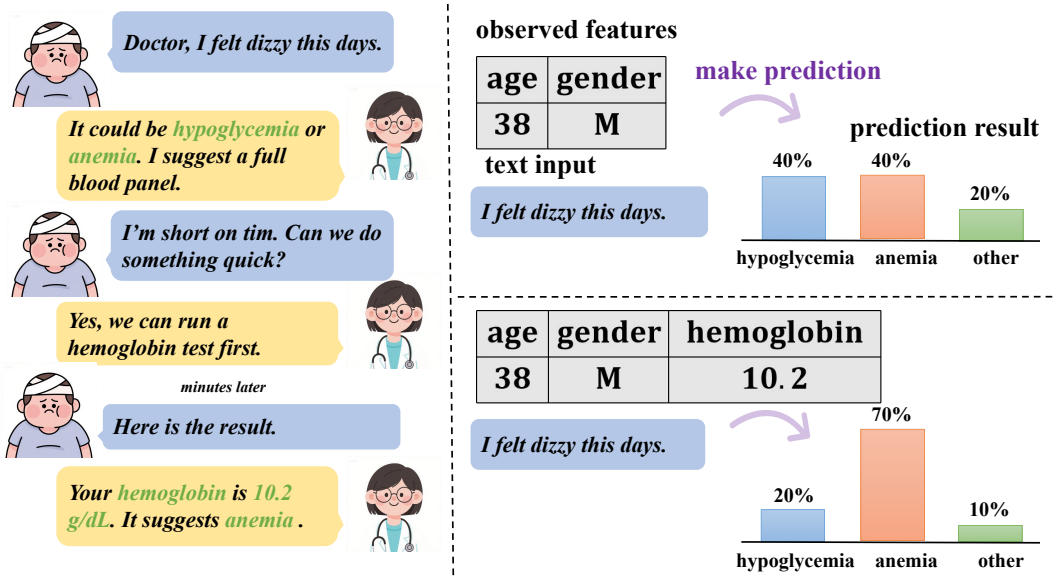

Figure 1: The medical diagnostic process can be formulated as a dynamic feature selection problem. Based on the patient information (e.g., age, gender, and symptoms) and preferences (e.g., time constraints, financial concerns), physicians arrange specific examinations, a process that corresponds to selecting new features.

A *scorer* evaluates all unobserved features based on the current input and the features acquired so far. (2) A *selector* then determines which feature to acquire next by jointly considering the feature scores, acquisition costs, and the patient's individual cost sensitivity. (3) A *predictor* produces diagnostic predictions using the currently available feature set. The scorer and predictor are jointly trained to ensure strong predictive performance under incomplete information. Once trained, the selector is fine-tuned to reflect instance-specific cost preferences and to avoid acquiring redundant or overly expensive features. We evaluate DISC on a real-world hospital dataset encompassing approximately 680,000 patients with 43 million medical examination records. DISC significantly reduces the number of acquired features—down to roughly 30%—while maintaining comparable diagnostic performance to models using all features. We further validate the generalizability of DISC on five public datasets from diverse domains.

Our main contributions are summarized as follows:

- We formulate real-world medical diagnosis as a multimodal dynamic feature selection problem, integrating text and structured data under cost constraints.
- We propose DISC, a novel instance-specific and cost-sensitive dynamic feature selection framework tailored to the diagnostic setting.
- We validate DISC on a real hospital dataset and five public datasets, achieving strong performance and reducing the cost of acquired features.

Related Work

Medical Diagnosis on Structured Tabular Data

Machine learning achieves notable success in the medical domain, with strong performance in tasks such as circulatory failure prediction (Hyland et al. 2020), leukemia diagnosis (Alcazer et al. 2024), and parathyroid hormone-related peptide testing (Yang et al. 2023). In these settings, patient data, such as demographic information and medical examination, is typically represented in structured tabular form (Jiang et al. 2025; Cai and Ye 2025; Ye et al. 2024), where each row corresponds to a patient and is associated with a diagnostic label. Both tree-based models (Chen and Guestrin 2016; Ke et al. 2017; Prokhorenkova et al. 2018) and deep learning methods (Gorishniy et al. 2021; Gorishniy, Kotelnikov, and Babenko 2025; Ye et al. 2025; Hollmann et al. 2025; Liu and Ye 2025) demonstrate competitive performance on such tabular datasets. However, most existing methods treat feature selection and prediction as separate steps, with a primary focus on building predictive models under the assumption that all features are available at inference time. This assumption is often unrealistic in clinical practice, where certain features may be missing, unavailable, or costly to acquire. So it is challenging for conventional approaches to deploy in real-world diagnostic workflows that require adaptive and cost-aware feature acquisition.

Feature Selection

Feature selection aims to identify a small subset of relevant features that can improve classification performance, reduce dimensionality and storage requirements, enhance computational efficiency, and facilitate data visualization and interpretation (Li et al. 2017; Dhal and Azad 2022; Jiao et al.

2023). It plays a crucial role in data mining, pattern recognition, and machine learning. Unlike feature construction or extraction methods (Ding et al. 2012; Liu et al. 2008), feature selection preserves the original semantics of features, thereby offering better interpretability.

Static feature selection Traditional feature selection methods typically focus on selecting a static feature subset shared across all instances. These methods rank features by their importance using techniques such as attribution analysis (Shrikumar, Greenside, and Kundaje 2017), correlation analysis (Hall 1999), information theory (Peng, Long, and Ding 2005), or game theory (Lundberg and Lee 2017), and then select the top-ranked features based on this ranking.

Dynamic feature selection Dynamic feature selection, which adapts the selected features to each individual instance, receives increasing attention. Existing approaches span several methodological families. A widely used line of work (Dulac-Arnold et al. 2011; Shim, Hwang, and Yang 2018; Li and Oliva 2021) formulates the problem as a reinforcement learning (RL) task (Sutton and Barto 1998), modeling feature acquisition as a Markov Decision Process (Bellman 1966). Although theoretically appealing (Dulac-Arnold et al. 2011), RL-based methods are often difficult to train and do not consistently outperform static feature selection techniques (Ghosh and Lan 2023).

Beyond RL, the EDDI framework (Ma et al. 2019) employs a variational autoencoder with arbitrary conditioning (Ivanov, Figurnov, and Vetrov 2019) to estimate conditional mutual information, but its estimates can be unreliable and computationally expensive. Other recent methods take greedy or amortized optimization approaches (Covert et al. 2023; Bae et al. 2022), while differentiable acquisition strategies such as DIFA (Ghosh and Lan 2023) avoid zeroth-order optimization and enable efficient learning. However, these techniques overlook feature acquisition costs. The AACO method (Valancius, Lennon, and Oliva 2024) incorporates cost-aware decision-making by estimating expected loss under partial observations via neighboring samples. Yet, it cannot capture individualized cost sensitivities.

Problem Definition and Notations

To enable cost-aware medical diagnosis, we aim to dynamically select a subset of informative and affordable features for each patient, based on both textual symptoms and structured examination data. We now introduce the notations used in our formulation.

We posit that physicians rely on two sources of information when making diagnostic predictions. The first is the self-reported symptoms of patients, which we denote as $\boldsymbol{x}_{\mathrm{t}}$. The second information source includes patient metadata (e.g., gender, age) and examination results (e.g., blood pressure, blood glucose, blood lipids). These medical examination results are collected sequentially and differ across patients. To model this process, we construct a feature vector $\boldsymbol{x}_{\mathrm{f}}$ of length d that represents patient metadata together with all possible examinations, even though each patient typically observes only a subset of them. We further introduce a binary mask $\boldsymbol{m} \in \{0,1\}^d$, which indicates which features are acquired. The masked feature vector actually observed for a given patient is therefore denoted as $\boldsymbol{x}_{\mathrm{m}}$:

$$\boldsymbol{x}_{\mathrm{m}} = (\boldsymbol{x}_{\mathrm{f}} \odot \boldsymbol{m}) \oplus \boldsymbol{m}, \tag{1}$$

where $\odot$ denotes element-wise multiplication and $\oplus$ denotes vector concatenation. $\odot$ is used to set all unacquired features to zero, while $\oplus$ is used to distinguish whether a zero indicates that the feature is not acquired or that the feature value itself is zero. We aim to learn a predictor f to produce accurate diagnostic predictions based on the provided self-reported symptoms $\boldsymbol{x}_{\mathrm{t}}$ and the masked feature vector $\boldsymbol{x}_{\mathrm{m}}$:

$$\hat{y} = f(\boldsymbol{m} \mid \boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{f}}), \tag{2}$$

where $\boldsymbol{x}_{\mathrm{t}}$ and $\boldsymbol{x}_{\mathrm{f}}$ are fixed for a given patient, representing all available patient information. $\boldsymbol{m}$ is a dynamic feature mask that controls which parts of this information are actually observable, and unobservable features wait for selection.

Since acquiring medical examinations in real-world settings often incurs non-negligible costs, we additionally introduce a cost vector $\boldsymbol{c}$ that specifies the acquisition cost of each feature. Prior work in dynamic feature selection (Ma et al. 2019; Covert et al. 2023; Ghosh and Lan 2023) learns a selection policy π and determines which feature to acquire next by jointly considering both its acquisition cost and its potential contribution to improving prediction accuracy:

$$\boldsymbol{m}^{\mathrm{oh}} = \pi(\boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{m}}, \boldsymbol{c}). \tag{3}$$

Here, $\boldsymbol{m}^{\mathrm{oh}}$ is a one-hot vector indicating the selected feature. It is used to update the feature mask $\boldsymbol{m}$, thereby enabling the sequential feature acquisition process.

Therefore, until a total of n features are selected, we can define an evaluation metric for the selection policy π and the predictor f as:

$$v(\pi, f) = \mathbb{E}_{p(\boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{f}}, y)} \Big[\sum_{i=0}^{n} \ell\Big(y, f(\boldsymbol{m}_i \mid \boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{f}})\Big)\Big], \tag{4}$$

where $p(\boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{f}}, y)$ denotes the data distribution across patients, and ℓ denotes the loss function corresponding to the task; for diagnostic classification tasks, we use the cross-entropy loss. $\boldsymbol{m}_i$ denotes the mask vector after the i-th selection step, with $i \leq n$. The optimization is performed by updating the parameters of both the selection policy π and the predictor f. A lower value of this metric indicates that the predictor achieves higher accuracy at each step of the selection process, implying that the selected features are more informative and discriminative.

DISC

To account for both feature acquisition costs and instance-specific cost sensitivity during the feature selection process, we propose DISC. In DISC, we decompose the selection policy π into two components:(1) a scorer π_1 and (2) a selector π_2. The scorer evaluates the utility of each unacquired feature and the selector decides whether to acquire a feature based on its utility and acquisition cost. DISC also retains (3) a predictor f to generate real-time diagnostic predictions using the currently available information.

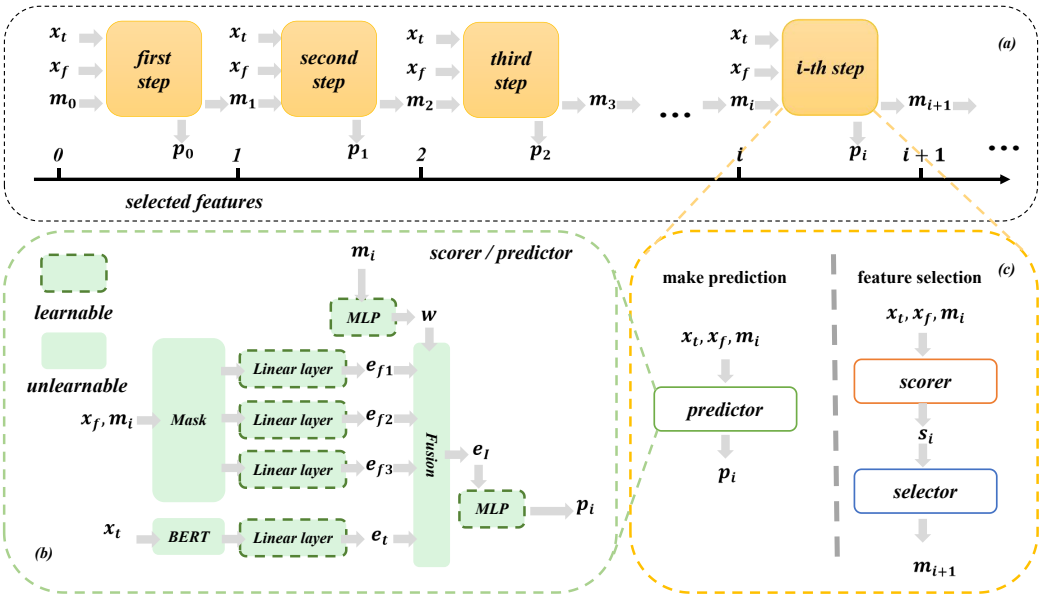

Figure 2: Overview of the DISC framework. (a) DISC starts with an all-zero feature mask and incrementally acquires features step-by-step to make predictions. (b) Detailed architecture of the predictor; the scorer shares the same structure, differing only in output dimensionality. (c) At each step, the model uses a shared backbone with a *scorer* and a *selector* to update the feature mask, and a *predictor* to generate diagnostic predictions. Selection and prediction are decoupled but use shared components.

In this section, we first present a detailed explanation of the operational mechanism of DISC, with a focus on how it incorporates instance-specific cost sensitivity. We then describe the training procedure, including implementation details of the model architecture. The operating mechanism of DISC is illustrated in Figure 2.

Selection Policy and Prediction

As mentioned before, DISC consists of three components: a scorer, a selector, and a predictor. Both the scorer and the predictor are learnable modules, with their parameters denoted as $\pi_1(\cdot;\phi)$ and $f(\cdot;\theta)$, respectively. The selector is implemented as a gating mechanism, denoted by $\pi_2(\cdot;\beta)$, where β is a separately tunable parameter that requires independent optimization. Together, the scorer and the selector constitute the selection policy π.

When we select $i{-}1$ features. Let $\boldsymbol{m}_{i-1}$ denote the corresponding feature mask. The scorer takes all currently available information as input and produces a score vector $\boldsymbol{s}$:

$$\boldsymbol{s} = \pi_1(\boldsymbol{m}_{i-1} \mid \boldsymbol{x}_\mathrm{t}, \boldsymbol{x}_\mathrm{f}; \phi). \tag{5}$$

To ensure that already selected features are not reselected and that both feature acquisition costs and instance-specific cost sensitivity are properly considered, DISC applies a selector to adjust the score vector $\boldsymbol{s}$ as follows:

$$\begin{aligned} \boldsymbol{s}' &= \pi_2(\boldsymbol{s} \mid \boldsymbol{c}, \gamma; \beta) \\ &= g\Big(\boldsymbol{s} - \alpha \times \boldsymbol{m}_{i-1} - \beta \times \gamma \times \boldsymbol{c}\Big) \end{aligned} \tag{6}$$

where function $g(\cdot)$ refers to the Gumbel-Softmax function (Jang, Gu, and Poole 2017), which transforms a score vector into a one-hot vector by resampling trick. Within function $g(\cdot)$, the second term involves a large constant α and sets the scores of selected features to $-\infty$. This prevents the re-selection of already acquired features. The third term penalizes feature scores based on instance-specific cost sensitivity γ and feature acquisition costs $\boldsymbol{c}$ mentioned before. Here, $\gamma \in [0, 1]$ quantifies a patient's tolerance for feature acquisition cost: a larger γ increases the penalization of expensive features and results in a more pronounced adjustment of the score vector. The modified score vector $\boldsymbol{s}'$ then updates the feature mask vector $\boldsymbol{m}_{i-1}$ as follows:

$$\boldsymbol{m}_i = \boldsymbol{m}_{i-1} + \boldsymbol{s}'. \tag{7}$$

The predictor and the scorer share the same underlying architecture, as both receive all selected features and text descriptions as input. The only difference lies in their outputs, the predictor generates a probability vector representing the predicted diagnosis result:

$$\boldsymbol{p}_i = f(\boldsymbol{m}_i \mid \boldsymbol{x}_\mathrm{t}, \boldsymbol{x}_\mathrm{f}; \theta). \tag{8}$$

Multi-step feature selection starts at $i{=}0$ and iteratively performs feature selection, acquiring one feature and updating the feature mask $\boldsymbol{m}_i$ at each step until the predefined maximum number of features is reached.

For simplicity, in the main text, we assume each feature is associated with a single cost and do not consider potential dependencies among features (e.g., multi-cost settings, group structures, feature acquisition with preconditions).

Structure of Scorer and Predictor

The feature scorer and predictor share the same underlying architecture, as both take the same input type: textual information $\boldsymbol{x}_\mathrm{t}$, feature vector $\boldsymbol{x}_\mathrm{f}$ and feature mask vector $\boldsymbol{m}$. The

only distinction lies in the output layer: the scorer produces a score vector $\boldsymbol{s}$, which is used to determine the next feature to acquire, while the predictor outputs a prediction vector $\boldsymbol{p}$, which is used to compute the final classification loss. A key challenge in the model design is to handle the dynamically expanding subset of selected features and to effectively integrate these features with the associated textual information.

Text embedding For Discriminative tasks involving textual information (in our dataset, each sample is associated with one corresponding sentence), we employ a pretrained language model, BERT (Devlin et al. 2019), as the text encoder. BERT maps each sentence to a 768-dimensional representation. To align the embedding dimension with other modalities, we add a linear projection layer that transforms the 768-dimensional vector into an h-dimensional text embedding. This process is defined as:

$$\boldsymbol{e}_{\mathrm{t}} = \mathrm{Linear}(\mathrm{BERT}(\boldsymbol{x}_{\mathrm{t}})), \tag{9}$$

where $\boldsymbol{x}_{\mathrm{t}}$ denotes the input textual information, and $\boldsymbol{e}_{\mathrm{t}}$ is the h-dimensional text embedding.

MoE embedding To enhance the representational capacity of DISC, we adopt a MoE framework (Jordan and Jacobs 1994)for embedding construction. The previously obtained text embedding $\boldsymbol{e}_{\mathrm{t}}$ serves as one of the experts in the MoE, specializing in textual information. In addition, we introduce three more experts to process the masked feature vector $\boldsymbol{x}_{\mathrm{m}}$, each producing an h-dimensional embedding:

$$\boldsymbol{e}_{\mathrm{f1}}, \boldsymbol{e}_{\mathrm{f2}}, \boldsymbol{e}_{\mathrm{f3}} = \mathrm{Linear}(\boldsymbol{x}_{\mathrm{m}}). \tag{10}$$

A router module selects and weights these expert outputs based on the feature mask $\boldsymbol{m}$, enabling dynamic routing. The final mixed embedding $\boldsymbol{e}_{\mathrm{I}}$ is computed as a weighted sum of all expert embeddings, guided by the routing weights. This fused representation is then passed through a multi-layer perceptron (MLP) to generate the final output:

$$\mathrm{output} = \mathrm{MLP}(\boldsymbol{e}_{\mathrm{I}}). \tag{11}$$

For the scorer, output is the score vector $\boldsymbol{s}$, and for the predictor, output is the predicted logits $\boldsymbol{p}$.

This framework assumes that each sample is accompanied by a piece of textual information, which, along with the selected features, guides both feature acquisition and prediction. In practice, such textual data is often easy to obtain and can take the form of any sample-relevant description. However, since these descriptions are inherently subjective—and may be vague or even inaccurate—the influence of the text embedding is gradually attenuated as more informative features are acquired during the selection process.

Training Recipe

Joint training of the scorer and predictor DISC jointly trains the scorer and predictor. Given a maximum feature budget n, DISC performs n iterations, each involving a feature acquisition step followed by a prediction. We denote predictions at each step as $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$, where $\boldsymbol{p}_0$ corresponds to the prediction based solely on textual information. Each prediction is compared to the ground-truth label y to compute a loss. Referring to Equation 4, we rewrite $v(\pi, f)$ as the empirical loss. The overall training loss is defined as:

$$\mathcal{L} = \mathbb{E}_{p(\boldsymbol{x}_{\mathrm{t}}, \boldsymbol{x}_{\mathrm{f}}, y)} \sum_{i=0}^{n} \ell(\boldsymbol{p}_i, y). \tag{12}$$

where ℓ denotes the cross-entropy loss for classification tasks, and D is the dataset of patients. The computed loss is used to update the parameters ϕ for scorer and θ for predictor. During joint training, we set γ=0, meaning the selector ignores feature acquisition costs and only avoids selecting observed features. This allows DISC to learn a selection strategy prioritizing predictive performance.

Adjust the constrained feature selector After joint training, we freeze the parameters of both the scorer and the predictor. We then tune only the parameter β, aiming to select a value that enables the model to produce distinct feature selections under varying cost sensitivity values γ. This resembles hyperparameter tuning: fix β, then observe how feature selections change with γ for a given instance.

Dataset	# Instances	# Features	# Classes
Clinical Datasets			
Respiratory Medicine	8,193	130	7
Endocrinology	7,922	86	5
Urinary Surgery	5,314	106	4
Public Datasets			
Attrition	938	29	2
House Prices	2,919	52	4
Credit Default	4,399	101	2
Diabetes	12,580	42	3
Developer Survey	6,745	135	2

Table 1: Overview of datasets used in our experiments. The top three are real-world clinical datasets from a top-tier hospital, containing examination records and symptom descriptions. The bottom five are public tabular datasets with synthetic text generated by DeepSeek-R1.

Experiments

This section presents the datasets used in our experiments, along with the baseline methods for comparison.

Datasets

To align with the medical scenario discussed throughout this paper, we collected real-world patient examination records from three departments, Respiratory Medicine, Endocrinology, and Urinary Surgery of a top-tier hospital. These datasets contain authentic patient data gathered during actual clinical visits. For each patient visiting a department, the textual component of the dataset records the self-reported symptoms from patient (e.g., recent dietary patterns or reported areas of pain). In addition, the dataset includes a examination sequence for each patient, documenting all medical examinations conducted and their corresponding results. Based on these sequences, we construct a unified ta-

Dataset	CAE	mRMR	DeepLIFT	EDDI	JAFA	Amo	DISC
Respiratory Medicine	75.35	79.77	74.91	76.87	62.47	78.97	**80.89**
Endocrinology	83.07	83.07	83.97	71.64	74.52	83.48	**88.85**
Urinary Surgery	74.21	83.06	73.71	78.45	72.80	79.25	**83.37**
Attrition	68.50	77.00	75.70	72.13	64.61	74.37	**77.75**
House Prices	85.91	85.71	85.98	82.30	82.85	89.87	**90.10**
Credit Default	62.18	71.97	69.08	66.20	58.52	73.55	**73.69**
Diabetes	58.63	61.10	60.89	58.86	56.42	61.91	**62.90**
Developer Survey	73.69	82.48	81.72	68.80	68.71	**83.49**	82.95

Table 2: Comparison of average AUROC scores across seven feature selection methods on eight datasets. For each dataset, we report the mean AUROC under varying feature subset sizes. The best-performing method per dataset is highlighted in bold. DISC achieves the highest AUROC on seven out of eight datasets and ranks second on the remaining one.

ble where each row represents a patient, and each column corresponds to a possible medical examination. The values in the table indicate the actual results of the examinations if conducted, or remain empty otherwise. This table, combined with basic demographic information (e.g., age and gender), constitutes the tabular part of the dataset.

To ensure the reproducibility of our experimental results and to assess the generalizability of our method to other domains, we additionally selected five publicly available tabular datasets for evaluation. For each of these datasets, we randomly selected a small subset of features and input their information into `DeepSeek-R1` (Guo et al. 2025) to generate a unique textual description for each instance. The selected columns are then removed from the original dataset.

For each dataset, we employ a large language model to estimate the acquisition cost of each feature. Specifically, we input the column names along with their corresponding descriptions into the large language model, which outputs a relative cost estimate for each feature, forming the cost vector $\boldsymbol{c}$. In our experiments, the predicted costs are constrained to integer values ranging from 0 to 5. For the real hospital dataset, We instruct `DeepSeek-R1` to estimate the acquisition cost of each feature based on the difficulty and expense involved in obtaining it. For example, age and gender are easily accessible through patient inquiries or hospital records, and thus their acquisition cost is set to zore. In contrast, blood pressure measurements are common and relatively convenient, typically performed at no additional cost during standard consultations. However, they still require specialized medical equipment and trained personnel, and thus incur a relatively low acquisition cost. Features such as red blood cell count, which can only be obtained through blood examination, involve both financial expenses and waiting time for results, and are therefore assigned a higher acquisition cost.

All datasets are classification tasks, with instance counts ranging from 938 to 12,580 and feature counts ranging from 29 to 135. Table 1 presents key statistics for each dataset.

Baselines

Feature selection We evaluate our method by comparing it with both dynamic and static feature selection methods. As static baselines, we use a mutual information selection method mRMR (Peng, Long, and Ding 2005), a deep model attribution method DeepLIFT (Shrikumar, Greenside, and Kundaje 2017), and the Concrete Autoencoder (CAE) (Balın, Abid, and Zou 2019) for differentiable feature selection. As dynamic baselines, we use the efficient dynamic discovery method EDDI (Ma et al. 2019), a reinforcement learning method JAFA (Shim, Hwang, and Yang 2018), and an amortized optimization method Amo (Covert et al. 2023) to dynamically sample unselected features.

Multimodal methods for text and tabular data We consider the following three approaches to process multimodal data consisting of both tabular features and text descriptions. BERT (Devlin et al. 2019): We use a pre-trained language model BERT to encode each textual input into a vector representation. The resulting text embedding is concatenated with its corresponding tabular feature vector, and the combined representation is passed through an MLP for prediction. CatBoost (Prokhorenkova et al. 2018): We utilize the CatBoost library, which supports raw text fields. CatBoost internally applies text embedding and tokenization mechanisms, enabling it to directly process both tabular features and text descriptions. TabPFN (Hollmann et al. 2025): We adopt TabPFN, a state-of-the-art tabular foundation model. In its recent versions, TabPFN supports the inclusion of designated text columns, allowing for joint modeling of structured features and instance-specific textual information.

Results

Feature selection We compare seven different feature selection methods, including DISC. For each method, a maximum of n features is selected, and a set of prediction results is generated each time a feature is selected. These results are compared with the true labels to calculate the AUROC values, and the average of the AUROC values across the n selection steps is computed as the metric.

We compare seven feature selection methods. For each method, a maximum of n features is selected, and prediction results are generated after each feature selection step. AUROC is calculated at each step, and the overall metric is computed as the average AUROC across all n selection steps. To ensure fairness, these experiments do not use any textual information, nor do they consider the costs of feature acquisition. As shown in Table 2, DISC achieves

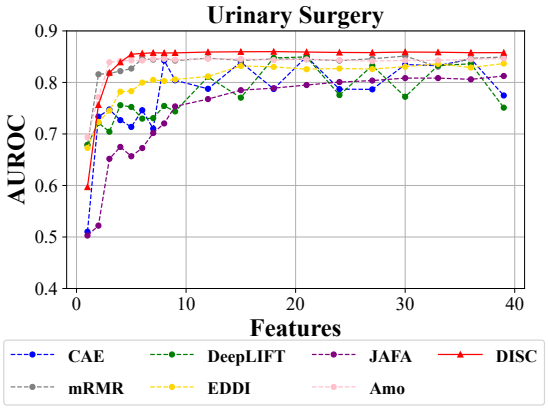

Figure 3: Features-AUROC curves with seven feature selection methods on the Urinary Surgery dataset.

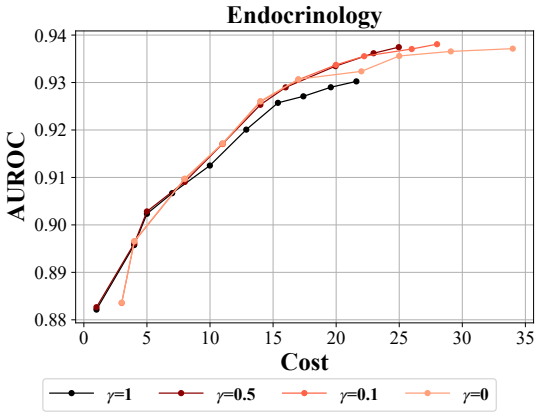

Figure 5: Cost-AUROC curves of DISC under different cost-sensitivity coefficient γ on the Endocrinology dataset.

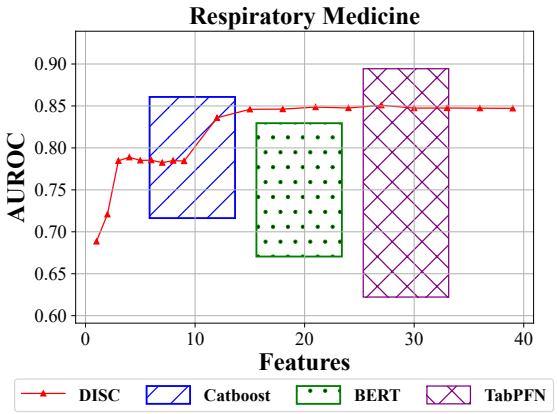

Figure 4: Features-AUROC curve of DISC and the AUROC of the prediction result from three multimodal methods on the Respiratory Medicine dataset. The lower and upper rectangles in the figure represent the results using text only and using both text and tabular information, respectively.

the best performance among all compared methods under these settings. Although motivated by the medical diagnostic process, DISC generalizes well to five additional public datasets. Figure 3 illustrates the Features-AUROC curves of the seven methods on the Urinary Surgery dataset.

Multimodal methods for text and tabular data We conduct a study of three multimodal prediction approaches: BERT combined with an MLP, CatBoost, and TabPFN. For each method, we conduct two experiments: (1) using only the textual modality as input, and (2) using both the textual and tabular modalities. Figure 4 shows the Features-AUROC of DISC and the AUROC of prediction results from three multimodal methods on the Respiratory Medicine dataset. For each bar in Figure 4, the lower segment represents the prediction performance using only textual modality information, while the top segment represents the performance when using both the textual and tabular modalities.

DISC achieves better prediction performance than the BERT-based method using all features after selecting only 15 out of 130 features. Its performance is comparable to CatBoost, though about 0.05 lower than that of TabPFN.

Feature cost sensitivity After jointly training the scorer and predictor, we tune the selector parameter β. Our goal is to ensure that for the same patient, varying the cost-sensitivity coefficient γ results in different feature selection behaviors. To evaluate this, we vary γ from 0 to 1, and for each value, allow all instances to select 10 features. We then analyze the resulting Cost-AUROC curves to assess the trade-off between cost and performance. Figure 5 shows the Cost-AUROC curves of DISC on the Endocrinology dataset, corresponding to different γ values.

A smaller γ indicates less sensitivity to feature acquisition costs. As shown in Figure 5, this is reflected by larger horizontal gaps between adjacent points, implying that DISC tends to select more expensive but more predictive features. When the number of selected features is fixed at 10, the total cost under γ=1 is approximately 22, while under γ=0 it rises to nearly 35. However, the resulting performance gain is marginal, less than 0.01 in AUROC.

Conclusion

We formulate medical diagnosis as a dynamic feature selection problem and propose DISC, a framework that integrates multimodal inputs and models instance-specific sensitivity to feature acquisition costs. By tailoring feature selection to individual patients, DISC enables personalized and cost-effective diagnostic decisions, addressing key limitations of existing approaches that assume uniform cost tolerance. Extensive experiments on three real-world hospital datasets and five public benchmarks demonstrate DISC's strong generalizability and superior predictive performance. As medical images are also critical in clinical practice, future work will explore extending DISC to incorporate imaging modalities for more comprehensive diagnostic support.

Acknowledgments

This work is partially supported by the Key Program of Jiangsu Science Foundation (BK20243012) and the National Natural Science Foundation of China (82402736, 82172348, 82102468); Zhongshan Hospital (2023ZSQN30).

References

Alcazer, V.; Le Meur, G.; Roccon, M.; Barrière, S.; Le Calvez, B.; Badaoui, B.; Spaeth, A.; Kosmider, O.; Freynet, N.; Eveillard, M.; et al. 2024. Evaluation of a Machine-Learning Model Based on Laboratory Parameters for the Prediction of Acute Leukaemia Subtypes: A Multicentre Model Development and Validation Study in France. *The Lancet Digital Health*, 6(5): e323–e333.

Bae, J.; Vicol, P.; HaoChen, J. Z.; and Grosse, R. B. 2022. Amortized Proximal Optimization. In *Advances in Neural Information Processing Systems*, 8982–8997.

Balın, M. F.; Abid, A.; and Zou, J. 2019. Concrete Autoencoders: Differentiable Feature Selection and Reconstruction. In *Proceedings of the 36th International Conference on Machine Learning*, 444–453.

Bellman, R. 1966. Dynamic Programming. *Science*, 153(3731): 34–37.

Bernardino, G.; Jonsson, A.; Loncaric, F.; Martí Castellote, P.-M.; Sitges, M.; Clarysse, P.; and Duchateau, N. 2022. Reinforcement Learning for Active Modality Selection During Diagnosis. In *Medical Image Computing and Computer-Assisted Intervention*, 592–601.

Cai, H.; and Ye, H.-J. 2025. Feature-aware Modulation for Learning from Temporal Tabular Data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Covert, I. C.; Qiu, W.; Lu, M.; Kim, N. Y.; White, N. J.; and Lee, S.-I. 2023. Learning to Maximize Mutual Information for Dynamic Feature Selection. In *Proceedings of the 40th International Conference on Machine Learning*, 6424–6447.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dhal, P.; and Azad, C. 2022. A Comprehensive Survey on Feature Selection in the Various Fields of Machine Learning. *Applied Intelligence*, 52(4): 4543–4581.

Ding, S.; Zhu, H.; Jia, W.; and Su, C. 2012. A Survey on Feature Extraction for Pattern Recognition. *Artificial Intelligence Review*, 37(3): 169–180.

Dulac-Arnold, G.; Denoyer, L.; Preux, P.; and Gallinari, P. 2011. Datum-Wise Classification: A Sequential Approach to Sparsity. In *Machine Learning and Knowledge Discovery in Databases*, 375–390.

Foster, K. R.; Koprowski, R.; and Skufca, J. D. 2014. Machine Learning, Medical Diagnosis, and Biomedical Engineering Research–Commentary. *Biomedical Engineering Online*, 13(1): 94.

Ghosh, A.; and Lan, A. 2023. DiFA: Differentiable Feature Acquisition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7705–7713.

Gorishniy, Y.; Kotelnikov, A.; and Babenko, A. 2025. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. In *Proceedings of the 13th International Conference on Learning Representations*.

Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Revisiting Deep Learning Models for Tabular Data. In *Advances in Neural Information Processing Systems*, 18932–18943.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948. ArXiv:2501.12948.

Hall, M. A. 1999. *Correlation-Based Feature Selection for Machine Learning*. Ph.D. thesis, The University of Waikato.

Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S. B.; Schirrmeister, R. T.; and Hutter, F. 2025. Accurate Predictions on Small Data with a Tabular Foundation Model. *Nature*, 637(8045): 319–326.

Hyland, S. L.; Faltys, M.; Hüser, M.; Lyu, X.; Gumbsch, T.; Esteban, C.; Bock, C.; Horn, M.; Moor, M.; Rieck, B.; et al. 2020. Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning. *Nature Medicine*, 26(3): 364–373.

Ivanov, O.; Figurnov, M.; and Vetrov, D. 2019. Variational Autoencoder with Arbitrary Conditioning. In *Proceedings of the 7th International Conference on Learning Representations*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of the 5th International Conference on Learning Representations*.

Janisch, J.; Pevnỳ, T.; and Lisỳ, V. 2020. Classification with Costly Features as a Sequential Decision-Making Problem. *Machine Learning*, 109(8): 1587–1615.

Jiang, J.-P.; Liu, S.-Y.; Cai, H.-R.; Zhou, Q.; and Ye, H.-J. 2025. Representation Learning for Tabular Data: A Comprehensive Survey. *CoRR*, abs/2504.16109. ArXiv:2504.16109.

Jiao, R.; Nguyen, B. H.; Xue, B.; and Zhang, M. 2023. A Survey on Evolutionary Multiobjective Feature Selection in Classification: Approaches, Applications, and Challenges. *IEEE Transactions on Evolutionary Computation*, 28(4): 1156–1176.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2): 181–214.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient

Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2017. Feature Selection: A Data Perspective. *ACM Computing Surveys*, 50(6): 94.

Li, Y.; and Oliva, J. 2021. Active Feature Acquisition with Generative Surrogate Models. In *Proceedings of the 38th International Conference on Machine Learning*, 6450–6459.

Liu, L.-P.; Yu, Y.; Jiang, Y.; and Zhou, Z.-H. 2008. TEFE: A time-efficient approach to feature extraction. In *2008 Eighth IEEE International Conference on Data Mining*, 423–432.

Liu, S.-Y.; and Ye, H.-J. 2025. TabPFN Unleashed: A Scalable and Effective Solution to Tabular Classification Problems. In *Proceedings of the 42nd International Conference on Machine Learning*.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 4766–4775.

Ma, C.; Tschiatschek, S.; Palla, K.; Hernández-Lobato, J. M.; Nowozin, S.; and Zhang, C. 2019. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, 4234–4243.

Peng, H.; Long, F.; and Ding, C. 2005. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8): 1226–1238.

Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems*.

Shim, H.; Hwang, S. J.; and Yang, E. 2018. Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding. In *Advances in Neural Information Processing Systems*, 1368–1378.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. *CoRR*, abs/1704.02685. ArXiv:1704.02685.

Sutton, R. S.; and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Valancius, M.; Lennon, M.; and Oliva, J. 2024. Acquisition Conditioned Oracle for Nongreedy Active Feature Acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, 48957–48975.

Yang, H. S.; Pan, W.; Wang, Y.; Zaydman, M. A.; Spies, N. C.; Zhao, Z.; Guise, T. A.; Meng, Q. H.; and Wang, F. 2023. Generalizability of a Machine Learning Model for Improving Utilization of Parathyroid Hormone-Related Peptide Testing Across Multiple Clinical Centers. *Clinical Chemistry*, 69(11): 1260–1269.

Ye, H.-J.; Liu, S.-Y.; Cai, H.-R.; Zhou, Q.-L.; and Zhan, D.-C. 2024. A closer look at deep learning methods on tabular datasets. *arXiv preprint arXiv:2407.04057*.

Ye, H.-J.; Yin, H.-H.; Zhan, D.-C.; and Chao, W.-L. 2025. Revisiting Nearest Neighbor for Tabular Data: A Deep Tabular Baseline Two Decades Later. In *Proceedings of the 13th International Conference on Learning Representations*.