scientific reports

Check for updates

OPEN

Large language models versus classical machine learning performance in COVID-19 mortality prediction using high-dimensional tabular data

Mohammadreza Ghaffarzadeh-Esfahani[1,2], Mahdi Ghaffarzadeh-Esfahani[2], Aryan Salahi-Niri[1], Hossein Toreyhi[1], Zahra Atf[3], Amirali Mohsenzadeh-Kermani[2], Mahshad Sarikhani[4], Zohreh Tajabadi[5], Fatemeh Shojaeian[6], Mohammad Hassan Bagheri[2], Aydin Feyzi[7], Mohamadamin Tarighat-Payma[4], Narges Gazmeh[7], Fateme Heydari[4], Hossein Afshar[7], Amirreza Allahgholipour[7], Farid Alimardani[7], Ameneh Salehi[4], Naghmeh Asadimanesh[4], Mohammad Amin Khalafi[4], Hadis Shabanipour[7], Ali Moradi[7], Sajjad Hossein Zadeh[7], Omid Yazdani[4], Romina Esbati[4], Moozhan Maleki[7], Danial Samiei Nasr[4], Amirali Soheili[4], Hossein Majlesi[4], Saba Shahsavan[4], Alireza Soheilipour[4], Nooshin Goudarzi[1], Erfan Taherifard[8], Hamidreza Hatamabadi[9], Jamil S. Samaan[10], Thomas Savage[11], Ankit Sakhuja[12], Ali Soroush[12], Girish Nadkarni[12], Ilad Alavi Darazam[13,14]✉, Mohamad Amin Pourhoseingholi[1,15]✉ & Seyed Amir Ahmad Safavi-Naini[1,12]✉

This study compared the performance of classical feature-based machine learning models (CMLs) and large language models (LLMs) in predicting COVID-19 mortality using high-dimensional tabular data from 9,134 patients across four hospitals. Seven CML models, including XGBoost and random forest (RF), were evaluated alongside eight LLMs, such as GPT-4 and Mistral-7b, which performed zero-shot classification on text-converted structured data. Additionally, Mistral-7b was fine-tuned using the QLoRA approach. XGBoost and RF demonstrated superior performance among CMLs, achieving F1 scores of 0.87 and 0.83 for internal and external validation, respectively. GPT-4 led the LLM category with an F1 score of 0.43, while fine-tuning Mistral-7b significantly improved its recall from 1% to 79%, yielding a stable F1 score of 0.74 during external validation. Although LLMs showed moderate performance in zero-shot classification, fine-tuning substantially enhanced their effectiveness, potentially bridging the gap with CML models. However, CMLs still outperformed LLMs in handling high-dimensional tabular data tasks. This study highlights the potential of both CMLs and fine-tuned LLMs in medical predictive modeling, while emphasizing the current superiority of CMLs for structured data analysis.

Keywords COVID-19 mortality, Large language models, Machine learning, Structured data, Zero-shot classification, Fine-tuning

Abbreviations

LLM	Large language model
CML	Classical machine learning model
LR	Logistic regression
SVM	Support vector machine
DT	Decision tree
KNN	K-nearest neighbor
RF	Random forest
XGBoost	Extreme gradient boosting
MLP	Multilayer perceptron

ZSC	Zero-shot classification
LASSO	Least absolute shrinkage and selection operator
SMOTE	Synthetic minority oversampling technique
QLoRA	Quantized low-lanking adaptation
MICE	Multiple imputation by chained equations
ReLU	Rectified linear unit
KBit	Knowledge bit
CRP	C-reactive protein
LDH	Lactate dehydrogenase
NLP	Natural language processing
CoT	Chain-of-thought

[1]Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Faculty of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran. [3]Faculty of Business and Information Technology, Ontario Tech University, Oshawa, Canada. [4]School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [5]Digestive Disease Research Institute, Tehran University of Medical Sciences, Tehran, Iran. [6]Department of Surgery, The Johns Hopkins University, Baltimore, MD, USA. [7]Student Research Committee, School of Nursing and Midwifery, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [8]MPH department, Shiraz University of Medical Sciences, Shiraz, Iran. [9]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. [10]Karsh Division of Gastroenterology and Hepatology, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, CA 90048, USA. [11]Department of Medicine, Stanford University, Stanford, CA, USA. [12]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. [13]Infectious Diseases and Tropical Medicine Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [14]Department of Infectious Diseases, Loghman Hakim Hospital, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [15]Nottingham Biomedical Research Centre, Hearing Sciences, Mental Health and Clinical Neurosciences, School of Medicine, National Institute for Health and Care Research (NIHR), University of Nottingham, Nottingham, UK. ✉email: ilad13@yahoo.com; aminphg@gmail.com; sdamirsa@ymail.com

The rapid advancement of large language models (LLMs) has revolutionized their practical applications across various domains, including medicine. These sophisticated models, trained on vast datasets, excel in a wide array of natural language processing tasks, demonstrating remarkable adaptability in assimilating specialized information from diverse medical fields[1]. While primarily designed for next-word prediction, LLMs have emerged as powerful, evidence-based knowledge assistants for healthcare providers, offering valuable insights and support in clinical decision-making processes[2–4]. While their main training centers on predicting the next word, LLMs can act as evidence-based knowledge helpers for healthcare providers, offering valuable insights and assistance[5].

In medical and clinical practice, machine learning models, particularly classical machine learning (CML) models (i.e., feature-based algorithms that learn patterns from preprocessed data rather than raw inputs), have gained significant traction in predicting patient outcomes, prognoses, and mortality rates. These models typically employ supervised and unsupervised learning methods, which primarily utilize structured data[6]. However, clinical datasets often present a complex interplay of structured and unstructured information, with clinical notes serving as prime examples of the latter. Traditionally, patient information management via machine learning has followed a two-step approach: transforming unstructured textual data into a structured format, followed by training CML models on these structured datasets. This process, however, often leads to potential information loss and introduces complexities in model deployment, hindering practical application in clinical settings[7].

While the efficacy of LLMs in handling unstructured text is well documented[8], their performance in handling structured data and their comparative effectiveness against CML models remain a critical area of investigation. This is particularly relevant given that much of the historical medical data are stored in structured formats that are often difficult to integrate[9]. Table 1 summarizes previous studies comparing the performance of LLMs and CML approaches in medicine[10–14]. Studies reported varied results, primarily due to differences in evaluated tasks (number of input features, sample size, and prediction complexity) and transformation techniques (e.g., transforming tables into textual prompts). However, they focus on tasks with a limited number of features (< 12), fail to represent real-world medical decisions, and train instances for the models (< 1000), limiting the CMLs to reach their maximum performance.

Our study aims to address this knowledge gap by evaluating LLMs' predictive capabilities in the context of COVID-19 mortality prediction via a high-dimensional dataset and simple table-to-text transformation. By utilizing a sufficient number of training instances, we provide the opportunity for CMLs to reach their maximum performance, enabling a more robust comparison with LLMs. This investigation is designed to provide insights into CML versus LLM comparisons in real-world, time-sensitive, and complex clinical tasks.

Methods

Ethical consideration

The study was approved by the Institutional Review Board (IRB) of Shahid Beheshti University of Medical Sciences (IR.SBMU.RIGLD.REC.004 and IR.SBMU.RIGLD.REC.1399.058). The IRB exempted this study from informed consent. Data were pseudonymized before analysis; patients' confidentiality and data security were prioritized at all levels. The study was completed under the Helsinki Declaration (2013) guidelines and all

experiments were performed in accordance with Iran Ministry of Health regulations. Informed consents were collected from all individuals or their legal guardians. During the generation of LLM predictions, using the OpenAI API and Poe Web interface, we opted out of training on OpenAI and used no training-use models in Poe to maintain the data safety of patient information.

Study aim and experimental summary

The objective of this research is to evaluate the efficacy of CMLs in comparison to LLMs, utilizing a dataset characterized by high-dimensional tabular data. We employed a previously compiled dataset and focused our experimental efforts on the task of classifying COVID-19 mortality. As illustrated in Fig. 1, the primary experiment encompasses the following:

- Assessment of the performance of seven CML models on both internal and external test sets.
- The assessment of eight LLMs and two pretrained language models on the test set.
- Assessment of a trained LLM's performance on both internal and external tests.

Additionally, we investigate the performance of models necessitating training (CML and trained LLM) across varying sample sizes, coupled with an elucidation of model prediction mechanisms through SHAP analysis.

Study context, data collection, and dataset

This study was conducted as part of the Tehran COVID-19 cohort, which included four tertiary centers with dedicated COVID-19 wards and ICUs in Tehran, Iran. The study period was from March 2020 to May 2023 and included two phases of data collection. The protocol and results of the first phase have been published previously. The four COVID-19 peaks during this period covered the alpha, beta, delta, and Omicron variants.

All admitted patients with a positive swab test during the first two days of admission or those with CT scans and clinical symptoms were included in the study. A medical team collected the patients' symptoms, comorbidities, habitual history, vital signs at admission, and treatment protocol through the hospital information system and reviewed the medical records. Laboratory values during the first and second days of admission were collected and organized from the hospitals' electronic laboratory records using pandas (v1.5.3), and NumPy (v1.24.1). Patients with a negative PCR result in the first two days of admission or with one missing clinical record in the HIS were excluded.

The dataset included the records of 9,134 patients with COVID-19. The data were filtered to include demographic information, comorbidities, vital signs, and laboratory results collected at the time of admission (first two days).

Computational environment

All classical machine learning (CML) experiments were performed on a workstation equipped with an Intel Core i9-12900 K CPU, 64 GB of RAM, and an NVIDIA RTX 3090 GPU (24 GB VRAM), running Ubuntu 22.04 and Python 3.10. The primary packages utilized include scikit-learn (version 1.2.2), XGBoost (version 1.7.5), pandas (version 1.5.3), and NumPy (version 1.24.1).

The fine-tuning of the Mistral-7b-Instruct model was conducted using an NVIDIA A100 80GB GPU via a cloud-based environment (Google Cloud Platform), utilizing the transformers (v4.37.2), peft (v0.9.0), and bitsandbytes (v0.41.1) libraries. The QLoRA fine-tuning procedure was implemented using 4-bit quantization, gradient accumulation steps, and mixed-precision training to optimize memory usage and reduce computational cost. All LLM zero-shot experiments were conducted via the OpenAI API and Poe interface under controlled sessions to ensure reproducibility.

Data preprocessing

Supplementary Figure S1 illustrates summary of the pipeline from raw data preprocessing through feature engineering and cleaning, to the final training–test data split used for model development and evaluation.

Imputing and normalization

The features in the dataset were divided into categorical and numerical categories. To address the missing values in the numerical features, we used an iterative imputer from the scikit-learn library. This method employs iterative prediction for each feature, considering the multiple imputation by chained equations (MICE) method (16,17). Missing values in the categorical features were imputed via KNN from the scikit-learn library. For optimal model performance, the dataset was normalized via a standard scaler (18). These preprocessing steps were executed independently for the input features of the training, test, and external validation sets, ensuring a consistent approach for handling missing values across the experimental sets without information leakage.

Feature selection

The dataset comprised 81 on-admission features. The dataset was separated into external and internal validations using patient hospitals. Patients from Hospital-4 were used for external validation, whereas patients from the remaining hospitals were used for internal validation. For internal validation, we split the data with a test size of 20% and allocated 80% for training.

The output features in this study include "in-hospital mortality," "ICU admission," and "intubation," with a focus solely on "hospital mortality" as the targeted feature, excluding other output features. Of the 81 features initially available, 76 were employed for training, comprising 53 categorical features and the remaining numerical values. During data wrangement, two duplicate features were dropped.

First author; year	Aim and task	Dataset (sample size; #feature)	Transformation techniques	Model, experiment, and metric	Zero-shot performance	Training size: performance
Hegselmann; 2023 (TabLLM)[10]	To transform table-to-text for binary classification of coronary artery disease and diabetes	Diabetes (768; #7); Heart (918, #11)	Template; Billion parameter LLM; Million parameter LLM			
				TabLLM-Diabetes (AUC)	0.82	32: 0.68 512: 0.78
				XGBoost- Diabetes (AUC)	-	32: 0.69 512: 0.80
				TabLLM-Heart (AUC)	0.54	32: 0.87 512: 0.92
				XGBoost-Heart (AUC)	-	32: 0.88 512: 0.92
Wang; 2024 (MediTab)[11]	To evaluate MediTab (GPT3.5) on seven medical classification tasks and compare it with TabLLM and CML	Seven datasets of breast, lung, and colorectal cancer from clinical trials (average 1451 ranging from 53 to 2968; average #3 categorical, #15 binaries; #7 numerical)	BioBERT-based model fine-tuned on transformation and GPT3.5 for sanity check			
				MediTab (Average AUC)	0: 0.82	200: 0.84
				XGBoost (Average AUC)	10: 0.64	200: 0.79
Cui; 2024 (EHR-CoAgent)[12]	To investigate the efficacy of LLMs-based disease prediction using structured EHR data generated from clinical encounters (MIMIC: acute care condition in the next hospital visit; CRADLE: CVD in diabetic patients)	MIMIC-III (11,353; #?); CRADLE (34,404; #?)	Disease, medicatin, and procudere codes by mapping the code value to code name (+ prompt engineering techniques)			
				EHR-CoAgent-GPT4 – MIMIC (Accuracy; F1)	0.79; 0.73	-
				GPT-4 – MIMIC (Accuracy; F1)	ZSC: 0.51; 52% Prompt engineered: 0.62; 0.58	Few-shot ($N=6$): 0.65; 0.64
				GPT-3.5 – MIMIC (Accuracy; F1)	ZSC: 0.78; 0.68 Prompt engineered: 0.72; 0.42	Few-shot ($N=6$): 0.76; 0.63
				RF – MIMIC (Accuracy; F1)	$N=6$: 0.69; 0.63	11,353: 0.78; 0.70
				LR – MIMIC (Accuracy; F1)	$N=6$: 0.48; 0.56	11,353: 0.79; 0.73
				DT – MIMIC (Accuracy; F1)	$N=6$: 0.71; 0.51	11,353: 0.81; 0.76
				EHR-CoAgent-GPT4 – CRADLE (Accuracy; F1)	0.70; 0.60	-
				GPT-4 – CRADLE (Accuracy; F1)	ZSC: 0.21; 0.22; Prompt engineered: 0.30; 0.29	Few-shot ($N=6$):0.41; 0.40
				GPT-3.5 – CRADLE (Accuracy; F1)	ZSC: 0.56; 0.52 Prompt engineered: 0.62; 0.54	Few-shot ($N=6$): 0.40; 0.40
				RF – CRADLE (Accuracy; F1)	$N=6$: 0.66; 0.51	34,404: 0.80; 0.57
				LR – CRADLE (Accuracy; F1)	$N=6$: 0.54; 0.48	34,404: 0.80; 0.59
				DT – CRADLE (Accuracy; F1)	$N=6$: 0.31; 0.31	34,404: 0.80; 0.52

Continued

First author; year	Aim and task	Dataset (sample size; #feature)	Transformation techniques	Model, experiment, and metric	Zero-shot performance	Training size: performance
Nazary; 2024 (XAI4LLM)[13]	To evaluate the diagnostic accuracy and risk factors, including gender bias and false negative rates using LLM. In addition, a comparison with CML approaches	Heart Disease Dataset (920; #11)	Using feature name-values or transforming to simple manual textual template			
				Best XAILLM: LLM + RF (F1	ZSC: 0.741	
				XGBoost (F1)	-	920: 0.91
				RF (F1)	-	920: 0.74

Table 1. Summary of studies comparing the performance of large Language models and classical machine learning methods in medicine using structured data. Cui et al.'s and Nazary et al.'s studies were preprint publications

We strategically employed the lasso method for feature selection because of its effectiveness in handling high-dimensional data. The Lasso method introduces regularization by adding a penalty term to the linear regression objective function, which encourages sparsity in the feature coefficients[15,16]. This approach proved to be 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 on the basis of their importance, aiding dimensionality reduction. We subsequently ranked and selected the top 40 features for further analyses.

Oversampling

To address the issue of class imbalance in our dataset, we employed the synthetic minority oversampling technique (SMOTE), a widely used method in machine learning, particularly for medical diagnosis and prediction tasks[17]. By applying SMOTE, we mitigated dataset imbalances, resulting in a more robust and reliable analysis for predicting mortality. SMOTE works by creating synthetic samples for the minority class instead of simply duplicating existing samples. It selects samples from the minority class and their nearest neighbors and then generates new synthetic samples by interpolating between these samples and their neighbors. This approach not only increases the number of samples in the minority class but also introduces new data points, improving dataset diversity. In our experiments, the SMOTE technique was applied to the training set (X_train), increasing the number of samples from 6118 to 9760.

Preparing data for the LLM

To prepare the data for input into the LLM, we completed all the previous steps for feature selection and sampling, but normalization was not performed. As shown in Fig. 1, we converted the dataset into text. We categorized the dataset features into symptoms, past medical history, age, sex, and laboratory data. For symptoms and medical history, we considered only positive data. For age, we added 'the patient's age is' before the age number. For sex, we used 'male' and 'female.' We used the normal range of laboratory data to classify the data into the normal range, higher than the normal range, and lower than the normal range. For example, if blood pressure and oxygen saturation were higher than the normal range, we used the sentence 'blood pressure and oxygen saturation are higher than the normal range.' We considered only laboratory data that were higher or lower than the normal range. The exclusion of negative features in symptoms and past medical history, or the normal range in laboratory data, is due to limitations in LLM context windows. We then concatenated the dataset into a single paragraph for each patient, indicating their medical history.

CML predictive performance

We employed five CML algorithms: logistic regression (LR), support vector machine (SVM), decision tree (DT), k-nearest neighbor (KNN), random forest (RF), multilayer perceptron neural network (MLP), and XGBoost. The hyperparameters were optimized via a grid search and cross-validation. The full details of training and hyperparameters are provided in **Supplementary Sect. 1.**

LLM predictive performance

We utilized open-source and proprietary LLMs to test their predictive power on clinical texts transformed from tabular data. First, we tested different prompts to determine the most efficient prompt to use, as well as the temperature (between 0.1 and 1). Full prompts are listed in Supplementary Table S1. We then sent clinical text and commands, received the unstructured output, and extracted the selected outcome, which could be either "survive" or "die." We used different sessions for each prediction, limiting the memory of the LLM to remembering previous generations.

We tested open-source, open-weight models of Mistral-7b, Mixtral 8 × 7 B, Llama3-8b, and Llama3-70b via the Poe Chat Interface. OpenAI models, including GPT-3.5T, GPT-4, GPT-4T, and GPT-4o, were utilized via the OpenAI API. We also tested the performance of two pretrained language models, BERT[18] and ClincicalBERT[19], which are fine-tuned versions of BERT on medical text. A list of all LLMs and times of use, as well as model parameters, is available in Supplementary Table S2.

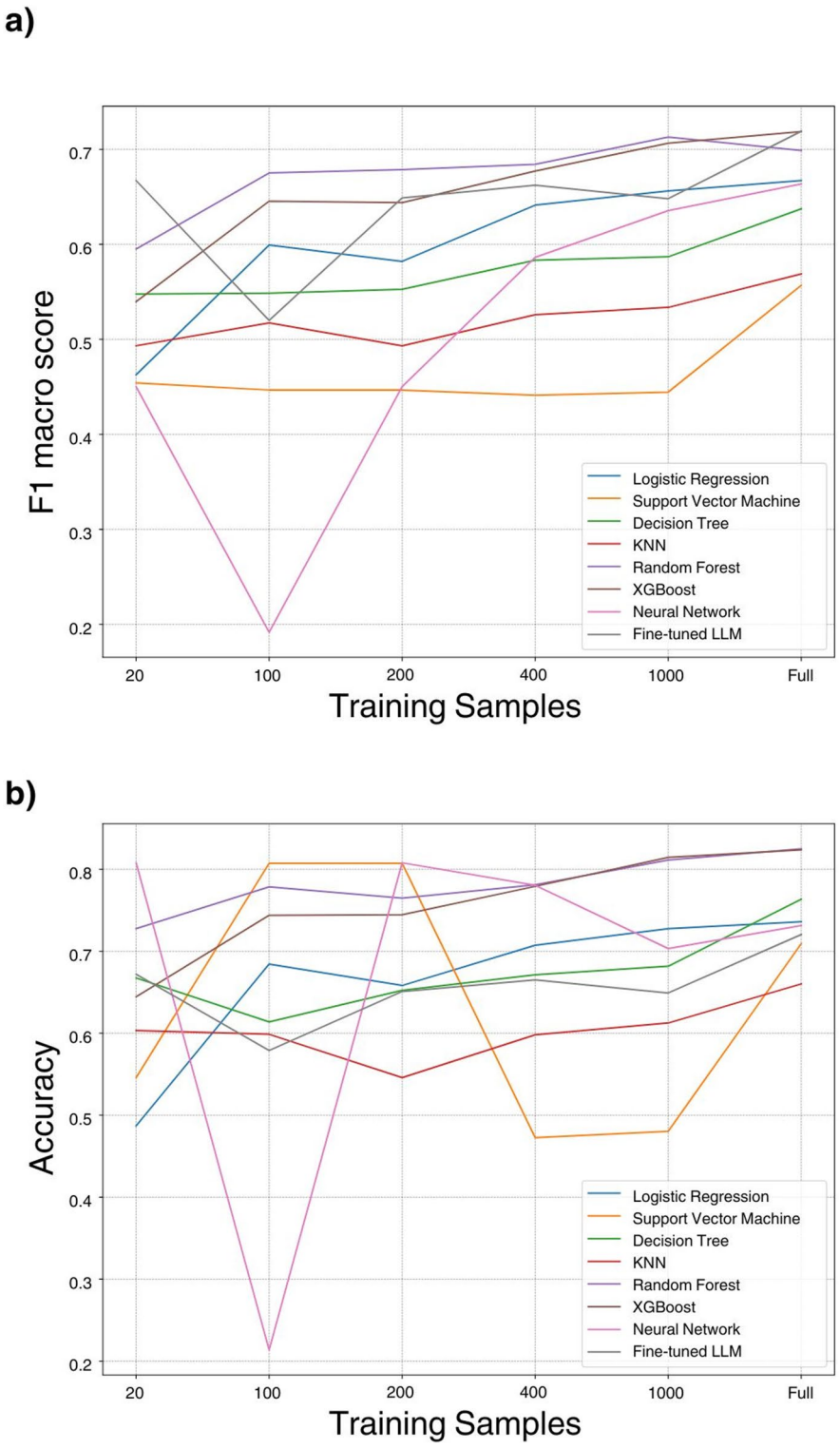

Fig. 1. Study Design and Experimental Summary. *Image caption: This figure illustrates the design and workflow of the study, which compared large language models (LLMs) and conventional machine learning (CML) approaches for the prediction of COVID-19 patient mortality. The patient data included demographics, symptoms, past medical history, and laboratory results. The data undergo preprocessing before being structured into input–output instances. The CML pipeline involves training and validation via various algorithms, such as logistic regression, support vector machine, and random forest. Moreover, the LLM pipeline involves a prompt engineering loop. We also fine-tuned one LLM, Mistral-7b, by giving the input and ground truth. The study aims to predict patient outcomes (survival or death) on the basis of the provided information.*

Zero-shot classification

Zero-shot classification is an approach in prompt engineering in which the prompt is given to the model without any training. This approach is used in transfer learning, where a model used for different purposes is employed instead of fine-tuning a new model, thereby reducing the cost of training the new model. To perform zero-shot classification, we used eight different LLMs and two LMs. We provided each patient's history as input to predict whether the patient would die or survive and then stored the results.

Fine-tuning LLM

We fine-tuned one of the open-source LLMs, Mistral-7b-Instruct-v0.2, which is a GPT-like large language model with 7 billion parameters. It is trained on a mixture of publicly available and synthetic data and can be used for natural language processing (NLP) tasks. It is also a decoder-only model that is used for text-generation tasks. Fine-tuning an LLM is usually considered time-consuming and expensive; recently, several methods have been introduced to reduce costs. We implemented the QLoRA fine-tuning approach to optimize the LLM while minimizing computational resources[20].

The model was configured for 4-bit loading with double quantization, utilizing an "nf4" quantization type and torch.bfloat16 compute data type. A 16-layer model architecture with Lora attention and targeted projection modules was employed. We used the PEFT library to create a LoraConfig object with a dropout rate of 0.1 and task type 'CAUSAL_LM'. The training pipeline, established via the transformer library, consisted of 4 epochs with a per-device batch size of 1 and gradient accumulation steps of 4. We utilized the "paged_adamw_32bit" optimizer with a learning rate of 2e-4 and a weight decay of 0.001. Mixed-precision training was conducted via fp16, with a maximum gradient norm of 0.3 and a warm-up ratio of 0.03. A cosine learning rate scheduler was employed, and training progress was logged every 25 steps and reported to TensorBoard. This methodology, which combines QLoRA with the Bitsandbytes library, enables efficient enhancement of our language model while significantly reducing resource requirements, demonstrating superior performance across various instruction datasets and model scales. A more detailed description is provided in **Supplementary Section S2**.

CML and LLM performance on different sample sizes

To investigate the influence of training sample sizes on model performance, we conducted a series of experiments using varying sample sizes: 20, 100, 200, 400, 1000, and 6118. Multiple models were trained using these sample sizes, and their performance was evaluated on the basis of the F1 score and accuracy metrics via an internal test set. The objective of this exploration was to gain valuable insights into the correlation between the volume of training data and the accuracy of predictive models.

Evaluation and cross-validation

The accuracy of the outputs was assessed by comparing them against a ground truth that categorized outcomes as either mortality or survival. Outputs from the LLM were similarly classified. If an LLM initially produced an undefined result, the prompt was repeatedly presented up to five times to elicit a defined prediction; these instances are documented in **Supplementary Table S2**. We evaluated the models' performance via five critical metrics: specificity, recall, accuracy, precision, and F1 score. To optimize our models, we employed a grid search strategy with accuracy as the primary criterion.

We further implemented 5-fold cross-validation on the training dataset ($n = 6{,}118$). The training data were randomly partitioned into five equal-sized subsets. For each fold, four subsets were used for training while the remaining subset served as a validation set. This process was repeated five times, with each subset serving as the validation set once. We calculated performance metrics (accuracy, precision, recall, specificity, F1 score, and AUC) for each fold and reported the mean and standard deviation across all five folds.

Statistical analysis

Baseline characteristics were compared between patients who died and those who survived using appropriate statistical tests based on variable type and distribution. Continuous variables were analyzed using the Mann-Whitney U test (chosen over parametric alternatives due to non-normal distributions typical of clinical data) and presented as mean ± standard deviation. Categorical variables were compared using Pearson's chi-square test. The area under the receiver operating characteristic curve (AUC) was used to illustrate the predictive capacity of each model. All statistical tests were two-sided with significance set at $P < 0.05$. Statistical analyses were performed using Python 3.12 with SciPy (v1.16.2).

Explainability

In our study, we employed SHAP (SHapley Additive exPlanations) values to examine both the total (global) and individual (granular) impacts of features on model predictions. We normalized the numerical data via a standard scaler and adopted a model-agnostic methodology. Our model-agnostic approach involved employing XGBoost as the explainer model for LLMs prediction, which was chosen for its robust performance, as demonstrated in prior research and our own findings. SHAP values provide a clear, quantitative assessment of how each feature influences individual predictions, enhancing transparency in the model's decision-making process.

For our analysis, we used the test set for each model, generated SHAP values for every prediction, and computed the mean and standard deviation of the absolute SHAP scores. We then converted SHAP scores from a range of 0 to 1 into "global impact percentages" by dividing each feature's score by the total score of all features and multiplying by 100. We calculated the average impact percentages for both CMLs and LLMs by first averaging the SHAP scores and then determining the impact percentages. To compute the standard deviation of the impact percentages, we adjusted the average standard deviation of CML/LLM via a multiplication factor derived from the ratio of the impact score to the SHAP mean. The global impact percentage represents the proportion of each

feature's impact on the predicted class across the entire dataset. A violin plot visually represents the variability of each input feature's effect on the output.

Results

Our study initially included a dataset of 9,057 patients, with a mean age of 58.40 ± 19.81 years and a male–female ratio of 1.19. The overall mortality rate in this group was 25.11% (N = 1818). Table 2 shows the distribution of variables and missing data for both the survived and mortality cohorts. We utilized an internal validation test set and an external validation set comprising 2,470 and 2,248 participants, respectively, each with a mortality rate of 50%. Additionally, the validation set for zero-shot classification included 590 patients randomly selected from the internal validation test set, with a mean age of 63.85 ± 18.37 years, a male-to-female ratio of 355:255, and a mortality rate of 50% (mortality count = 295). Table 3 details the performance metrics of all models across internal, external, and cross-validation.

Classic machine learning predictive performance

As shown in Fig. 2, XGBoost and RF were the top-performing models in terms of accuracy, achieving scores of 86.28% and 86.52%, respectively. These models also excelled in precision, recall, specificity, and F1 scores, all surpassing 85%. The MLP also delivered an acceptable performance, with an accuracy of 75.87%. When the models were applied to the external validation set, a slight decline in the AUC of 2–5% was observed. Supplementary Figures S2 and S3 depict the confusion matrix of the CMLs on the internal validation test set and the external validation set, respectively. SVM, KNN, and DT showed consistent performance across both validation sets, confirming their reliability in generalizing to unseen data.

LLM: zero-shot classification and fine-tuned mistral-7b

The zero-shot classification results showed variability among the models, with GPT-4 outperforming the other models by achieving an accuracy of 0.62 and an F1 score of 0.43 and recording the highest recall at 0.28 among the LLMs. Generally, LLMs exhibited low recall rates, predominantly classifying predictions as "mortality." The open-source models, including Llama-3-70B, Llama-3-8B, Mistral-7b-Instruct, and Mistral-8 × 7b-Instruct-v0.1, had F1 scores ranging from 0.03 (Mistral-7b) to 0.15 (Llama3-8B and Llama3-70b). Notably, the gpt-4o model showed limited effectiveness, with an F1 score of 0.01, indicating a challenge in distinguishing between true positives and true negatives. The pretrained language models – BERT and ClinicalBERT – also labeled all outcomes as dies, failing to provide predictive power. Supplementary Figure S4 shows confusion matrix of LLM and language models.

Fine-tuning Mistral-7b significantly improved its performance, increasing the F1 score from 0.03 to 0.74 in the internal test set and to 0.69 in the external test set. This fine-tuned version also demonstrated a high recall rate of 78.98%, a substantial increase from 1% in zero-shot classification, showing its ability to accurately identify a greater proportion of actual survival instances. This consistency between internal and external validations highlights the generalizability of the fine-tuned Mistral-7b in mortality prediction. The confusion matrix of fine-tuned and zero-shot Mistral-7b is presented at Supplementary Figure S5.

Comparing models on different training sample sizes

To evaluate the impact of training sample size on model efficacy, experiments were conducted across various sample sizes. Figure 3 shows that the performance of all CMLs increased as the size of the training set increased. XGBoost demonstrated the strongest performance across all categories: small (100 samples), medium (400–1000 samples), and full training set sizes (6118 samples). Notably, the MLP neural network and SVM exhibited the most significant performance improvements, with accuracies increasing from 55% with 20 training samples to 73% and 77%, respectively.

In contrast, while the zero-shot performance of GPT-4 reached an F1 score of 0.43, CMLs still surpassed both zero-shot classification and fine-tuned LLMs in predicting COVID-19 mortality. During the fine-tuning of Mistral-7b, notable performance degradation occurred in scenarios with small training sizes, leading to a loss of broader model understanding, an effect termed "negative transfer."

Explainability: impact of features on prediction

As shown in Supplementary Figure S5 while the global impact of features among CMLs exhibits similar patterns, with many of the top 10 impactful features being consistent, the granular impact differs significantly. For example, in the context of O2 saturation levels in patients, XGBoost, RF, DT, and MCP consider both high (increasing mortality risk) and low (increasing survival chance) levels to be significant, whereas KNN and LR focus only on low saturation levels. According to Fig. 4.a, the most influential features are age (11.18%) and O2 saturation (9.89%), followed by LOC (4.83%), lymphocyte count (4.79%), dyspnea (3.76%), and sex (3.68%).

Conversely, the influence of features in LLMs, particularly in lower-performing models such as Mistralb-7b and GPT4o, appears less coherent, as illustrated in Supplementary Figures S6 and S7. This inconsistency contributes to noise in the average feature impact among LLMs (Fig. 4.d). Nonetheless, age (6.58%) and O2 saturation (5.51%) remained the most significant features, with a series of laboratory tests, including neutrophil count, PT, ALP, MCV, K, Na, ESR, and Cr, revealing impacts in the 4%–5% range.

When comparing the top performers among CMLs and LLMs—XGBoost and GPT4—the patterns of global (Fig. 4.b and Fig. 4.e) and granular (Fig. 4.c and Fig. 4.f) impacts diverge, with XGBoost displaying more specific impacts and GPT4 showing broader ranges of impact.

Figure 5 illustrates how fine-tuning Mistral-7b altered the impact of features at both the global and granular levels. This refinement in prediction logic aligned the top 10 most important features more closely with those of CMLs, resulting in more equitable impact percentages among features and enhanced granularity.

Variable Demographics	Total (*N* = 9,057)	Mortality (*n* = 1,818)	Survived (*n* = 7,239)	*P*-value
Age (years)	58.4 ± 19.8	69.9 ± 23.5	55.5 ± 17.6	< 0.001***
Male gender	54.5% (4,932)	59.3% (1,078)	53.2% (3,854)	< 0.001***
Vital signs				
Diastolic BP (mmHg)	75.3 ± 11.7	74.1 ± 14.2	75.6 ± 10.8	< 0.001***
O2 Saturation without support (%)	85.1 ± 17.0	79.2 ± 18.1	86.6 ± 16.4	< 0.001***
Pulse rate (bpm)	88.0 ± 15.5	90.6 ± 19.3	87.3 ± 14.2	< 0.001***
Respiratory rate (breaths/min)	19.3 ± 5.1	20.6 ± 6.9	19.0 ± 4.4	< 0.001***
Systolic BP (mmHg)	120.4 ± 18.8	119.9 ± 23.4	120.6 ± 17.3	0.003**
Temperature (°C)	37.1 ± 1.3	37.1 ± 1.5	37.1 ± 1.3	0.334
Symptoms				
Abdominal pain	5.5% (502)	5.2% (94)	5.6% (408)	0.473
Anorexia	16.1% (1,456)	15.6% (284)	16.2% (1,172)	0.579
Chest pain	8.9% (810)	6.8% (124)	9.5% (686)	< 0.001***
Chills	26.6% (2,409)	22.0% (400)	27.8% (2,009)	< 0.001***
Cough	48.0% (4,344)	42.6% (774)	49.3% (3,570)	< 0.001***
Diarrhea	9.3% (842)	7.3% (132)	9.8% (710)	< 0.001***
Dyspnea	58.1% (5,260)	63.2% (1,149)	56.8% (4,111)	< 0.001***
Ear Pain	0.1% (12)	0.1% (1)	0.2% (11)	0.480
Fever	42.4% (3,843)	38.4% (699)	43.4% (3,144)	< 0.001***
Headache	10.5% (949)	6.0% (109)	11.6% (840)	< 0.001***
Hemiparesis	0.7% (48)	1.1% (15)	0.6% (33)	0.097
Hemorrhage	0.4% (33)	0.8% (14)	0.3% (19)	0.003**
Joint pain	1.0% (87)	1.0% (19)	0.9% (68)	0.780
Loss of consciousness	8.0% (726)	22.6% (410)	4.4% (316)	< 0.001***
Myalgia	29.0% (2,628)	19.6% (356)	31.4% (2,272)	< 0.001***
Nausea/vomiting	20.8% (1,888)	17.1% (311)	21.8% (1,577)	< 0.001***
Olfactory dysfunction	1.2% (112)	0.4% (7)	1.5% (105)	< 0.001***
Rhinorrhea	0.9% (86)	0.9% (16)	1.0% (70)	0.837
Seizure	0.5% (49)	0.8% (15)	0.5% (34)	0.095
Skin lesion	0.2% (21)	0.6% (11)	0.1% (10)	< 0.001***
Sore throat	2.5% (224)	1.5% (27)	2.7% (197)	0.003**
Weakness	36.7% (3,326)	41.9% (762)	35.4% (2,564)	< 0.001***
Comorbidities				
Alcohol use	0.6% (53)	0.9% (16)	0.5% (37)	0.095
Alzheimer's disease	1.9% (170)	5.0% (90)	1.1% (80)	< 0.001***
Anemia	1.0% (94)	1.3% (24)	1.0% (70)	0.231
Asthma	2.4% (221)	2.3% (42)	2.5% (179)	0.752
COPD	1.6% (144)	2.4% (43)	1.4% (101)	0.004**
Cancer	4.9% (446)	8.8% (160)	4.0% (286)	< 0.001***
Chronic kidney disease	3.7% (331)	6.4% (116)	3.0% (215)	< 0.001***
Current smoker	5.0% (453)	5.9% (107)	4.8% (346)	0.061
Diabetes mellitus	26.3% (2,384)	35.3% (642)	24.1% (1,742)	< 0.001***
Fatty liver disease	0.7% (61)	0.5% (9)	0.7% (52)	0.379
GI disorder	1.3% (115)	1.7% (31)	1.2% (84)	0.082
Heart failure	1.6% (146)	2.9% (53)	1.3% (93)	< 0.001***
Hepatitis C	0.1% (11)	0.2% (4)	0.1% (7)	0.248
Hookah use	0.6% (53)	0.4% (7)	0.6% (46)	0.280
Hyperlipidemia	4.6% (421)	4.4% (80)	4.7% (341)	0.618
Hypertension	32.0% (2,899)	44.3% (806)	28.9% (2,093)	< 0.001***
Immunocompromised	0.2% (19)	0.1% (2)	0.2% (17)	0.399
Ischemic heart disease	13.5% (1,224)	21.1% (384)	11.6% (840)	< 0.001***
Opium use	3.9% (352)	5.5% (100)	3.5% (252)	< 0.001***
Parkinson's disease	0.8% (76)	2.0% (37)	0.5% (39)	< 0.001***
Pregnancy	0.6% (38)	0.1% (2)	0.7% (36)	0.033*
Prior CABG	3.5% (316)	5.9% (108)	2.9% (208)	< 0.001***

Continued

Variable Demographics	Total (*N* = 9,057)	Mortality (*n* = 1,818)	Survived (*n* = 7,239)	*P*-value
Prior pneumonia	0.4% (39)	0.9% (17)	0.3% (22)	< 0.001***
Prior stroke	4.5% (404)	9.9% (180)	3.1% (224)	< 0.001***
Psychiatric disorder	1.7% (156)	2.5% (46)	1.5% (110)	0.004**
Rheumatoid arthritis	0.9% (82)	1.3% (24)	0.8% (58)	0.051
Seizure disorder	1.3% (114)	1.4% (26)	1.2% (88)	0.538
Thyroid disorder	4.5% (405)	4.1% (75)	4.6% (330)	0.462
Tuberculosis	0.3% (23)	0.2% (4)	0.3% (19)	1.000
Laboratory values				
Alkaline phosphatase (U/L)	226.4 ± 181.4	264.1 ± 223.4	215.5 ± 165.7	< 0.001***
CPK (U/L)	333.2 ± 996.3	505.5 ± 1668.4	283.7 ± 683.7	< 0.001***
Creatinine (mg/dL)	1.9 ± 2.5	2.6 ± 3.2	1.7 ± 2.3	< 0.001***
ESR (mm/hr)	58.1 ± 1387.2	40.7 ± 26.8	62.9 ± 1567.8	0.027*
Hemoglobin (g/dL)	12.4 ± 2.2	11.9 ± 2.5	12.5 ± 2.1	< 0.001***
INR	5.4 ± 22.1	7.4 ± 25.7	4.9 ± 20.8	< 0.001***
Lymphocytes (%)	17.6 ± 10.7	13.0 ± 9.8	18.8 ± 10.6	< 0.001***
MCV (fL)	84.7 ± 7.4	85.7 ± 8.0	84.4 ± 7.2	< 0.001***
Neutrophils (%)	77.8 ± 12.0	83.2 ± 10.5	76.4 ± 12.0	< 0.001***
PT (seconds)	19.1 ± 27.1	22.9 ± 32.4	17.9 ± 25.3	< 0.001***
PTT (seconds)	45.1 ± 61.7	51.8 ± 77.2	43.1 ± 56.0	< 0.001***
Platelets ($\times 10^9$/L)	201.9 ± 91.2	193.2 ± 96.2	204.2 ± 89.6	< 0.001***
Potassium (mmol/L)	7.0 ± 17.8	8.3 ± 22.0	6.7 ± 16.5	< 0.001***
Sodium (mmol/L)	137.3 ± 5.7	137.7 ± 7.3	137.2 ± 5.2	0.025*
WBC ($\times 10^9$/L)	8.3 ± 7.0	10.5 ± 7.5	7.7 ± 6.8	< 0.001
Outcomes and treatment				
Symptom to referral time (days)	6.9 ± 8.7	6.5 ± 8.0	7.0 ± 8.8	< 0.001***
Dialysis	3.3% (303)	10.1% (183)	1.7% (120)	< 0.001***
ICU admission	16.3% (1,474)	48.2% (876)	8.3% (598)	< 0.001***
Mechanical ventilation	9.7% (876)	41.7% (758)	1.6% (118)	< 0.001***

Table 2. Data are presented as mean ± standard deviation for continuous variables and n (%) for categorical variables. P-values were calculated using Mann-Whitney U test for continuous variables and chi-square test for categorical variables. *BP*, Blood pressure *CABG*, Coronary artery bypass grafting *CKD*, Chronic kidney disease *COPD*, Chronic obstructive pulmonary disease *CPK*, Creatine phosphokinase *ESR*, Erythrocyte sedimentation rate *GI*, Gastrointestinal *ICU*, Intensive care unit *INR*, International normalized ratio *MCV*, Mean corpuscular volume *PT*, Prothrombin time *PTT*, Partial thromboplastin time *WBC*, White blood cell

Pipeline validation

Supplementary Table S3 presents the XGBoost model's F1 scores for external validation, showing a result of 0.82 (AUC: 0.92) with imputation and 0.89 (AUC: 0.60) without imputation. Supplementary Table S4 presents data on external and internal validation using SMOTE to address class imbalance in CMLs. Application of SMOTE resulted in increased performance metrics for both validation sets across CMLs; for example, the XGBoost AUC rose from 0.60 to 0.92 in external validation.

Discussion

Our study reveals a notable performance gap between CML models and LLMs in predicting patient mortality via tabular data. RF and XGBoost emerged as the top CML performers, achieving over 80% accuracy and an F1 score of 0.86. In contrast, the best-performing LLM, GPT-4, achieved 62% accuracy and an F1 score of 0.43 in zero-shot classification. This disparity highlights the challenges LLMs face when dealing with purely tabular data. Notably, increasing our sample size from 5,000 patients in our previous study to 9,000 patients in this study significantly improved the performance of CML models. The AUC of RF improved from 0.82 to 0.94, underscoring the importance of large and diverse datasets in realizing the full potential of CMLs in medical tasks.

LLM performance heavily relies on the knowledge embedded within model weights, the complexity of input data, and the table-to-text transformation technique. Our approach, which uses a simple prompt and transformation to resonate with current clinical use, achieved results comparable to those of similar studies, with F1 scores of 0.50–0.60 across different medical tasks using LLMs such as the GPT-4 or GPT-3.5[10,11,21]. However, in line with many previous studies, we found that CMLs can outperform this zero-shot performance with even fewer than 100 training samples[10,13].

	Approach	Model	Accuracy	Precision	Recall	Specificity	F1	AUC
Cross validation								
	CML	LR	0.74 ± 0.02	0.91 ± 0.01	0.75 ± 0.03	0.71 ± 0.03	0.82 ± 0.02	0.73 ± 0.02
	CML	SVM	0.75 ± 0.02	0.91 ± 0.0	0.76 ± 0.02	0.72 ± 0.02	0.83 ± 0.01	0.74 ± 0.02
	CML	DT	0.72 ± 0.01	0.86 ± 0.01	0.77 ± 0.01	0.49 ± 0.03	0.81 ± 0.01	0.63 ± 0.01
	CML	KNN	0.65 ± 0.02	0.88 ± 0.01	0.64 ± 0.02	0.67 ± 0.03	0.74 ± 0.02	0.65 ± 0.01
	CML	RF	0.82 ± 0.01	0.88 ± 0.01	0.90 ± 0.02	0.50 ± 0.04	0.89 ± 0.01	0.70 ± 0.01
	CML	Xgboost	0.80 ± 0.01	0.88 ± 0.01	0.86 ± 0.02	0.55 ± 0.03	0.87 ± 0.01	0.71 ± 0.01
	CML	MLP	0.77 ± 0.02	0.88 ± 0.01	0.83 ± 0.02	0.53 ± 0.05	0.85 ± 0.01	0.68 ± 0.02
Internal validation								
	CML	LR	0.74	0.74	0.75	0.74	0.74	0.82
	CML	SVM	0.76	0.73	0.81	0.71	0.77	0.85
	CML	DT	0.76	0.76	0.76	0.76	0.76	0.76
	CML	KNN	0.69	0.71	0.66	0.72	0.68	0.74
	CML	RF	0.86	0.84	0.89	0.83	0.87	0.94
	CML	Xgboost	0.87	0.88	0.84	0.89	0.86	0.95
	CML	MLP	0.76	0.73	0.81	0.70	0.77	0.83
	Fine-tuned LLM	Fine-tuned mistral-7b	0.72	0.69	0.79	0.65	0.74	0.72
External validation								
	CML	LR	0.73	0.73	0.74	0.72	0.73	0.80
	CML	SVM	0.74	0.73	0.75	0.72	0.74	0.81
	CML	DT	0.73	0.72	0.73	0.72	0.73	0.72
	CML	KNN	0.68	0.71	0.62	0.75	0.66	0.73
	CML	RF	0.82	0.79	0.86	0.77	0.83	0.91
	CML	Xgboost	0.82	0.85	0.78	0.86	0.82	0.92
	CML	MLP	0.71	0.69	0.76	0.67	0.72	0.79
	Fine-tuned LLM	Fine-tuned mistral-7b	0.69	0.69	0.68	0.69	0.69	0.67
ZSC validation*								
	LLM	Mistral-7b	0.51	0.80	0.01	1.00	0.03	0.51
	LLM	Mixtral-8 × 7b	0.52	0.94	0.05	1.00	0.10	0.52
	LLM	Llama-3-8b	0.54	1.00	0.08	1.00	0.15	0.54
	LLM	Llama-3-70b	0.54	0.89	0.08	0.99	0.15	0.54
	LLM	gpt-3.5-turbo	0.50	0.49	0.17	0.83	0.25	0.50
	LLM	gpt-4	0.62	0.84	0.28	0.95	0.43	0.62
	LLM	gpt-4-turbo	0.57	0.82	0.19	0.96	0.31	0.57
	LLM	gpt-4o	0.50	1.00	0.00	1.00	0.01	0.50
	LM	BERT	0.50	1.00	0.00	1.00	0.01	0.50
	LM	ClinicalBERT	0.50	1.00	0.00	1.00	0.01	0.50

Table 3. Model results on the internal validation test set, external validation dataset and cross validation (The result for cross validation shows average and standard deviation across five models). * ZSC validation dataset was created using a random sample of the internal validation dataset. *AU-ROC*, Area under the receiver operating characteristic curve *ZSC,* Zero-shot classification *CML*, Classical machine learning *LLM*, Large language model *LR*, Logistic regression *SVM*, Support vector machine *DT*, Decision tree *KNN*, K-nearest neighbors *RF*, Random forest *XGBoos*t, eXtreme gradient boosting *MLP*, Multilayer perceptron

Given the performance gap between CMLs and LLMs, researchers have explored two main approaches for improving LLM performance: pipeline improvements and fine-tuning. Previous studies have shown that LLMs can close the gap in CML performance via pipeline improvements such as prompt engineering techniques (XAI4LLM), few-shot approaches (XAI4LLM, EHR-CoAgent, TabLLM), multiple runs of LLM to double-check results (EHR-CoAgent), the addition of a tree-based explainer alongside the LLM (XAILLM), or novel LLM-based text-to-table transformation (Medi-TAB, TabLLM)[21–23]. However, many of their evaluated tasks may not resonate with real-world use, as they have low-dimensional datasets (8–15 features) that do not reflect real-world complex medical data and limited sample sizes (< 500 instances in rare classes) that restrict CMLs from reaching their maximum performance.

The alternative approach, fine-tuning or in-context learning, aims to modify the model weights to teach the model a new task, which has been evaluated on name entity recognition and text extraction[24,25]. We validated this approach in our high-dimensional task, where fine-tuning Mistral increased the F1 score from 0.03 to 0.69,

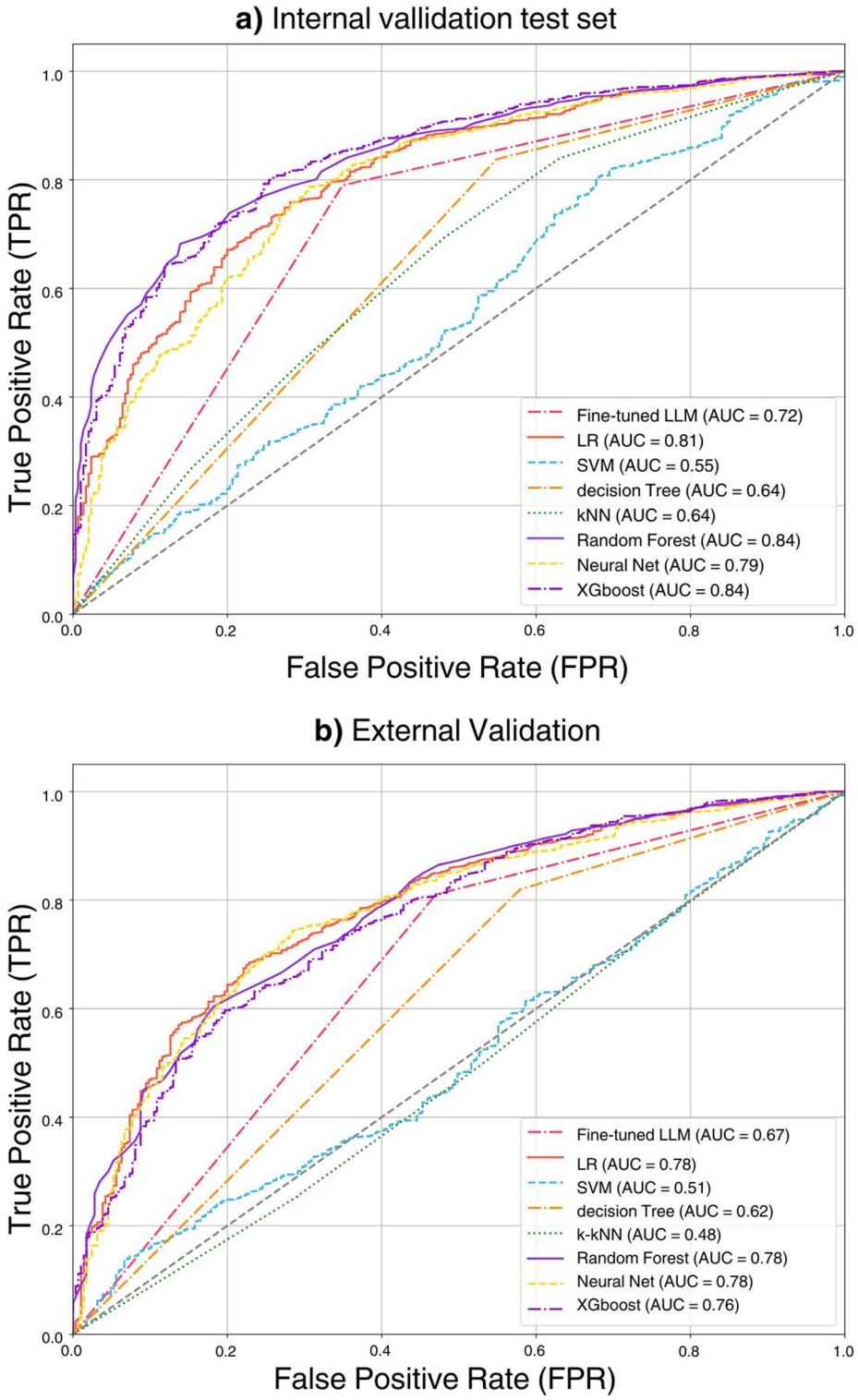

Fig. 2. ROC curves and AUC scores for COVID-19 mortality prediction models in internal and external validation. Caption: Models include fine-tuned LLM (Mistral 7b) and seven classical machine learning algorithms: logistic regression (LR), support vector machine (SVM), decision tree, k-nearest neighbors (kNN), random forest, neural network, and XGBoost. Upper panel: internal validation; Lower panel: external validation. Curves show true positive rate (TPR) versus false positive rate (FPR). AUC scores indicate each model's discriminative power.

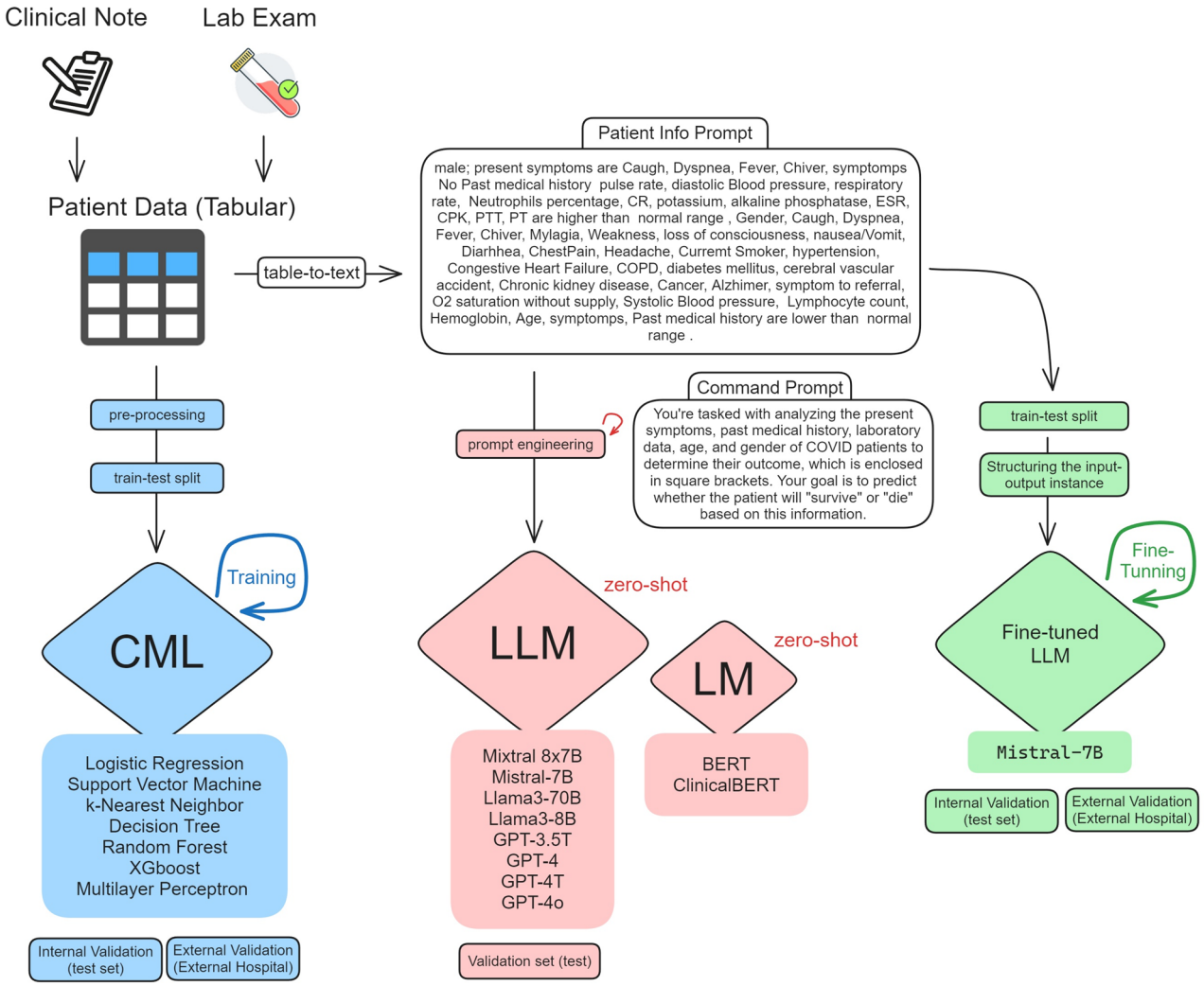

Fig. 3. Performance comparison of classical machine learning (CML) models and fine-tuned large language models (LLMs) in COVID-19 mortality prediction across varying sample sizes. F1 scores and accuracy are shown for seven CML models (logistic regression, support vector machines, decision trees, k-nearest neighbors, random forests, XGBoost, and neural networks) and a fine-tuned LLM. Training sample sizes range from 20 to 2047. XGBoost consistently outperforms other models, with performance improving as sample size increases.

even with a resource-efficient QLoRa method. Our SHAP analysis provides initial evidence of an improved rationale after fine-tuning, as the top 10 features more closely align with XGBoost and clinician decision-making.

Despite these advancements, LLMs still face significant limitations that affect their applicability in medical settings. Their vulnerability to hallucination raises concerns about producing harmful information[26], whereas computational constraints impose token limits that can truncate responses and diminish interaction quality[27,28].. Data privacy is another crucial concern, particularly in medical contexts, as many powerful LLMs are proprietary or require cloud-based computations, increasing the risk of data leaks[28]. Moreover, the cost of using LLM APIs for large clinical databases can disproportionately impact low- and middle-income communities[29]. While open-source models present a more affordable alternative, they may not match the capabilities of proprietary models.

In light of these challenges, alternative approaches have emerged, including the use of small pretrained language models and rule-based systems. These offer resource-efficient alternatives to large LLMs. Previous studies have shown that rule-based and gradient boosting algorithms can achieve strong overall performance in specific tasks, such as extracting physical rehabilitation exercise information from clinical notes[30,31]. Additionally, fine-tuning pretrained BERT-like models has yielded promising results in some medical applications. However, our brief experiment with the zero-shot performance of pretrained models (BERT and ClinicalBERT) revealed their limitations, suggesting that further research is needed to optimize these approaches for complex medical tasks.

It is important to acknowledge several limitations of our study. Although the fine-tuning method used was resource efficient, it may not have been the most effective for achieving maximum performance. Fine-tuning for conversational responses instead of classification tasks with models similar to BERT may result in less

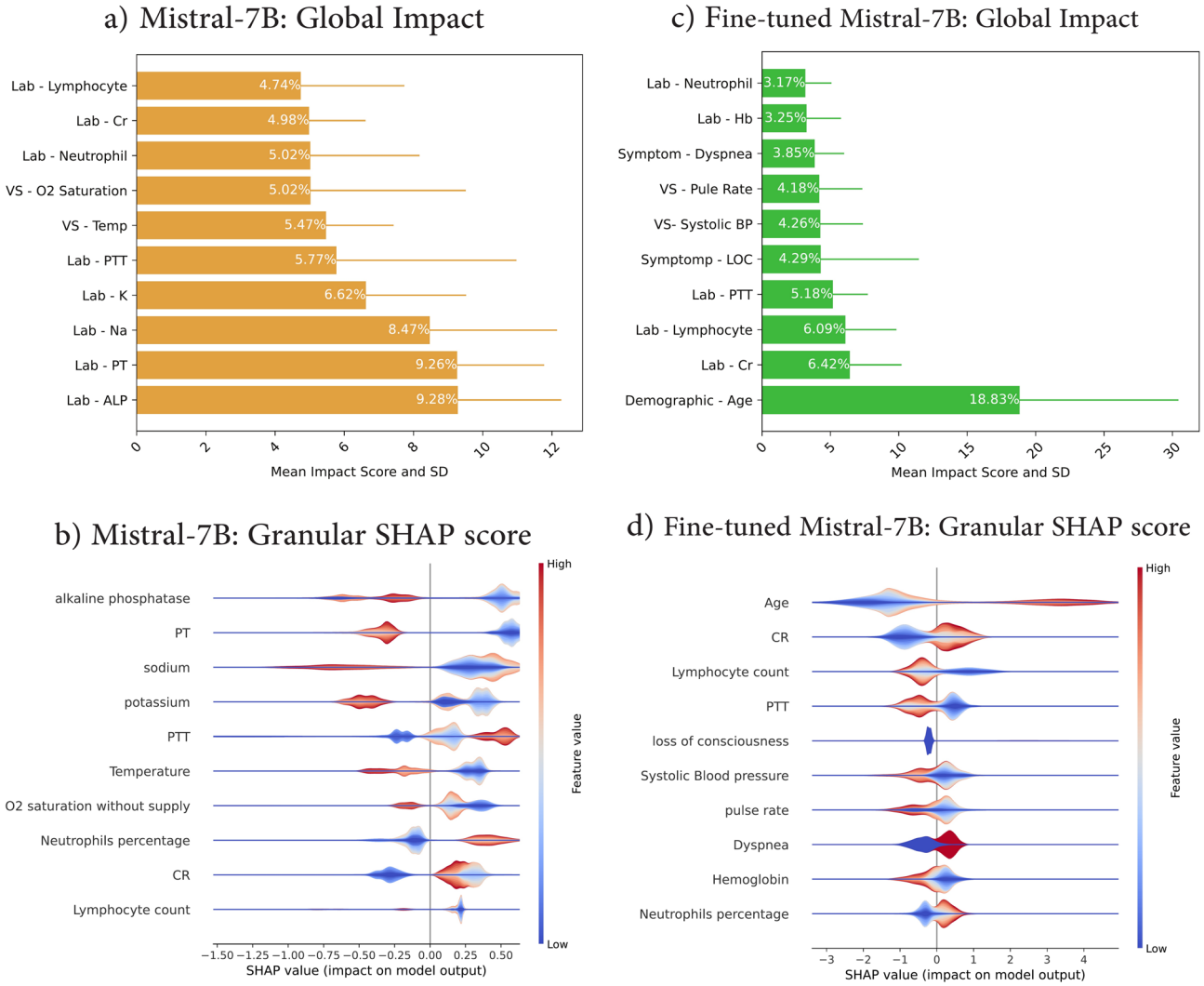

Fig. 4. SHAP analysis comparing feature importance in COVID-19 mortality prediction models. (**a**) Average global feature impact for classical machine learning (CML) models. (**b**) Global impact scores for XGBoost (best-performing CML). (**c**) SHAP score distribution for XGBoost. (**d**) Average global feature impact for large language models (LLMs). (**e**) Global impact scores for GPT-4 (best-performing LLM). (**f**) SHAP score distribution for GPT-4. Key features include age, O2 saturation (VS - O2 Sat), and creatinine (Cr) levels. In panels c and f, red indicates higher feature values; positive SHAP values increase mortality prediction, negative values decrease it. VS: vital sign.

reliable predictions; however, this approach mirrors how clinicians interact with AI tools. Future research could investigate ways to balance conversational accessibility and prediction accuracy. Our fine-tuned model was a small LLM with the lowest performance among our eight tested LLMs, indicating that fine-tuning larger and more accurate models could yield better results. Furthermore, our table-to-text transformation and prompts were designed to resonate with a medical user context, but more robust approaches (e.g., few-shot learning, advanced prompt engineering, and sophisticated transformation techniques) may achieve higher accuracies, especially in zero-shot classification[11,13]. Although our sample size was substantial, the retrospective nature of our investigation necessitates prospective validation to confirm the generalizability of these findings. As all participating hospitals operated within our specific resource context, variations in healthcare access and quality may have influenced the generalizability of the models to other countries and settings.

Our findings highlight several critical areas for future research in the application of LLMs to medical data analysis. We propose the following research questions to advance the field:

- Does the LLM explanation of the prediction (death or survival) in human language align with the feature importance analysis? Can LLMs accurately explain their rationale?
- What would be the performance of fine-tuning pretrained models and large LLMs compared to small LLMs?
- Could we create a model to distinguish correct answers from incorrect answers via LLM output? How can we measure the certainty of the given answer?

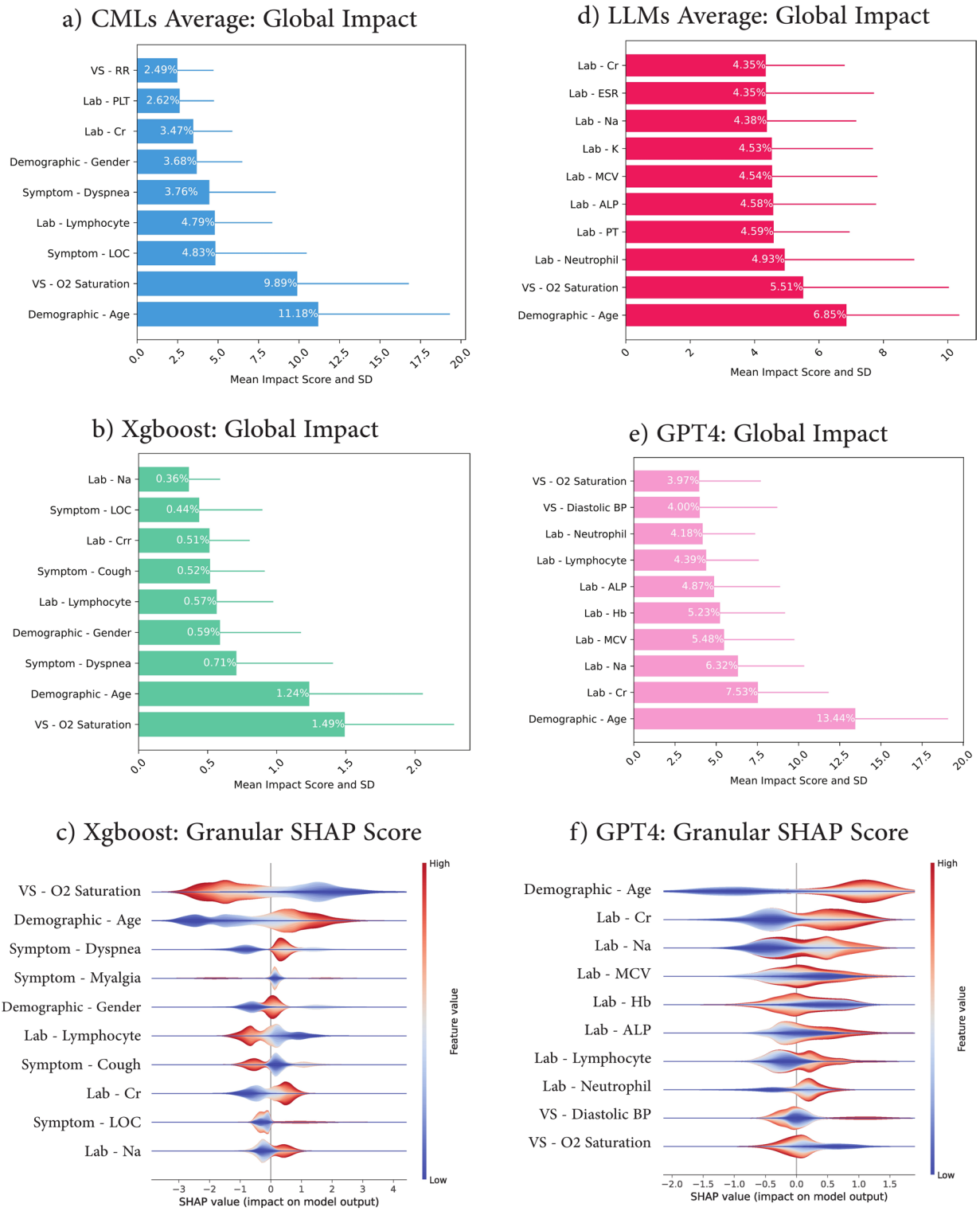

Fig. 5. SHAP analysis for Mistral-7b and fine-tuned Mistral-7b models in COVID-19 mortality prediction. (**a**) Global feature impact for base Mistral-7b, with alkaline phosphatase (ALP), prothrombin time (PT), and sodium (Na) as top features. (**b**) SHAP score distribution for base Mistral-7b, showing individual feature value influences. (**c**) Global feature impact for fine-tuned Mistral-7b, highlighting age and creatinine (Cr) as most influential. (d) SHAP score distribution for fine-tuned Mistral-7b, emphasizing the impact of age, Cr, and lymphocyte count. In panels b and d, higher SHAP values indicate increased mortality prediction.

Conclusion

The efficacy of LLMs versus CML approaches in medical tasks appears to be contingent upon data dimensionality and data availability. In low-dimensional scenarios with limited samples, LLM-based methodologies may offer superior performance; however, as dimensionality increases and diverse sample sizes become available, CML techniques tend to outperform the zero-shot capabilities of LLMs. Notably, fine-tuning LLMs can substantially enhance their pattern recognition and logical processing, potentially achieving performance levels comparable to those of CMLs. The potential of LLMs to process both structured and unstructured data may outweigh marginally lower performance metrics than CMLs do. Ultimately, the choice between LLMs and CMLs should be guided by careful consideration of task complexity, data characteristics, and clinical context demands, with further research warranted to elucidate the precise conditions under which each methodology excels.

Data availability

The code and information for generating the output are available at https://github.com/mohammad-gh009/Large-Language-Models-vs-Classical-Machine-Learning and https://github.com/Sdamirsa/Tehran_COVID_Cohort. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request (sdamirsa@ymail.com). We would welcome researchers to build upon our evaluation of LLMs in the context of using structured tabular dataset.

Received: 2 September 2024; Accepted: 30 October 2025
Published online: 28 November 2025

References

1. Karabacak, M. & Margetis, K. Embracing large Language models for medical applications: opportunities and challenges. *Cureus* https://doi.org/10.7759/cureus.39305 (2023).
2. Dathathri, S. et al. Plug and play language models: a simple approach to controlled text generation. in *8th International Conference on Learning Representations, ICLR 2020* (2020).
3. Han, J. M. et al. Unsupervised neural machine translation with generative language models only. *arXiv preprint arXiv:2110.05448* (2021).
4. Petroni, F. et al. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).
5. Vaid, A. et al. Generative Large Language Models are autonomous practitioners of evidence-based medicine. *arXiv preprint arXiv:2401.02851* (2024).
6. Zhang, D., Yin, C., Zeng, J., Yuan, X. & Zhang, P. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Med. Inf. Decis. Mak.* **20**, 1–11 (2020).
7. Sedlakova, J. et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS Digit. Health*. **2**, e0000347 (2023).
8. Zhou, H. et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).
9. Wornow, M. et al. The shaky foundations of large Language models and foundation models for electronic health records. *NPJ Digit. Med.* **6**, 135 (2023).
10. Hegselmann, S. et al. Tabllm: Few-shot classification of tabular data with large language models. in *International Conference on Artificial Intelligence and Statistics* 5549–5581 (2023).
11. Wang, Z., Gao, C., Xiao, C. & Sun, J. MediTab: Scaling Medical Tabular Data Predictors via Data Consolidation, Enrichment, and Refinement. *arXiv preprint arXiv:2305.12081* (2023).
12. Cui, H. et al. LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction. *arXiv preprint arXiv:2403.15464* (2024).
13. Nazary, F., Deldjoo, Y., Di Noia, T. & di Sciascio, E. XAI4LLM. Let Machine Learning Models and LLMs Collaborate for Enhanced In-Context Learning in Healthcare. *arXiv preprint arXiv:2405.06270* (2024).
14. Patel, D. et al. Comparative Analysis of a Large Language Model and Machine Learning Method for Prediction of Hospitalization from Nurse Triage Notes: Implications for Machine Learning-based Resource Management. *medRxiv* https://doi.org/10.1101/2023.08.07.23293699 (2023).
15. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
16. Simpson, L., Combettes, P. L. & Müller, C. L. c-lasso–a Python package for constrained sparse and robust regression and classification. *arXiv preprint arXiv:.00898* (2020). (2020). (2011).
17. Liu, X. Y., Wu, J. & Zhou, Z. H. Exploratory undersampling for Class-Imbalance learning. *IEEE Trans. Syst. Man. Cybernetics Part. B (Cybernetics)*. **39**, 539–550 (2009).
18. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *North American Chapter of the Association for Computational Linguistics* (2019).
19. Wang, G. et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat. Med.* **29**, 2633–2642 (2023).
20. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. & Qlora,. Efficient finetuning of quantized llms. *Adv Neural Inf. Process. Syst* **36**, 10088–10115 (2024).
21. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
22. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
23. Han, Z., Gao, C., Liu, J. & Zhang, S. Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).
24. Kanzawa, J., Yasaka, K., Fujita, N., Fujiwara, S. & Abe, O. Automated classification of brain MRI reports using fine-tuned large Language models. *Neuroradiology* https://doi.org/10.1007/s00234-024-03427-7 (2024).
25. Akbasli, I. T., Birbilen, A. Z. & Teksam, O. Human-Like Named Entity Recognition with Large Language Models in Unstructured Text-based Electronic Healthcare Records: An Evaluation Study. (2024).
26. O'Neill, M. & Connor, M. Amplifying Limitations, Harms and Risks of Large Language Models. *arXiv preprint arXiv:2307.04821* (2023).
27. Savage, T. et al. Large Language Model Uncertainty Measurement and Calibration for Medical Diagnosis and Treatment. *medRxiv* https://doi.org/10.1101/2024.06.06.24308399 (2024).
28. Wirth, F. N., Meurers, T., Johns, M. & Prasser, F. Privacy-preserving data sharing infrastructures for medical research: systematization and comparison. *BMC Med. Inf. Decis. Mak.* **21**, 242 (2021).

29. Gangavarapu, A. & Introducing L2M3, A Multilingual Medical Large Language Model to Advance Health Equity in Low-Resource Regions. *arXiv preprint arXiv:2404.08705* (2024).
30. Sivarajkumar, S. et al. Mining clinical notes for physical rehabilitation exercise information: natural Language processing algorithm development and validation study. *JMIR Med. Inf.* **12**, e52289 (2024).
31. Chen, S. et al. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J. Am. Med. Inf. Assoc.* **31**, 940–948 (2024).

Acknowledgements
We used ChatGPT with the following prompt: "Is this paragraph grammatically correct, and can you make it sound scientific? Improve the grammar to improve the English style, understanding, and coherence". Two authors, SAASN and MG, reviewed the suggestions and accepted relevant changes. All the authors are responsible for the validity of the final draft.

Author contributions
MoGE: Conceptualization, Methodology, Programming, Investigation, Writing Original Draft; MaGE: Investigation, Methodology; ASN: Investigation, Methodology; HT: Writing Original Draft, Methodology; ZA: Investigation, Programming; AMK: Investigation; MS: Investigation; ZT: Investigation; FS: Investigation; MHB: Investigation; AF: Investigation; MT: Investigation; NG: Investigation; FH: Investigation; HA: Investigation; AA: Investigation; FA: Investigation; AS: Investigation; NA: Investigation; MAK: Investigation, Project Administration; HS: Investigation; AM: Investigation; SHZ: Investigation; OY: Investigation; RE: Investigation; MM: Investigation; DSN: Investigation; ALS: Investigation; HM: Investigation; SS: Investigation; ARS: Investigation; NG: Investigation; ET: Investigation, Validation; HH: Investigation; JSS: Reviewing and Editing the Manuscript; TS: Reviewing and Editing the Manuscript; AKS: Reviewing and Editing the Manuscript; ALS: Methodology, Validation, Reviewing and Editing the Manuscript; GN: Reviewing and Editing the Manuscript; IAD: Data Acquisition, Reviewing and Editing the Manuscript, Administration, Supervision; MAP: Data Acquisition, Reviewing and Editing the Manuscript, Administration, Supervision, Validation; SAASN: Conceptualization, Methodology, Programming, Data Curation, Writing and Editing the Original Draft, Project Administration, Supervision.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-26705-7.

Correspondence and requests for materials should be addressed to I.A.D., M.A.P. or S.A.A.S.-N.