Large Language Models vs Classical Machine Learning: Predictive Power on Structured Data

Running Title:

Mohammadreza Ghaffarzadeh-Esfahani1, XX, XX, Ali Soroush, Hamidreza Hatamabadi, Ilad Alavi Darazam, Seyed Amir Ahmad Safavi-Naini \*1,2,3 , Mohamad Amin Pourhoseingholi\*1

1- Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran

2-

3- Division of Data Driven and Digital Health (D3M), The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

4-

5-

X- Department of Emergency Medicine, School of Medicine, Safety Promotion and Injury Prevention Research Center, Imam Hossein Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

X- Infectious Diseases and Tropical Medicine Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran

X- Department of Infectious Diseases, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran

\* Corresponding to: Seyed Amir Ahmad Safavi-Naini ([sdamirsa@ymail.com](mailto:sdamirsa@ymail.com)) and Mohamad Amin Pourhoseingholi (xxx)

Authors’ Detail

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Full Name | Position | Email; ORCID | Affiliation | Contribution |
| Mohammadreza Ghaffarzadeh-Esfahani |  |  | (a) Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran  (b) xxxx | Conceptualization, Methodology, Programming, Investigation, Writing Original Draft |
|  |  |  |  |  |
| Ali Soroush | MD, Assistant Professor of Gastroenterology | [ali.soroush.2012@gmail.com](mailto:ali.soroush.2012@gmail.com" \t "_blank); 0000-0001-6900-5596 | Division of Data Driven and Digital Health (D3M), The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA | Methodology, Validation |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| Hamidreza Hatamabadi |  |  | Department of Emergency Medicine, School of Medicine, Safety Promotion and Injury Prevention Research Center, Imam Hossein Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran |  |
| Ilad Alavi Darazam |  | [ilad13@yahoo.com](mailto:ilad13@yahoo.com); 0000-0002-4440-335X | (a) Infectious Diseases and Tropical Medicine Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran  (b) Department of Infectious Diseases, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran |  |
| Seyed Amir Ahmad Safavi-Naini | MD, Research Fellow | [sdamirsa@ymail.com](mailto:sdamirsa@ymail.com" \t "_blank); 0000-0001-9295-9283 | (a) Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran  (b) Division of Data Driven and Digital Health (D3M), The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA | Conceptualization, Methodology, Data Curation, Editing the Original Draft, Project Administration |
| Mohamad Amin Pourhoseingholi |  |  |  |  |

Introduction

Recent advancements in large language models (LLMs) have led to a surge in their practical applications, notably in the medical field. These models, capable of rapidly assimilating specialized knowledge across various medical domains, offer versatility in language and context adaptation, thereby broadening global access to medical expertise ([ref](https://pubmed.ncbi.nlm.nih.gov/37378099/)). LLMs, having undergone extensive training on vast datasets, excel in numerous natural language processing tasks, including language generation, machine translation, and question-answering ([2](https://arxiv.org/abs/1912.02164), [3](https://arxiv.org/abs/2110.05448), [4](https://arxiv.org/abs/1909.01066)). While their primary training focuses on predicting subsequent words, LLMs can function akin to an evidence-based knowledge assistant for practitioners, offering valuable insights and support (ref).

In the realm of medical and clinical practice, machine learning models are increasingly used for predicting patient outcomes, prognoses, and mortality rates. These models typically include supervised (e.g., classification models) and unsupervised learning (e.g., clustering algorithms) methods, primarily utilizing structured data ([5](https://pubmed.ncbi.nlm.nih.gov/33121479/)). Clinical datasets often contain a mix of structured and unstructured data, with clinical notes being a prime example of the latter.

In the domain of patient information management, traditional machine learning paradigms often adopt a bifurcated procedural approach. The initial phase involves the transformation of textual information, inherently unstructured, into a structured tabular format. The subsequent phase entails the utilization of these structured datasets for the training of various machine learning models (ref). However, this dual-phase process is frequently associated with potential pitfalls, including the loss of critical information and the introduction of complexities in model deployment, thereby posing significant challenges in the practical application of these technologies in clinical settings (ref).

The power of LLMs on using clincal text, such as discharge summaries and ward notes are previously showed in detail (ref). However, the LLM performance in handling structured data and to determine whether they outperform classic machine learning models is an area for study. This is important since much of the previous medical data is stored in a structured format. Although these structured data originate from unstructured sources, such as clinical notes, the predictive power of LLMs on tabular data is an unanswered debate ([ref](https://pubmed.ncbi.nlm.nih.gov/37516790/)).

This study seeks to address this knowledge gap by conducting a rigorous experiment on LLMs' predictive capabilities in the context of clinical outcomes. Specifically, we intend to evaluate their effectiveness in utilizing tabular structured data, which has been manually extracted from clinical notes recorded by healthcare professionals. Our second experiment is focused on evaluating the effect of missing values, and the use of LLMs as imputer. This pilot study can guide the future use of LLMs on mixed or structured-only datasets.

# Method

## Ethical Consideration

## Dataset

In this study we tested our experiments on the previously collected dataset of X patients

### Dataset: Data Collection and Outcomes

## Experiment 1: Seven CML Predictive Performance

In the first experiment, we employed five classical machine learning methods: Logistic Regression, Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbors (KNN), Random Forest plus  Neural Networks, and XGBoost. These models were implemented on our dataset to predict mortality outcomes.

To enhance the robustness of our analysis, we initially partitioned our dataset into two categories: internal validation and external validation. Subsequently, we divided the dataset  to train set and test set using a test size ratio of 0.2. This process was undertaken to ensure a rigorous evaluation of the models and to establish a clear distinction between internal and external validation for comprehensive performance assessment.

### Imputing and normalizing

To address missing values in the dataset we used iterativeimputer from skitlearn library.this method uses iterative prediction for each feature and then impute it considering Multiple Imputation by Chained Equations (MICE) method. for optimal model performance, the dataset underwent normalization using StandardScaler.these preprocessing steps were executed independently for both the training and test sets, ensuring a consistent approach in handling missing values across the experimental sets.

### Feature selection and sampling

Given the substantial number of features in our dataset (81), we strategically employed the Lasso method for feature selection due to its effectiveness in handling high-dimensional data. The Lasso method introduces regularization by adding a penalty term to the linear regression objective function, encouraging sparsity in feature coefficients. This approach proved superior to alternative methods, facilitating notable enhancements in our results. Through the application of Lasso, we derived a refined dataset that highlighted the most impactful features based on their importance, aiding in dimensionality reduction. Subsequently, we ranked and selected the top 40 features for further analysis.

To address imbalanced data in the dataset., characterized by uneven class distribution, we applied the Random undersampling technique.The successful application of Random undersampling underscored its effectiveness in mitigating imbalances within our dataset, thereby contributing to a more robust analysis of mortality prediction.

### model implamentation

we employed seven classical machine learning algorithms, namely Logistic Regression, Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbors (KNN), Random Forest, Neural Network, and XGBoost. Utilizing the grid search technique from the Scikit-Learn library, we conducted a thorough exploration of hyperparameter space, creating optimized models with a 5-fold cross-validation strategy on the dataset.

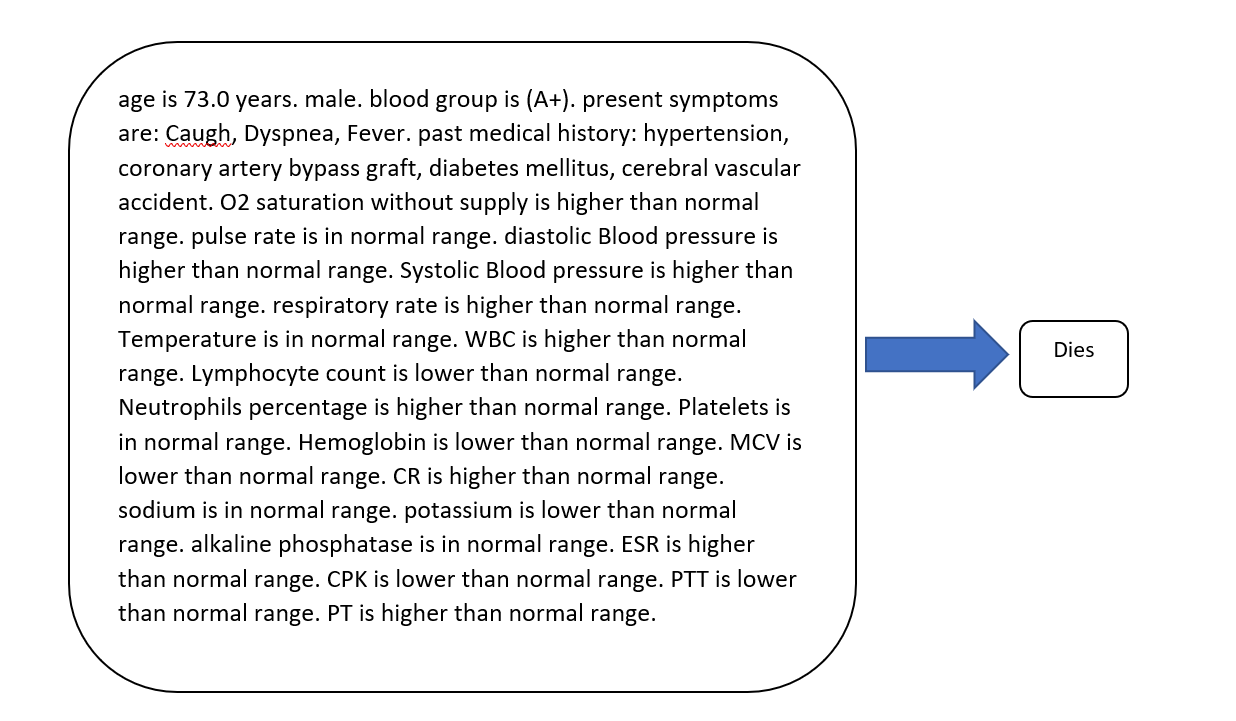
The models were evaluated using five key metrics: Specificity, Recall, Accuracy, Precision, and F1 Score. These metrics provide a comprehensive assessment of the model's performance across different aspects, ensuring a thorough examination of their effectiveness in capturing specific nuances within the dataset. The use of grid search and cross-validation enhances the reliability of our results and underscores the robustness of our model selection process.

## Experiment 1: LLMs Predictive Performance

The primary objective of this experiment is to transform our dataset into a format compatible with large language models and subsequently compare their performance with classical machine learning models. The dataset was partitioned into two categories: external validation and internal validation. As previously mentioned, we divided the dataset and our test set into two-tenths based on the test size ratio. Subsequently, KNN imputation was applied to handle missing values.

For the utilization of large language models, which inherently require text as input, a conversion process was imperative. Our original dataset, organized in tabular form, needed to be transformed into textual data to create meaningful narratives for patients. Given the inherent limitations of large language models in processing numerical data accurately, we strategically divided numerical features into categories of higher-than-normal and lower-than-normal. These categorical representations were then converted into text, facilitating a more effective integration of numerical information into the language model. This comprehensive transformation of our dataset sets the stage for a nuanced comparison between the performance of large language models and classical machine learning approaches.

(عکس از نمونه شرح حال)



### fine tuning LLMs

We aimed to improve the fine-tuning process for large language models by using recent efficient methods that save time and costs. We specifically explored QLORA, combined with the bitsanbdbytes library, to enhance language models while requiring fewer resources. Our study focused on fine-tuning the zephyr-7b-beta model, utilizing the QLORA framework with a 4-bit quantization approach.

### vector database similarity search

In this method, we first converted the dataset to embeddings using huggingfaceEmbeddings and then created a vector database  for the training set using FAISS. Then we performed a similarity search for each patient in the test set and identified the most relevant result as the patient's outcome.

### zero shot classification

We aimed to give each patient medical history individually to facebook/bart-large-mn model model and assess the model’s performance in zero shot classification

# Results

## Experiment 1: Seven CML

By using machine learning methods, we were able to achieve the results shown in the table below. The random forest model had the best performance (internal validation: Accuracy=0.76, Precision=0.79, Recall=0.71, Specificity=0.82, F1=0.75, AUC=0.76; External validation: Accuracy=0.69, Precision=0.92, Recall=0.66, Specificity=0.78, F1=0.77, AUC=0.72). To assess the efficacy of this model, we not only evaluated its performance on the test dataset but also subjected it to external validation.

## fine tuning LLMs and vector database similarity search

By fine-tuning large language models and performing vector database similarity search.We achieved the best performance among the reviewed models(Both model had internal validation: Accuracy=1, Precision=1, Recall=1, Specificity=1, F1=1, AUC=1; External validation: Accuracy=1, Precision=1, Recall=1, Specificity=1, F1=1, AUC=1).

## zero shot classification

We used face book model for zero shot classification which had worst results among the all mentioned methods(internal validation: Accuracy=0.19, Precision=0, Recall=0, Specificity=1, F1=0, AUC=0; External validation: Accuracy=0.20, Precision=0, Recall=0, Specificity=1, F1=0, AUC=0)

