Explainable Feature Engineering in Health Data Science: Empirical Comparison of ChatGPT-4o and Classical Machine Learning Methods

Behshid Behkamal, PhD
Western University
ON, Canada

Amin Rezaei, MS
University of Pittsburgh
PA, United States

Nickolas Littlefield, MS
University of Pittsburgh
PA, United States

Samarth Bhardwaj
University of Pittsburgh
PA, United States

Leah Reid, MD
University of Pittsburgh
PA, United States

Nicole Myers, MS, RN
University of Pittsburgh
PA, United States

Soheyla Amirian, PhD
Pace University
NY, United States

Ahmad P. Tafti, PhD
University of Pittsburgh
PA, United States

Abstract

Machine learning (ML) is demonstrating remarkable success in various healthcare applications. The success of ML in healthcare is inherently linked to the rigorous process of feature engineering and feature selection, which truly forms the backbone of ML model development. This study investigates the role of a well-known large language model (LLM), the ChatGPT-4o, in feature selection and classification processes for healthcare data, focusing on the explainability of ML. The performance of ChatGPT-4o is evaluated and compared to traditional ML methods—such as information gain (IG), correlation-based feature selection (CFS), and principal component analysis (PCA) for identifying relevant features in predictive modeling. This comparison is conducted using two widely recognized healthcare datasets, SEER and NSQIP. After evaluating the features selected by classical ML methods and LLMs through expert review, the results indicate that while ChatGPT-4o aligns closely with expert evaluations and effectively provides contextual information on healthcare datasets, traditional ML methods such as IG, CFS, and PCA outperform in systematic feature ranking due to their structured and data-driven nature. Furthermore, anonymization did not significantly affect the feature selection process, highlighting the robustness of ChatGPT-4o under privacy-preserving conditions. ChatGPT-4o's strength lies in complementing these methods by providing interpretability and facilitating exploratory analysis, rather than serving as a standalone solution for precise feature ranking.

Corresponding authors: Behshid Behkamal (behshid.behkamal@uwo.ca), Soheyla Amirian (samirian@pace.edu), and Ahmad P. Tafti (tafti.ahmad@pitt.edu).

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Behshid Behkamal, PhD. 2024. Explainable Feature Engineering in Health Data Science: Empirical Comparison of ChatGPT-4o and Classical Machine Learning Methods. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In recent years, artificial intelligence (AI) and ML have emerged as transformative tools in healthcare care, achieving significant advances in various clinical applications, such as early diagnosis, shared decision making, personalized treatment, and prediction of patient outcomes [3, 11, 28, 36]. A key part of building accurate and meaningful ML methods is to choose the right features from large and complex datasets. This involves feature engineering and feature selection. While feature engineering helps to create useful input variables, feature selection ensures that only the most important features are used in building ML models [15, 34, 39].

In the healthcare domain, where obtaining high-quality datasets is both costly and complex, feature selection plays a significant role in analyzing and extracting relevant information [23, 24, 30]. However, despite the successful adoption of AI/ML models in healthcare, challenges persist regarding the interpretability and explainability of these computational models [25, 35]. Understanding how AI/ML algorithms make decisions, especially in post-hoc analyses, provides essential insights into their reliability and accountability, helping to advance the implementation and uptake of AI/ML models.

This study investigates the emerging role of ChatGPT-4o, as a novel tool for feature selection in healthcare. Using two widely recognized healthcare datasets, including the Surveillance, Epidemiology, End Results Program (SEER) [1] and the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) [18]. We explore ChatGPT-4o's capability in feature selection and ranking to enhance the interpretability and explainability of predictive ML models. In this study, we define two key research questions to guide our investigation:

- **RQ (1):** Can ChatGPT-4o serve as an effective method for feature selection comparable to classical ML-based feature selection techniques, such as IG, CFS, and PCA?
- **RQ (2):** Does ChatGPT-4o rely on its pre-trained knowledge to identify and select important features?

To address these questions, we design two experimental scenarios. The first scenario evaluates the performance of ChatGPT-4o against traditional feature selection methods by assessing the relevance of features selected for two healthcare datasets, including SEER and NSQIP. This comparison provides insights into how well ChatGPT-4o aligns with established techniques. The second scenario investigates whether ChatGPT-4o's pre-trained knowledge influences its feature selection. To test this, we evaluate its performance using original and anonymized datasets, where feature names are replaced with generic labels to remove contextual cues. These scenarios allow us to comprehensively assess ChatGPT-4o's effectiveness, robustness, and reliance on prior knowledge in the feature selection process. With that, the significance of this work lies in the following key contributions:

- **Clinical Significance:** By focusing on the crucial steps of feature selection and feature importance, we aim to enhance the applicability of ML models in healthcare settings. Our exploration promises to contribute to more effective ML models and holds the key to uncovering complex patterns within healthcare data.
- **Technical Significance:** Our work specifically focuses on the underexplored domain of ChatGPT-4o as an LLM. The technical significance lies in exploring the potential of ChatGPT-4o to rank important features, contributing to the interpretability and explainability of ML models. This exploration expands the toolkit available to data scientists working with healthcare data, deepening our understanding of the synergy between large language models (LLMs) and feature selection, and highlighting its significance in machine learning applications.

The structure of this work is organized as follows. Section 2 provides a review of related literature. Section 3 details the materials and methods employed. Experimental validation and results are presented in Section 4. Finally, Section 5 discusses the contributions, concluding the study with proposed future directions.

2 Related Work

This section primarily explores the utilization of LLMs in healthcare and includes a brief comparison of feature selection methods relevant to our study.

2.1 LLMs Applications in Healthcare

In applying LLMs to healthcare settings, there are various ways to classify the state of the art. For example, Li et al. [19] conducted a systematic review of existing publications on the use of ChatGPT-4 in healthcare and proposed a two-sided taxonomy: application-oriented and user-oriented. The taxonomy is based on the nature of medical tasks, including triage, translation, medical research, clinical workflow, medical education, consultation, and multimodal mechanisms. Each task targets one or multiple end-user groups, such as patients, healthcare professionals, and researchers. In another study, Yu et al. [38] reviewed the literature on the integration of generative AI and LLMs into healthcare and medical practices. They presented their findings across various aspects, including technological approaches to generative AI applications, methods for training LLMs, model evaluation, current applications of generative AI and LLMs in healthcare and medicine, and regulatory considerations. Furthermore, Yang et al. [37] provided a practical guide for practitioners and end-users, demonstrating how to harness the power of LLMs for various downstream Natural Language Processing (NLP) tasks.

In this review, we classify previous works into two main groups.

- **Development and Utilization of Domain-Specific Healthcare LLMs.** This category focuses on building or fine-tuning LLMs specifically for healthcare applications. Singhal et al. [31] introduced MultiMedQA, a benchmark that combines six existing datasets with a newly developed dataset, HealthSearchQA, to evaluate the performance of LLM models, such as PaLM [9] and its instruction-tuned variant, Flan-PaLM [10]. The benchmark spans tasks ranging from professional medical exams to consumer health queries. Another benchmark, Med-HALT, was proposed by Pal et al. [33] to assess and mitigate hallucinations in medical LLMs. Med-HALT introduces two key test categories: Reasoning Hallucination Tests (RHTs), which measure a model's logical coherence and ability to avoid generating false information, and Memory Hallucination Tests (MHTs), which evaluate a model's accuracy in retrieving biomedical information. The dataset includes over 18,000 samples sourced from diverse medical exams and PubMed, ensuring comprehensive topic coverage and geographic representation. The study compared several models, including GPT-3.5, Falcon, and LLaMA-2, highlighting notable differences in performance. For instance, Falcon models excelled at fake question detection, while all models demonstrated room for improvement in memory-based tasks. The research emphasized the importance of benchmarks like Med-HALT in developing safer, more reliable LLMs for healthcare applications. Similarly, Han et al. [14] introduced MedAlpaca, an open-source collection of medical conversational AI models and training datasets specifically for healthcare applications. The framework fine-tunes Meta's LLaMA models (7B and 13B parameters) using a curated dataset of over 160,000 medical entries, including medical flashcards, Stack Exchange forums, WikiDoc content, and the USMLE [1] and CORD-19 [2].
  In addition to these benchmarks, BioGPT [21] is a generative pre-trained transformer model specifically designed for biomedical text generation and mining tasks. Built on the GPT-2 architecture, BioGPT is pre-trained on 15 million PubMed abstracts and fine-tuned for downstream tasks, including relation extraction, question answering, document

[1] https://www.usmle.org/sites/default/files/2021-10/Step_1_Sample_Items.pdf
[2] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

classification, and biomedical text generation. It introduces novel target sequence formats that leverage natural language semantics, enhancing its ability to produce structured and meaningful biomedical output. BioGPT outperforms state-of-the-art models across several benchmarks, achieving top performance in relation extraction and question answering, while also demonstrating superior text generation capabilities compared to general-purpose GPT-2.
Similarly, Guan et al. [12] proposed CohortGPT, an LLM-based system designed to assist in clinical trial participant recruitment by analyzing unstructured medical text, such as radiology reports. CohortGPT combines medical knowledge graphs with a Chain-of-Thought (CoT) reasoning approach, which improves its ability to classify medical reports accurately. By inferring disease-related details step-by-step, this approach enables the system to enhance performance even with limited labeled data. CohortGPT shows strong potential for streamlining participant recruitment and other medical text analysis tasks while requiring minimal training data. Additionally, the ChatDoctor model [20], a fine-tuned version of Meta's LLaMA-7B large language model, was specifically developed for medical applications. ChatDoctor was trained using a dataset comprising 100,000 real-world patient-doctor dialogues and an additional 10,000 dialogues for evaluation. It improves the understanding of patient inquiries and delivers accurate medical advice by incorporating a self-directed information retrieval system. This system enables access to real-time data from reliable online and offline sources, such as medical databases and Wikipedia, allowing the model to address newer medical terms or conditions effectively. ChatDoctor outperforms ChatGPT in terms of precision, recall, and F1 scores, particularly excelling in tasks that require up-to-date medical knowledge or domain-specific expertise.

- **Integration of General-Purpose LLMs in Healthcare Workflows.** This category covers the practical deployment of LLMs in healthcare settings, emphasizing real-world applications. Yu et al. [38] provided a comprehensive roadmap for integrating generative AI and LLMs, such as ChatGPT, into healthcare and medicine. The study explored their potential applications, including improving decision-making, automating workflows, and enhancing communication between patients and clinicians. It also examined technological advancements, such as reinforcement learning from human feedback (RLHF) and fine-tuning, along with ethical considerations like bias and privacy, as well as challenges such as hallucinations and regulatory requirements. Similarly, Toufiq et al. [32] evaluated the use of LLMs, including GPT-3.5, GPT-4, Claude, and Bard, to prioritize genes for inclusion in biomarker panels derived from large-scale molecular profiling. Their method involved using LLMs to analyze and score candidate genes based on biological and clinical relevance, followed by selecting the best candidates using a structured workflow. The study demonstrated that LLMs could effectively assist with these tasks, requiring minimal human intervention, and confirmed their utility in knowledge-driven gene prioritization for clinical and research applications.

  In addition to deployment strategies, Reddy et al. [26] proposed a comprehensive evaluation framework to assess the applicability of LLMs in healthcare, emphasizing translational value and governance. The framework introduces a layered approach that combines traditional natural language processing (NLP) metrics, such as perplexity, BLEU, and ROUGE, with evaluations of capability, utility, and adoption in real-world healthcare settings. Additionally, it incorporates a governance layer that addresses critical aspects, including fairness, transparency, trustworthiness, and accountability, to ensure the ethical and safe implementation of LLMs. Furthermore, Banerjee et al. [4] investigated the integration of LLMs with classical ML methods, such as Random Forest models, in healthcare applications. The study utilized real-world data from the National Health and Nutrition Examination Survey (NHANES) to showcase the effectiveness of Random Forest classifiers in predicting health conditions like hypertension. Additionally, it evaluated how LLMs can enhance tasks such as medical record abstraction and clinical note summarization. The authors addressed key limitations of LLMs, including bias and misinformation, and proposed hybrid models that combined classical ML with advanced language modeling to enable ethical and effective healthcare decision-making. Expanding on feasibility, Cascella et al. [5] assessed the feasibility of using ChatGPT in healthcare across four scenarios: (1) supporting clinical practice by generating structured medical notes for patients in the Intensive Care Unit (ICU), (2) contributing to scientific writing by drafting conclusions for abstracts based on the background, methods, and results sections of research papers, (3) reasoning about public health topics, such as analyzing the concept of seniority and proposing methods for assessing biological age in perioperative contexts, and (4) examining potential misuse in medicine, including both intentional and unintentional exploitation of ChatGPT in clinical and research settings. ChatGPT demonstrated its capabilities in summarizing patient data, producing structured medical notes, and drafting scientific abstracts. However, the study highlighted limitations, including ChatGPT's lack of domain-specific knowledge, its inability to establish causal relationships and significant ethical concerns regarding potential misuse.

2.2 Comparing Feature Selection Methods

Feature selection is a critical task in preparing data for ML algorithms. It aims to identify a subset of features from the original dataset that are relevant while minimizing redundancy. This process involves constructing and selecting features that improve predictive performance and enhance the interpretability of ML models [22]. Feature selection methods can be categorized based on (1) ML paradigms, such as supervised, unsupervised, and semi-supervised approaches [34], (2) domain-specific applications, such as in medical contexts where selecting relevant predictors is crucial [6, 8, 27], or (3) evaluation criteria [15]. Among these, classification based

on evaluation criteria is the most prevalent, encompassing filter, wrapper, embedded, hybrid, and ensemble methods [17, 39].

Filter methods are a category of feature selection techniques that evaluate the relevance of features independently of a specific ML model. Rather than relying on model training to evaluate feature importance, these methods utilize the statistical or mathematical properties of the data itself. Filters can be univariate, where individual features are evaluated and ranked, or multivariate, where subsets of features are analyzed for their combined relevance. These methods are computationally efficient and highly scalable, making them particularly suitable for high-dimensional datasets. However, they may fail to capture interactions between features or account for how the features affect the performance of a learning algorithm [6, 34, 39].

Wrapper methods integrate feature selection with the ML process, evaluating feature subsets based on the model's performance metrics, such as accuracy or error rate [15]. Wrappers often yield better-performing subsets than filters because the evaluation involves actual model training. However, they are computationally expensive, as the model needs to be retrained repeatedly for different feature subsets.

Embedded methods incorporate feature selection directly into the ML algorithm.[39] By leveraging the algorithm's internal properties, embedded methods guide feature evaluation during model training, offering a balance between computational efficiency and performance. Unlike wrappers, embedded methods avoid repeated classifier execution, making them faster while maintaining high-quality feature selection.

Recent advancements in feature selection include hybrid methods and ensemble methods [15]. Hybrid methods combine the strengths of multiple approaches, such as pairing filter and wrapper methods, to leverage their complementary advantages. Ensemble methods, on the other hand, aim to improve stability and robustness by applying feature selection techniques to various subsamples of the dataset and aggregating the results. This approach creates more consistent and reliable feature subsets [39].

These feature selection strategies are essential for reducing dimensionality, improving model performance, and making sure that selected features are both relevant and interpretable for specific ML applications.

3 Materials and Methods

3.1 Data Collection and Dataset Preparation

This study employs two publicly available datasets, including Surveillance, Epidemiology, End Results Program (SEER) [1] and the American College of Surgeons National Surgical Quality Improvement (NSQIP) [18]. Regarding the SEER dataset, we utilized the data records available between the years 2004 to 2013. This study focuses on four different cancers within the SEER dataset, including bladder, kidney, pancreas, and prostate, and the classification task is cancer survivability prediction as discussed in Appendix A. The second dataset is a cohort of 19,055 patients in the NSQIP data repository who underwent primary total shoulder arthroplasty (TSA) between 2016 and 2020 [18]. In this dataset, a collection of 21 predictors, including basic demographics, preoperative and intraoperative variables, plus comorbidity and laboratory results have been employed to predict 30-day unplanned reoperation following primary TSA. The description of both datasets, including predictive variables along with outcomes are presented in the Appendix A.

3.2 Proposed Approach

As outlined earlier, the primary objective of this study is to evaluate the feasibility of integrating one of the widely used large language models (LLMs), ChatGPT-4o, into the feature selection phase. This section provides a comprehensive description of our proposed approach, detailing each step involved in the process. Figure 1 illustrates the pipeline framework developed for this study.

3.2.1 ***Preprocessing.*** Preprocessing is an important step in any data analysis pipeline, aimed at cleaning and transforming raw data into a format suitable for further analysis or the development of ML models. Due to the high quality of the current experimental datasets, which are derived from our prior works detailed in [18, 29], additional data-cleaning procedures are unnecessary in this study. Consequently, the preprocessing phase is limited to tasks such as data discretization and feature anonymization.

Data Discretization: To reduce the impact of small variations in the data, and to effectively reflect the distribution of target classes, we have converted continuous data into discrete intervals. It helps create more robust models by grouping similar values into intervals and reducing the influence of minor fluctuations that allow models to learn more generalized patterns.

Feature Anonymization: We began by examining all features to identify and eliminate redundancies that could potentially affect the performance of the ML classification models. Redundant features fail to contribute new information and may unnecessarily increase the model's complexity, potentially degrading its overall efficiency and interoperability. On the other hand, including features that are perfectly correlated with the target class or label, in our case *cancer survivability* or *readmission following TSA*, can lead to overfitting, where the ML model performs well on the training data but generalizes poorly to new, unseen data. For example, in SEER datasets, two features, *vital status code* and *cause of death code*, are highly correlated with the label/outcome (e.g., survived code). Removing these perfectly correlated features is essential, as it also helps improve the ML model's robustness, reliability, interpretability, and generalization ability, leading to better performance on unseen data. Subsequently, we implemented a rigorous feature anonymization process, replacing the original feature names with generic labels such as F1, F2, and so on. This step is designed to evaluate the capability of ChatGPT-4o in feature selection under two distinct conditions: (1) using the original dataset with meaningful feature names, and (2) an anonymized version where contextual information from feature names is removed. We will test how well the LLM can find important features by running experiments on both the original and anonymized datasets. This will help us see how it performs when feature details are hidden.

3.2.2 ***Feature Selection.*** Feature selection and engineering is a crucial stage in the data science pipeline, aimed at improving the quality of features to enhance the performance of AI models. This process involves applying various techniques to transform, create,

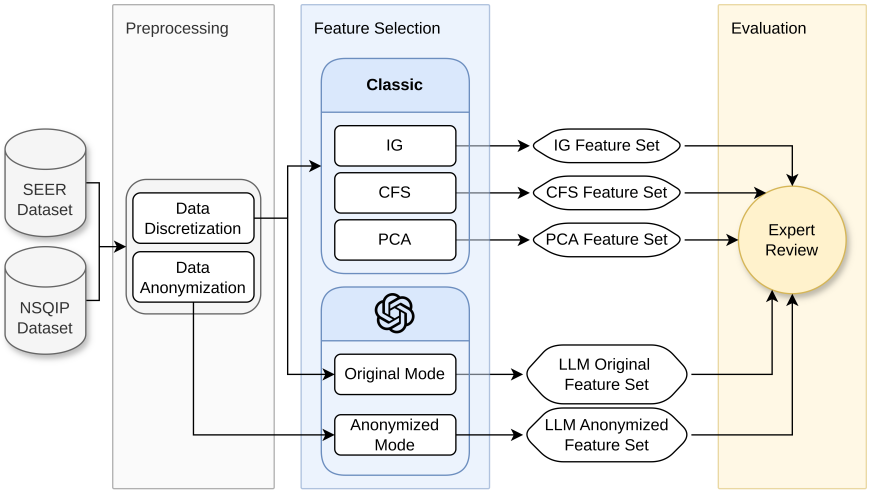

Figure 1: ***The proposed pipeline in this study illustrates the sequential stages of data processing, including preprocessing, feature selection (using both classic methods and ChatGPT-4o as an LLM-based approach), and evaluation through expert review.***

or select features that contribute to better predictive accuracy and ML model efficiency. In this section, we outline two experimental setups designed specifically for feature selection.

Classic Feature Selection: In the first approach, we aim to construct a subset of features as small as possible that represents the critical input features. We use a combination of feature selection and feature extraction methods to inherit the advantages of different metrics. These computational methods are:

- Information Gain (IG): IG serves as a univariate filter method that quantifies the information gained about a target variable by considering the values of a specific feature [16].
- Correlation-based Feature Selection (CFS): CFS is a multivariate statistical measure that assesses the strength and direction of the relationship between two variables, indicating how one variable changes with changes in another variable [13].
- Principal Component Analysis (PCA): PCA, is a popular dimensionality reduction technique that can transform high-dimensional data into a lower-dimensional space while retaining the essential information [2].

By using these methods, we generate various subsets of features, providing a robust foundation for comparing our results with those derived from the LLM.

LLM-based Feature Selection: ChatGPT has been adapted for various NLP tasks through different prompting methods. One notable method is *Instruction Prompting*, which involves giving LLMs specific task instructions to guide their responses. This method was explored in the LEAP Framework for clinical relation extraction, which combined instructional and example-based adaptive prompts to significantly improve F1 scores on clinical datasets [41]. Another innovative approach is *Chain-of-Thought Prompting*, which enhances temporal reasoning by breaking down complex tasks into smaller steps. This was demonstrated in the Grounding-Prompter method for temporal sentence grounding in long videos, which utilized a multiscale denoising chain-of-thought strategy to integrate global and local semantics for improved performance [7]. Additionally, the *Hierarchical Step-by-Step (HiSS) Prompting* method showed effectiveness in fact verification by separating claims into sub-claims and verifying each progressively, outperforming previous supervised models [40]. These methods highlight the diverse and evolving strategies in LLM prompting to enhance model accuracy and applicability.

In this study, we chose the *Instruction Prompting* method using ChatGPT-4o for several reasons. First, it aligns well with the nature of our task, which requires ChatGPT-4o to focus on specific features within the data and understand their relevance in the context of feature selection. It also provides a clear and concise direction that helps the model focus on the task at hand, ensuring that the outputs are relevant and accurate. Moreover, The versatility of *Instruction Prompting* allows for easier adaptation and fine-tuning of the prompts based on the initial responses from the LLM. This adaptability is crucial for iterative experimentation, enabling us to refine the prompts to better suit the specific requirements of our feature selection tasks.

Overall, *Instruction Prompting* offers a structured and effective approach for guiding the LLM, making it the preferred choice for our experiments in feature selection within health data science. In addition to experimenting with various prompting techniques, we perform analyses using both the original and anonymized datasets. The objective is to evaluate how the presence or absence of feature context influences the LLM's capability to identify relevant features, thereby examining its robustness and adaptability when feature metadata is obscured.

3.2.3 ***Evaluation.*** In this phase, we undertake a comprehensive evaluation of the features selected by various methods, with "domain experts review" serving as a central component. The evaluation compares the features identified by traditional techniques, including IG, CFS, and PCA, with those selected by ChatGPT-4o using both original and anonymized datasets. These are further compared against the ground truth established by the domain experts to assess the relevance, interpretability, and clinical significance of each feature within the context of the datasets and the classification task.

4 Experimental Validations and Results

This section provides a solid experimental framework that analyzes and compares the ability of an LLM, specifically ChatGPT-4o in performing feature selection as part of feature engineering. The experiments are designed to achieve two primary objectives. First, to evaluate ChatGPT-4o's effectiveness in feature selection compared to classical ML methods, including IG, CFS, and PCA. Second, to assess the impact of ChatGPT-4o's pre-trained knowledge on its feature selection performance, particularly by examining its behavior in scenarios using original data versus anonymized data.

To address these goals, we design and conduct two experiments. In the first experiment, we utilize traditional feature selection techniques, such as IG, CFS, and PCA, to identify key features from the SEER and NSQIP datasets. These methods provide a structured and data-driven baseline for comparison. In the second experiment, we leverage ChatGPT-4o capabilities to perform feature selection, analyzing its ability to rank features from the same datasets. Together, these experiments provide a comprehensive evaluation of ChatGPT-4o's performance in feature selections.

4.1 Experimental Setup and Test Bed

The experimental setup involved conducting feature selection tasks and subsequent analysis using the specified configuration. Python 3.11 was utilized as the primary programming language for implementing feature selection algorithms and performing data analysis.

The test bed comprised a series of experiments aimed at evaluating the performance of feature selection methods, particularly within the context of ML tasks. The Random Forest algorithm, configured with 100 estimators and using the Gini criterion, served as the benchmark model for comparison. The machine configuration comprises an Ubuntu 23.10 operating system running on an AMD Ryzen™ 9 5900HX processor with 16 cores and 32.0 GB of memory. For GPU capabilities, it employs an NVIDIA GeForce RTX 3060 and CUDA version 12.2. In terms of programming, Python 3.11 is the language of choice within the ChatGPT environment, accessed via WebUI.

4.2 First Experiment: Identifying Key Features Through Classical ML Methods

This experiment focuses on using well-established ML-based feature selection techniques, including IG, CFS, and PCA, to identify the most important features in the SEER and NSQIP datasets for a classification task. These methods apply systematic and data-driven approaches, each of which applies its unique criteria to evaluate and rank features based on their relevance to the predictive task. IG focuses on the information-theoretic relevance of each feature, CFS assesses the correlation between features and the target variable while minimizing redundancy, and PCA transforms the data to highlight components that capture the maximum variance. The results from all three classical ML methods serve as a baseline for comparison with AI-driven approaches using ChatGPT-4o examined in the second experiment.

For both experimental datasets, including (a) SEER, and (b) NSQIP, the results of this experiment are presented in Table 1. Initially, the top ten features were selected for each dataset based on their rankings as determined by IG, CFS, and PCA. Subsequently, the overlapping features among the top selections from these methods were identified and highlighted, showcasing the features consistently recognized as important across multiple techniques.

4.3 Second Experiment: Identifying Key Features Through ChatGPT-4o

This experiment investigates the capability of ChatGPT-4o to perform feature selection by analyzing and ranking the features in the SEER and NSQIP datasets. We also investigate whether ChatGPT-4o's feature selection relies on its pre-trained knowledge or the inherent structure of the data. Therefore, ChatGPT-4o is evaluated in two contexts: (1) using the original datasets with intact and original feature names, and (2) anonymized datasets where feature names are replaced with generic labels.

To extract important features of SEER datasets using ChatGPT-4o, we designed a structured prompt asking the model to analyze the features in each dataset and rank them based on their importance for predictive modeling. The prompt included relevant context about the dataset, the target variable (cancer survivability), and the objective of feature selection. The model then provided a ranked list of features, with the ranking reflecting its judgment of each feature's significance in predicting cancer survivability. Following this, we selected the top ten features from each of the four SEER datasets. Then, the selected features are compared across all four datasets, and those that appear in at least two of the feature sets are identified as consistently important. To test whether ChatGPT-4o's feature selection relies on its pre-trained knowledge, this process was repeated using an anonymized dataset, where feature names are replaced with generic labels (e.g., F1, F2).

After conducting experiments for both experimental datasets, SEER and NSQIP, the most important features are identified, as presented in Table 2.

4.4 Results Analysis

This section evaluates the experiment outcomes using qualitative analysis with domain expert reviews, which are explained below.

4.4.1 ***Qualitative Evaluation.*** The goal of this qualitative evaluation is to analyze and understand the most critical features identified by different feature selection methods, highlighting patterns of consistency and significance. In the first experiment, we identify the features that overlap across these methods, focusing on those consistently selected by all three. As summarized in Table 1, the performance of classical ML-based methods remains consistent across both datasets, with a significant degree of overlap in the features they prioritize. This indicates that these methods are robust

Table 1: ***The most important features identified by three classical ML-based feature selection methods, including IG, CFS, and PCA. Each method ranks the features based on its respective importance criteria, and the top ten features for each dataset are selected. The overlapping features among these top selections are then identified and presented. Checkmarks indicate features selected as important by all three methods.***

(a) ***Common important features of SEER datasets selected by IG, CFS, and PCA.***

Dataset	Age	Tumor size	Extension	Summary Stage	Regional Nodes Examined	Regional Nodes Positive	Lymph Nodes	Histologic Type
SEER Bladder	✓	✓	✓	✓		✓		
SEER Kidney	✓	✓	✓	✓			✓	✓
SEER Pancreas	✓		✓	✓	✓			
SEER Prostate		✓	✓					

(b) ***Common important features of NSQIP dataset selected by IG, CFS, and PCA.***

NSQIP	Sex	Age	Smoke	Steroid	BMI

Table 2: ***The most important features identified by ChatGPT-4o with numerical values, which indicate the rank of each feature. First, the features of each dataset are ranked based on their importance criteria. Next, the top ten features are identified for each dataset. Finally, features that appear in at least two of four selected feature sets are identified and presented as important features.***

(a) ***Common important features of SEER datasets selected by ChatGPT-4o.***

	Bladder		Kidney		Pancreas		Prostate	
Feature	Original	Anonymized	Original	Anonymized	Original	Anonymized	Original	Anonymized
Summary stage	1	1	1	1	2	1		
Lymph nodes	2	3	2	3	5	5	1	3
Regional nodes positive	3	4	5	5	7	6	3	1
Extension	4	2	3	4	3	3	5	7
Tumor size	5	5		6				5
Metastasis at diagnosis	6		4	2				
Age	7	6			6	4	4	4
Grade			6	6	1	2		
Histologic type					4	7	2	2

(b) ***Important features of NSQIP dataset selected by ChatGPT-4o***

Dataset	Mode	BMI	Age	Diabetes	Race	Ethnicity	Sex	Hypermed	Smoke
NSQIP	Original	1	2	3	4	5	6	7	8
	Anonymized	1	2	3	4	5	6	7	8

in identifying critical attributes within the datasets. In the second experiment, the model's behavior differed slightly between the two experimental datasets. For the SEER dataset, ChatGPT-4o ranked features for four types of cancer: bladder, kidney, pancreas, and prostate, with some variability observed between the original and anonymized datasets. Key features such as *Summary Stage*, *Lymph Nodes*, *Regional Nodes Positive*, and *Extension consistently* ranked among the top in both modes, with Summary Stage maintaining its position as the most important feature across all cancer types except pancreas, where it still held a high rank. However, other features, including *Tumor Size*, *Histologic Type*, and *Metastasis at Diagnosis*, showed differences in rankings between the original and anonymized datasets. For instance, *Tumor Size*, which was highly ranked in the original dataset for bladder cancer, presented a decrease in its rank in the anonymized mode, reflecting the influence of feature names on the model's prioritization process.

For the NSQIP dataset, ChatGPT-4o showed remarkable consistency in feature selection between original and anonymized

modes. Unlike the SEER dataset, which showed some variability, the NSQIP dataset results suggest that ChatGPT-4o can perform equally well without feature name metadata. This could be due to the stronger statistical signals within the NSQIP dataset or the less ambiguous nature of its features. For example, features such as *BMI*, *Age*, *Diabetes*, *Race*, and *Sex* were ranked identically in both modes, suggesting that ChatGPT-4o relied more on the inherent patterns within the data rather than the contextual information from feature names. This indicates robust performance in datasets where statistical relationships dominate over contextual cues.

In conclusion, while classical ML-based methods remain consistent across both datasets, ChatGPT-4o exhibited differing behaviors across the SEER and NSQIP datasets. The SEER dataset demonstrated some variability in feature rankings between original and anonymized modes. In contrast, the NSQIP dataset showed remarkable consistency between the two modes.

4.4.2 ***Expert Review.*** The goal of this evaluation with keeping domain experts-in-the-loop is to assess the performance of classical ML-based methods and ChatGPT-4O in the feature selection process through domain expert knowledge. We have targeted domain experts with relevant expertise in the following areas: (1) Healthcare professionals with relevant clinical expertise, and (2) Health data scientists who are experts in analyzing and interpreting complex medical data.

A panel of experts, consisting of three healthcare professionals and two health data scientists, has been invited to participate in the survey to provide their insights on the most critical features for predicting outcomes in the given datasets. For the NSQIP dataset, the experts were asked: "In your opinion, which of the features listed in the following table are the most important for predicting 30-day unplanned reoperation following primary Total Shoulder Arthroplasty (TSA)?" Similarly, for the SEER dataset, the experts addressed the question: "In your opinion, which of the features listed in the following table are the most important for predicting cancer survivability, including Bladder, Kidney, Pancreas, and Prostate cancer?"

To ensure unbiased evaluations, the experts independently assessed the features without discussing the questions or their responses with one another. Furthermore, they were not informed about the features selected by each of the feature selection methods. The features selected by different methods and by domain experts for the SEER dataset and the NSQIP dataset are presented in Tables 3 and 4, respectively. As shown in both tables, the majority of features selected by classical Ml-based methods, including IG, CFS, and PCA, are also recognized by domain experts as important and meaningful predictive variables. This alignment suggests that these methods are effective in identifying key features. For example, in Table 3, out of the nine most important features selected by all classical ML-based feature selection methods, seven features, such as *Age*, *Histologic Type*, *Summary Stage*, *Tumor Size*, *Extension*, and *Regional Nodes Positive*, were also identified as important by all the domain experts. Also, two features, such as *Lymph Nodes* and *Regional Nodes Examined*, were selected by 80% of the experts. Similarly, for the NSQIP dataset, as indicated in Table 4, the majority of features selected by classical ML-based feature selection methods were also recognized as important by the experts.

The analysis also reveals that ChatGPT-4o performs well in identifying key features, aligning closely with classical ML-based methods and expert evaluations, particularly for high-priority variables. However, its occasional variability in complex datasets (e.g., SEER) and limitations in capturing secondary but clinically relevant features suggest that it is best used as a complementary tool rather than as a standalone method for feature selection.

4.4.3 ***Analysis of Two Experiments Addressing the Research Questions.*** To address the first research question of our study, "Can ChatGPT-4o serve as an effective method for feature selection comparable to classical ML-based feature selection techniques, such as IG, CFS, and PCA??", ChatGPT-4o has been evaluated for its general applicability as a feature selection tool by comparing its performance with traditional methods such as IG, CFS, and PCA. The focus was on understanding whether the LLM could align with established approaches to select critical features in different datasets. This experiment revealed that ChatGPT-4o demonstrated strong potential as a complementary tool, particularly in identifying high-priority features that were also validated by domain expert reviews. However, its variability with less prominent features highlighted its limitations in systematic feature selection compared to the more deterministic and structured approaches of traditional ML-based methods. The results suggest that while ChatGPT-4o offers interpretive insights, it may not yet be suitable as a standalone method for feature selection, especially in high-stakes settings or clinical applications where consistency is crucial.

To address the second research question, "Does ChatGPT-4o rely on its pre-trained knowledge to identify and select important features?", we have evaluated the effect of feature anonymization on ChatGPT-4o's feature selection performance. The data anonymization experiment demonstrated that ChatGPT-4o could perform consistently well with anonymized datasets, particularly for NSQIP, where statistical signals were strong. This result suggests that the model relies less on feature names and pre-trained knowledge when data patterns are clear and unambiguous. However, in more complex datasets, such as SEER, some variability emerged, indicating that contextual information provided by feature names influenced the model's ability to rank features effectively. The two experiments provide complementary insights into ChatGPT-4o's capabilities. While the applicability experiment highlights ChatGPT-4o's potential for interpretive and exploratory feature selection, the anonymization experiment reveals its ability to generalize in privacy-preserving settings, albeit with some dependence on dataset-specific characteristics.

5 Discussion, Conclusion, and Outlook

This section further explores the implications of our current findings and outlines future directions to expand upon this work.

5.1 Key Insights and Observations

1. ChatGPT-4o demonstrates potential in identifying key features for healthcare datasets but exhibits variability and limitations in feature ranking that prevent its use as a standalone feature selection tool. The results from the SEER and NSQIP datasets reveal that ChatGPT-4o can identify key features relevant to predicting outcomes, with features like *Summary Stage*,

Table 3: ***The features selected by IG, CFS, PCA, ChatGPT-4o, and domain experts for the SEER dataset are presented, with the last column indicating the percentage of experts who identified each feature as important. The numbers under "LLM Original" and "LLM Anonymized", represent the frequency with which each feature appeared among the important features across the SEER datasets.***

Feature	IG	CFS	PCA	LLM Original	LLM Anonymized	Experts (%)
Age	✓	✓	✓	3	3	100
Histologic type	✓	✓	✓	2	2	100
Summary stage	✓	✓	✓	3	3	100
Tumor size	✓	✓	✓	2	3	100
Extension	✓	✓	✓	4	4	100
Metastasis at diagnosis		✓	✓	2	1	100
Regional nodes positive	✓	✓	✓	4	4	100
Radiation						80
Lymph nodes	✓	✓	✓	4	4	80
Regional nodes examined	✓	✓	✓			80
Grade				2	2	80
Primary site	✓					40
Survival months						0
Vital status code						0
Race						20
Site-specific surgery code	✓		✓			20
Year of diagnosis						20
Month of diagnosis						0
Cause of death						0
Marital status						0
Sex		✓				0
Behavior code						0

Lymph Nodes, and *Age* consistently ranked as important across multiple sub-datasets. While the model demonstrated robustness by maintaining similar rankings in original and anonymized versions of the datasets, variability in the importance of less prominent features highlights its reliance on implicit patterns and dataset context. For the NSQIP dataset, critical features such as *BMI*, *Age*, and *Diabetes* were consistently identified. However, ChatGPT-4o's selections occasionally lacked domain-specific reasoning. The variability and inconsistencies observed across datasets suggest that while ChatGPT-4o can provide useful exploratory insights, it is not yet suitable as a standalone tool for reliable feature selection in sensitive applications like healthcare. Its strengths lie in complementing traditional methods rather than replacing them.

2. ChatGPT-4o primarily relies on data-driven patterns for feature selection, focusing on intrinsic statistical relationships and structural properties within the dataset rather than leveraging its pre-trained knowledge. The evidence suggests that ChatGPT-4o does not heavily rely on pre-trained knowledge when feature names are anonymized, particularly for datasets with strong statistical patterns, such as NSQIP. In such cases, the model appears to prioritize features based on intrinsic data-driven patterns. These patterns involve analyzing statistical relationships, such as correlations between features and the target variable, feature variances, and inter-feature dependencies. For example, features like *BMI*, *Age*, and *Diabetes* were consistently ranked as important in both original and anonymized modes, indicating that their inherent statistical relevance within the dataset guided their selection, independent of contextual metadata.

However, the variability observed in the SEER dataset suggests that pre-trained knowledge may still play a role in contexts where datasets are more complex or feature relationships are subtle. For example, features like *Tumor Size* and *Histologic Type* showed differences in rankings between original and anonymized modes, implying that contextual cues from feature names influenced the model's prioritization in these cases. This contrast shows the delicate balance between using data-driven selection and relying on pre-trained semantic understanding in complex datasets. Thus, while it cannot be definitively stated that ChatGPT-4o is entirely independent of pre-trained knowledge, its performance demonstrates a significant reliance on data-driven patterns for feature selection. These patterns allow the model to identify important features based on their statistical properties and relationships within the data, especially in scenarios where explicit contextual information is unavailable.

3. The effectiveness of ChatGPT-4o's feature selection is influenced by the characteristics of a dataset, highlighting the importance of tailoring its application to specific data contexts. The variability in ChatGPT-4o's performance across the SEER and NSQIP datasets underscores the role of dataset characteristics in determining its effectiveness as a feature selection tool. In a dataset like NSQIP, which has clear statistical patterns and less

Table 4: ***The features selected by IG, CFS, PCA, LLM, and experts for the NSQIP dataset. The last column indicates the percentage of experts who identified each feature as important.***

Feature	IG	CFS	PCA	LLM Original	LLM Anonymized	Experts (%)
BMI	✓	✓	✓	✓	✓	100
Dialysis						100
Age	✓	✓	✓	✓	✓	80
Steroid	✓	✓	✓			80
Smoke	✓	✓	✓	✓	✓	80
Fnstatus2						80
Transfus						60
Bleedis		✓				60
Hxchf		✓				60
Diabetes	✓		✓	✓	✓	60
Dyspnea		✓	✓			40
Hxcopd	✓	✓				40
Discancr						40
Sex	✓	✓	✓	✓	✓	20
Race	✓		✓	✓	✓	0
Anesthes	✓					0
Inout						0
Hypermed		✓	✓	✓	✓	0
Wtloss						0
Ascites						0
Ethnicity			✓	✓	✓	0

ambiguous features, ChatGPT-4o performed consistently across original and anonymized modes, demonstrating its ability to prioritize features based on intrinsic data relationships. However, in the more complex SEER dataset, which involves complicated feature relationships and context-specific variables, the model showed inconsistencies in feature rankings between the original dataset and its anonymized modes. This suggests that while ChatGPT-4o excels in scenarios with well-defined data structures, its reliance on contextual metadata or pre-trained knowledge becomes more evident when handling intricate datasets. These findings highlight the need for careful consideration of dataset complexity and context when applying ChatGPT-4o for feature selection, as its strengths and limitations may vary depending on the nature of the data.

5.2 Limitation

A key limitation of this study lies in the variability of ChatGPT-4o's performance across datasets and experimental conditions. While the model showed consistency in identifying critical features in datasets with clear statistical patterns, such as NSQIP, it exhibited variability in more complex datasets like SEER, particularly when feature names were anonymized. This suggests that ChatGPT-4 may partially rely on contextual information or pre-trained knowledge to interpret feature importance, which could limit its generalizability in scenarios requiring strict anonymization or privacy.

Another limitation is the reliance on expert reviews, which, while valuable and well-established, may introduce subjectivity in assessing feature relevance. Furthermore, the absence of a systematic quantitative framework for evaluating the interpretability and robustness of ChatGPT-4o's feature selection amplifies this limitation, making it challenging to objectively compare its performance across different scenarios.

5.3 Outlook

Our future research will expand on this study by addressing several key areas to enhance the understanding and application of LLMs in feature selection. One important direction is testing other LLMs to compare their performance in feature selection tasks. This would provide valuable insights into whether the observed behaviors are specific to ChatGPT-4o or represent a broader characteristic of LLMs in general.

Additionally, integrating ChatGPT-4o with traditional feature selection methods could lead to hybrid approaches that leverage the strengths of both LLMs and data-driven techniques. For instance, combining statistical methods, such as PCA, with the interpretive capabilities of LLMs may improve both feature selection accuracy and explainability.

Finally, the development of quantitative frameworks for evaluating feature selection, such as standardized metrics for interpretability, consistency, and robustness, would provide a more objective basis for comparing LLMs (e.g., ChatGPT-4o) with traditional methods and other advanced AI-driven techniques. Together, these directions can contribute to a deeper understanding and broader applicability of LLMs in feature selection and downstream data science tasks.

6 Data Availability

All Python codes along with the supplementary materials are available at: https://github.com/pitthexai/LLMs_Explainable_Feature_Engineering. This GitHub repository is publicly and freely available for academic, research, and educational purposes.

7 Author contributions statement

B.B., S.A., A.P.T. conceived and designed the study. A.R., B.B., N.L., and S.B. implemented the ML-driven methods and developed scientific visualization components. B.B., A.R., S.B., N.L., L.R., N.M., S.A., and A.P.T. did the experimental validations and analyzed the results. All authors contributed to the interpretation of the results. B.B. and A.P.T. led the writing of this manuscript with all coauthors' comments. All authors read, reviewed, and approved the final manuscript. No competing interest is declared.

Disclosures: Ahmad P. Tafti is the Head of AI at Youki GmbH.

References

[1] [n. d.]. Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2023, based on the November 2022 submission. ([n. d.]).

[2] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

[3] A. Alanazi. 2022. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked Jan* 1 (2022), 30.

[4] Sri Banerjee, Pat Dunn, Scott Conard, and Roger Ng. 2023. Large language modeling and classical AI methods for the future of healthcare. *Journal of Medicine, Surgery, and Public Health* 1 (2023), 100026.

[5] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems* 47, 1 (2023), 33.

[6] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

[7] Houlun Chen, Xin Wang, Hong Chen, Zihan Song, Jia Jia, and Wenwu Zhu. 2023. Grounding-Prompter: Prompting LLM with Multimodal Information for Temporal Sentence Grounding in Long Videos. *arXiv.org* abs/2312.17117 (2023). https://doi.org/10.48550/arxiv.2312.17117

[8] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2021. Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 378–388.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* 24, 240 (2023), 1–113.

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[11] R. Das and DJ. Wales. 2017. Machine learning landscapes and predictions for patient outcomes. *Royal Society open science Jul* 26 (2017), 4.

[12] Zihan Guan, Zihao Wu, Zhengliang Liu, Dufan Wu, Hui Ren, Quanzheng Li, Xiang Li, and Ninghao Liu. 2023. Cohortgpt: An enhanced gpt for participant recruitment in clinical study. *arXiv preprint arXiv:2307.11346* (2023).

[13] Mark A Hall. 1999. *Correlation-based feature selection for machine learning*. Ph. D. Dissertation. The University of Waikato.

[14] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv:2304.08247 [cs.CL]

[15] Alan Jović, Karla Brkić, and Nikola Bogunović. 2015. A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, 1200–1205.

[16] John T Kent. 1983. Information gain and a general measure of correlation. *Biometrika* 70, 1 (1983), 163–173.

[17] Utkarsh Mahadeo Khaire and R Dhanalakshmi. 2022. Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences* 34, 4 (2022), 1060–1073.

[18] Christina Letter, Puneet Gupta, Annie Kim, Guang-Ting Cong, Hongfang Liu, and Ahmad P Tafti. 2023. Gender-Specific Machine Learning Models to Predict Unplanned Return to Operating Room Following Primary Total Shoulder Arthroplasty. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE, 717–721.

[19] Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine* (2024), 108013.

[20] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* 15, 6 (2023).

[21] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (09 2022), bbac409. https://doi.org/10.1093/bib/bbac409 arXiv:https://academic.oup.com/bib/article-pdf/23/6/bbac409/47144271/bbac409.pdf

[22] Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana, Elias B Khalil, and Deepak S Turaga. 2017. Learning Feature Engineering for Classification.. In *Ijcai*, Vol. 17. 2529–2535.

[23] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev. 2022. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics Nov* 1 (2022), 2.

[24] O. Rado, N. Ali, H. M. Sani, A. Idris, and D. Neagu. 2019. Performance analysis of feature selection methods for classification of healthcare datasets. InIntelligent Computing. In *Proceedings of the 2019 Computing Conference, Volume 1 . International Publishing*. Springer, 929–938.

[25] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and Qadir J. Explainable trustworthy. 2022. and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine Sep* 7 (2022). Article 106043.

[26] Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* (2023), 101304.

[27] Beatriz Remeseiro and Veronica Bolon-Canedo. 2019. A review of feature selection methods in medical applications. *Computers in biology and medicine* 112 (2019), 103375.

[28] L. Rubinger, A. Gazendam, S. Ekhtiari, and M. Bhandari. 2023. Machine learning and artificial intelligence in research and healthcare. *Injury May* 1, 54 (2023), S69–73.

[29] Tejasvi Sanjay Kamble, Hongtao Wang, Nicole Myers, Leah Reid, Cynthia S McCarthy, Young Ji Lee, Soheyla Amirian, Nickolas Littlefield, Hongfang Liu, Liron Pantanowitz, Hooman Rashidi, and AHmad P Tafti. 2024 (Under Review). Machine Learning Fairness and Explainability in Stage-Specific Cancer Survivability Prediction. In *ECCB 2024*. Oxford.

[30] K. Selvakuberan, D. Kayathiri, T. B. Harini, and MI. Devi. [n. d.]. An efficient feature selection method for classification in health care systems using machine learning techniques. In2011 3rd International Conference on Electronics Computer Technology 2011 Apr 8. *Vol.* 4,). IEEE ([n. d.]), 223–226.

[31] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[32] Mohammed Toufiq, Darawan Rinchai, Eleonore Bettacchioli, Basirudeen Syed Ahamed Kabeer, Taushif Khan, Bishesh Subba, Olivia White, Marina Yurieva, Joshy George, Noemie Jourde-Chiche, et al. 2023. Harnessing large language models (LLMs) for candidate gene prioritization and selection. *Journal of Translational Medicine* 21, 1 (2023), 728.

[33] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343* (2023).

[34] B Venkatesh and J Anuradha. 2019. A review of feature selection and its methods. *Cybernetics and information technologies* 19, 1 (2019), 3–26.

[35] P. Whig, S. Kouser, A. B. Bhatia, R. R. Nadikattu, and P. Sharma. [n. d.]. Explainable Machine Learning in Healthcare. InExplainable Machine Learning for Multimedia Based Healthcare Applications 2023 Sep 9. Springer International Publishing, Cham, 77–98.

[36] J. Wiens and ES. Shenoy. 2018. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical infectious diseases Jan* 1, 66 (2018), 1.

[37] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* 18, 6 (2024), 1–32.

[38] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. 2023. Leveraging generative AI and large Language models: a Comprehensive Roadmap for Healthcare Integration. In *Healthcare*, Vol. 11. MDPI, 2776.

[39] Rizgar Zebari, Adnan Abdulazeez, Diyar Zeebaree, Dilovan Zebari, and Jwan Saeed. 2020. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends* 1, 2 (2020), 56–70.

[40] Xuan Zhang and Wei Gao. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. *arXiv.org* abs/2310.00305 (2023). https://doi.org/10.48550/arxiv.2310.00305

[41] BMed Huixue Zhou, MS Mingchen Li, MS YongkangXiao, MS Han Yang, and Rui Zhang. 2023. LLM Instruction-Example Adaptive Prompting (LEAP) Framework for Clinical Relation Extraction. *medRxiv* (2023). https://doi.org/10.1101/2023.12.15.23300059

A Appendix

Table 5: *Features of the SEER datasets to predict cancer survivability.*

No.	Feature	Description	Data values (% frequency)
1	Marital status	Marital status at diagnosis	[2(75.7%), 1(9.9%), 4(7.5%), 5(6.2%), 3(0.7%)]
2	Race	Race	[1(85.6%), 2(8.05%), 6(1.4%), 5(1.3%), 4(0.9%), 96(0.5%), 8(0.5%), 7(0.4%), 3(0.3%), 10(0.3%), 16(0.2%), others(0.5%)]
3	Sex	The sex of the patient at diagnosis	[1(80.4%), 2(19.6%)]
4	Age	Age at diagnosis	[65(4.2%), 63(4.2%), 60(4.1%), 61(4.05%), 59(4.03%), 66(4.0%), 58(3.9%), 67(3.9%), 64(3.8%), 62(3.8%), 68(3.5%), others(56.5%)]
5	Year of diagnosis	Year of diagnosis	[2008(22.4%), 2007(21.7%), 2006(18.9%), 2005(18.6%), 2004(18.4%)]
6	Primary site	The site in which the primary tumor originated	[C619(50.9%), C649(14.8%), C250(14.1%), C679(4.2%), C678(2.7%), C252(2.3%), C672(2.2%), C659(1.6%), C251(1.2%), C674(1.1%), C258(1.0%), others(3.9%)]
7	Histologic type	Form of tumor	[8140(59.8%), 8310(8.2%), 8120(8.2%), 8500(6.3%), 8130(4.7%), 8312(2.6%), 8260(1.2%), 8550(1.2%), 8246(0.9%), 8480(0.8%), 8318(0.7%), others(5.5%)]
8	Behavior code	Code based on aggressiveness of tumor	[3(99.7%), 2(0.3%)]
9	Grade	Category based on the appearance of tumor	[3(51.0%), 2(33.7%), 4(11.05%), 1(4.3%)]
10	Site-specific surgery code	Code for surgery of primary site as first course of therapy	[50(63.9%), 37(9.8%), 61(4.8%), 30(4.3%), 60(4.1%), 71(2.5%), 40(2.05%), 70(2.05%), 36(1.9%), 0(1.1%), 35(1.0%), others(2.6%)]
11	Radiation	Method of radiation therapy used in the first course of treatment	[0(86.9%), 1(11.8%), 8(0.6%), 7(0.5%), 5(0.2%), 4(0.02%), 3(0.01%), 2(0.01%)]
12	Summary stage	Defined according to the spread of cancer	[1(46.8%), 2(45.3%), 7(7.4%), 0(0.4%)]
13	Tumor size	Tumor size in mm	[20(6.9%), 15(6.1%), 30(5.8%), 25(4.8%), 40(4.4%), 10(4.3%), 35(4.0%), 50(3.2%), 12(2.6%), 18(2.3%), 45(2.3%), others(53.2%)]
14	Extension	Information on extension of tumor	[150(27.8%), 100(10.1%), 300(8.6%), 400(7.4%), 600(6.04%), 230(5.9%), 440(5.8%), 200(5.03%), 210(4.2%), 411(3.1%), 220(3.1%), others(12.9%)]
15	Lymph nodes	The highest specific lymph node chain that is involved by the tumor	[0(77.5%), 100(17.9%), 200(1.9%), 400(1.2%), 800(0.5%), 110(0.3%), 500(0.3%), 210(0.1%), 700(0.1%), 300(0.1%), 250(0.1%), others(0.1%)]
16	Metastasis at diagnosis	Information on distant metastasis	[0(94.2%), 40(4.8%), 55(0.3%), 10(0.3%), 11(0.1%), 50(0.1%), 30(0.1%), 12(0.01%), 60(0.00%)]
17	Regional nodes positive	Number of regional lymph nodes that contained metastases	[0(77.9%), 1(8.5%), 2(4.6%), 3(2.6%), 4(1.8%), 5(1.2%), 6(0.8%), 7(0.7%), 8(0.5%), 9(0.3%), 10(0.2%), others(0.9%)]
18	Regional nodes examined	Number of regional lymph nodes removed and examined	[2(14.5%), 1(10.4%), 3(8.7%), 4(8.1%), 5(7.1%), 6(6.0%), 7(5.3%), 8(4.6%), 9(4.0%), 10(3.5%), 11(3.1%), others(24.7%)]
20	Month of diagnosis	Month of diagnosis	[3(8.8%), 7(8.7%), 10(8.6%), 4(8.5%), 11(8.3%), 1(8.3%), 6(8.3%), 5(8.3%), 12(8.2%), 9(8.1%), 8(8.04%), others(8.0%)]
21	Survival months	Survival months	Multiple categories
22	Vital status code	Vital status code	[1(69.7%), 4(30.3%)]
23	Cause of death	Cause of death to SEER site recode	[0(69.7%), 21100(16.4%), 29010(6.7%), 29020(6.7%), 28010(0.6%)]
25	Survived code (label)	Survival status (Yes/No)	[yes(69.7%), no(30.3%)]

Table 6: ***Features of the NSQIP dataset to predict 30-day unplanned reoperation following primary TSA.***

No.	Feature	Description	Data values (% frequency)
1	BMI	Body mass index (kg/m2)	Numeric values
2	Age	Age	[72(4.7%), 69(4.6%), 73(4.5%), 70(4.5%), 71(4.5%), 68(4.2%), 74(4.2%), 67(4.0%), 76(4.0%), 65(3.9%), 75(3.8%), others(53.2%)]
3	Sex	Sex	[f(55.8%), m(44.2%)]
4	Race	Race	[w(83.3%), unknown/not reported(10.8%), b(4.4%), asn(0.8%), a(0.5%), native Hawaiian or pacific islander(0.1%), others(0.03%), n(0.03%), combinations with low frequency(0.01%)]
5	Hypermed	Hypertension requiring medication	[yes(66.4%), no(33.6%)]
6	Dyspnea	Dyspnea	[no(93.2%), moderate exertion(6.5%), at rest(0.3%)]
7	Diabetes	Diabetes mellitus	[no(81.7%), non-insulin(13.0%), insulin(5.3%)]
8	Fnstatus2	Functional health status prior to Surgery	[independent(96.9%), partially dependent(1.9%), unknown(1.1%), totally dependent(0.1%)]
9	Hxcopd	History of severe chronic obstructive pulmonary disease	[no(92.9%), yes(7.1%)]
10	Smoke	Current smoker within the 12 months prior to surgery	[No:90.01%, Yes:9.99%]
11	Anesthesia	Principal anesthesia technique	[general(96.8%), regional(1.8%), mac/iv sedation(0.9%), other(0.3%), spinal(0.1%), local(0.04%), epidural(0.02%), unknown(0.02%)]
12	Inout	Inpatient or outpatient setting	[in(87.5%), out(12.5%)]
13	Bleedis	Bleeding disorders	[no(97.4%), yes(2.6%)]
14	Steroid	Steroid or immunosuppressant use for a chronic condition	[no(95.1%), yes(4.9%)]
15	Hxchf	Congestive heart failure	[no(99.2%), yes(0.8%)]
16	Discancer	Disseminated cancer	[no(99.8%), yes(0.2%)]
17	Dialysis	Currently on dialysis (pre-op)	[no(99.7%), yes(0.3%)]
18	Transfus	Transfusion >= 1 units PRBCs in 72 hours before surgery	[no(99.8%), yes(0.2%)]
19	Wtloss	10% loss body weight in last 6 months	[no(99.8%), yes(0.2%)]
20	Ascites	Ascites	[no(99.98%), yes(0.02%)]
21	Ethnicity	Ethnicity	[n(83.1%), u(12.5%), y(4.3%)]
22	Returnor (label)	Return to OR Status (Yes/No)	[no(98.6%), yes(1.4%)]