Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 363-372

Integrating tabular data and textual representations for clinical risk prediction using machine learning and large language models

M.Rafly Rahman[1], Setio Basuki[1], Muhammad Ilham Perdana[1], La Febry Andira Rose Cynthia*[2]
Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia[1]
Electrical Engineering, Universitas Muhammadiyah Malang, Indonesia[2]

Article Info
Keywords:
Medical Tabular Data, Large Language Models (LLM), Clinical Risk Prediction, Data Serialization, Few-shot Learning

Article history:
Received: November 07, 2025
Accepted: February 08, 2026
Published: May 01, 2026

Cite:
M. Rahman, S. Basuki, M. I. Perdana, and L. F. A. R. Cynthia, "Integrating Tabular Data and Textual Representations for Clinical Risk Prediction Using Machine Learning and Large Language Models", *KINETIK*, vol. 11, no. 2, May 2026.
https://doi.org/10.22219/kinetik.v11i2.2570

*Corresponding author.
La Febry Andira Rose Cynthia
E-mail address:
lafebryarc@umm.ac.id

Abstract
Global health is currently facing serious challenges due to the increasing number of chronic disease patients, such as those with heart failure, diabetes, and cancer. This issue arises from the limitations of electronic health record (EHR) systems, which are not yet fully capable of ensuring accurate clinical diagnoses because of potential data input errors and delays in symptom identification by medical personnel. In response to this issue, this paper focuses on the integration of medical tabular data with a classification approach based on classical machine learning (ML) and large language models (LLM) to improve the accuracy of patient diagnosis predictions. This paper aims to develop and compare the performance of various ML models, such as XGBoost, SVM, and logistic regression, as well as LLM models like Gemini, LLaMA, and Qwen in fine-tuning, few-shot, and zero-shot scenarios. The paper results show that the combination of Gemini and the few-shot approach (250 shots) achieved the highest accuracy of up to 99.8% in predicting heart failure risk. The main finding of this study is that the narrative text representation of tabular data processed with LLM significantly enhances contextual understanding and classification accuracy, making this approach highly potent for application in AI-based clinical decision-making.

1. Introduction

Global health is currently facing a serious crisis driven by the escalating number of deaths from chronic diseases. According to the World Health Organization (WHO), heart failure, diabetes, and cancer are among the world's most fatal illnesses, with their prevalence consistently rising. In 2022, the WHO reported that the number of global diabetes patients reached 828 million, with an estimated 3.4 million deaths globally attributed to the disease [1]. Similarly, data from the Global Cancer Observatory showed that cancer affected 20 million people, leading to an estimated 9.7 million deaths [2]. While the World Heart Federation estimated 64 million heart failure patients, of whom approximately 50% die within five years of diagnosis [3]. These alarming statistics highlight a critical challenge that requires an urgent response from the global health system. However, most healthcare services still rely on Electronic Health Records (EHR) which are susceptible to human errors like mistyping and data duplication, negatively affecting patient care [4],[5],[6]. This creates a significant gap between the need for accurate medical data and the capacity of conventional systems to manage it effectively, directly impeding the successful treatment of chronic diseases.

Amidst these challenges, the emergence of new technologies offers new hope. Artificial intelligence (AI) has emerged as a new pioneer aimed at helping to overcome weaknesses in modern healthcare systems. The use of AI brings great opportunities such as improved diagnostic accuracy, improved medical processes, and reduced workload for healthcare workers [7],[8]. In particular, machine learning models have demonstrated remarkable capabilities in processing large amounts of medical data, enabling faster and more efficient identification of disease patterns and symptoms [9],[10],[11]. This technology is also capable of overcoming the weaknesses of traditional Electronic Health Records (EHR) systems through data entry automation, inconsistency detection, and the provision of real-time decision-making support for medical personnel [12]. WithOne of the latest achievements of this evolution is the emergence of Large Language Models (LLMs), such as GPT, BERT, and BioBERT, which are designed with large-scale natural language understanding and generation capabilities. LLMs are the result of rapid advances in deep learning and Natural Language Processing (NLP), with transformer architectures that enable models to understand complex linguistic contexts and generate coherent text [13]. In the medical context, LLMs play an important role in extracting information from electronic health records, analyzing clinical reports, answering text-based medical questions, and supporting intelligent clinical decision-making processes [14],[15]. This development marks a significant transition from numeric data-based analysis to semantic understanding of textual data, expanding the scope of AI application in healthcare and strengthening the foundation for more adaptive, efficient, and data-driven healthcare systems.

Although artificial intelligence has shown significant progress in healthcare, fundamental challenges remain regarding its ability to produce accurate clinical risk classifications from complex patient histories. Many existing studies still rely on classical machine learning applied directly to structured medical tabular data, which limits the capture of richer clinical context and may lead to mismatches between model outputs and real clinical conditions due to data bias and limited contextual understanding [16],[17],[18].Previous natural language processing approaches primarily focus on unstructured clinical text and rarely explore the systematic serialization of medical tabular data into narrative representations. Moreover, recent studies involving large language models (LLMs) in healthcare often evaluate a single learning strategy without comprehensive comparison across fine-tuning, zero-shot, and few-shot scenarios. Consequently, a unified comparative evaluation of classical machine learning, classical NLP, and LLM-based approaches on serialized medical tabular data remains insufficiently explored.

This paper aims to explore approaches to medical data classification using classical machine learning and large language models (LLMs) on chronic disease datasets, including heart failure, cancer, and diabetes. The first approach utilizes medical tabular data that is analyzed directly using classic machine learning classifiers, such as logistic regression, Naive Bayes, support vector machine (SVM), K-nearest neighbors (KNN), and XGBoost, with performance evaluated through accuracy, precision, recall, and F1-score metrics. This tabular analysis allows for the identification of patterns of relationships between clinical attributes, such as age, gender, blood pressure, and laboratory results, and provides a baseline for model performance before further transformation. Next, the tabular data is converted into narratives using serialization techniques, resulting in textual descriptions that describe the patient's condition more comprehensively, including clinical attributes and vital parameters [19]. This text representation is evaluated using NLP methods, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, and Word2Vec embedding, which are then fed into a classical machine learning classifier. Furthermore, LLM is applied as a natural language-based classification system trained to understand serialized patient narratives, enabling it to recognize contextual relationships between medical attributes in a format resembling natural clinical documentation. With systematically organized narratives, LLM is expected to be able to classify the risk of chronic diseases in a yes/no format while providing clinical intervention recommendations. Thus, this paper not only tests the performance of machine learning-based classification models but also explores the potential of LLM as a natural language-based classification system that is capable of understanding clinical contexts more deeply and supporting medical decision-making. This paper delivers several contributions as follows:

- This study proposes an integrated diagnostic framework that combines classical machine learning and large language models (LLMs) by transforming medical tabular data into narrative text using BoW, TF-IDF, Word2Vec, and N-gram representations. This approach enables a comparative evaluation between tabular-based and text-based clinical risk prediction methods.
- Experimental results show that the Gemini LLM with a few-shot learning strategy achieves the best performance, reaching 99.8% accuracy in heart failure risk prediction using 250 training examples.
- The proposed approach is validated on multiple global priority disease datasets, including heart failure, diabetes, and cancer, in accordance with World Health Organization (WHO) reports.
- The findings demonstrate the potential of LLM-based few-shot learning on serialized medical data to support practical clinical decision-making systems.

2. Medical Data Prediction Method

The medical data prediction system is implemented through several stages, namely data collection, Classification on Tabular Dataset, Classification on Classical NLP Method, Classification using Large Language Model, and the Learning Scenario stage of the Large Language Model. The architecture of the proposed system is illustrated in Figure 1.

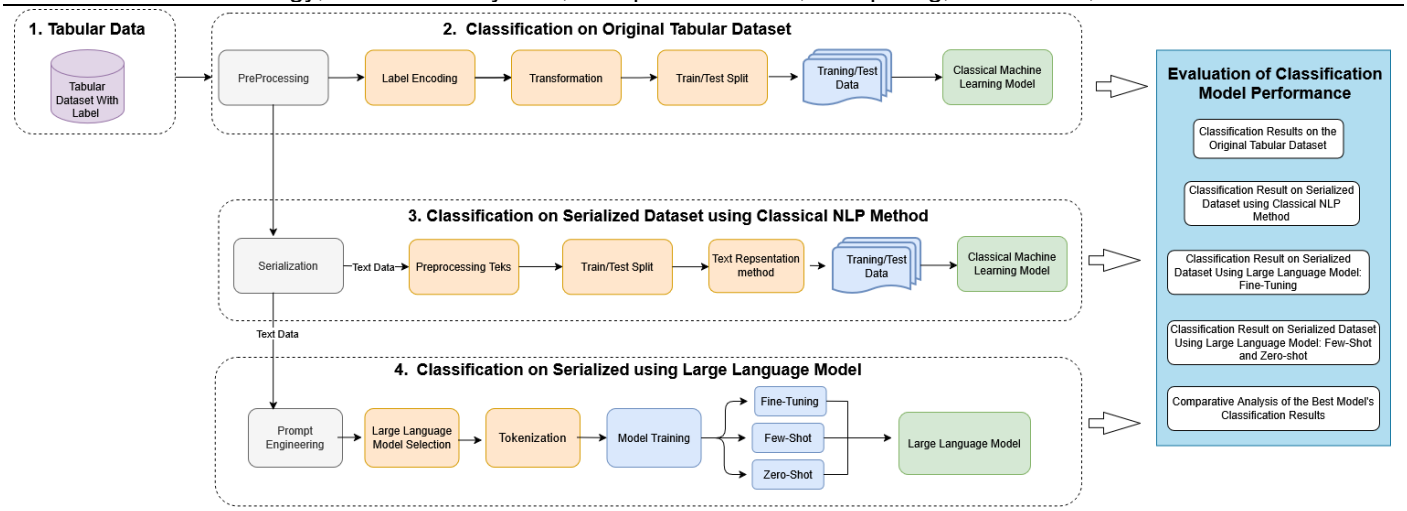

Figure 1. Framework of Tabular Dataset Classification using ML and LLM

2.1 Dataset of Medical Diseases

This paper utilizes tabular data-based classification in a medical context. Tabular data, common in systems like Electronic Health Records (EHR), is organized in rows and columns that represent individuals and clinical or demographic variables, including age, blood pressure, and medical history[19]. Each dataset also contains target labels such as disease status or severity level. Three datasets from Kaggle were used: "Cancer Prediction Dataset" (1,500 samples, 9 attributes), "Diabetes Dataset" (768 samples, 9 attributes), and "Heart Failure Prediction" (299 samples, 13 attributes). The selection refers to a WHO report that ranks these three diseases as the leading global causes of death [20]. Each dataset contains different clinical characteristics relevant for diagnosis and prognosis, such as age, gender, smoking status, and medical test results. Tables 1, 2, and 3 present the variables used in the respective datasets.

Table 1. Preview of Diabetes Dataset

Age	Gender	BMI	Smoking	GeneticRisk	Physical Activity	Alcohol Intake	Cancer History	Diagnosis (Target Class)
58	1	16.0853	0	1	8.1462	4.1482	1	1
71	0	30.8287	0	1	9.3616	3.5196	0	0
……	…….	…….	…….	…….	…….	…….	…….	…….
67	1	23.6631	0	0	2.5258	2.8566	1	0

Table 2. Preview of Cancer Prediction Dataset

Pregnancies	Glucose	Blood Pressure	SkinThickness	Insulin	BMI	Diabetes PedigreeFunction	Age	Outcome (Target Class)
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
……	……	…….	…….	……	……	…….	…..	…….
1	93	70	31	0	30.4	0.315	23	0

Table 3. Preview of Heart Failure Dataset

Age	Anaemia	Diabetes	High blood pressure	Ejection Fraction	CPK	Platelets	Serum Creatinine	Serum Sodium	Sex	Smoking	Time	Death Event (Target Class)
75	0	0	1	20	582	265000	1.9	140	0	0	4	1
55	1	0	1	38	7861	263358	1.1	136	1	0	6	1
…..	……	……	……	……	…..	…….	…….	…….	….	…….	…..	……
57	1	0	0	30	129	140	1.1	140	0	1	8	0

2.2 Classification on Original Tabular Dataset

Before tabular data is used in tabular data classification, the data must go through a preprocessing stage, including handling missing values through mean imputation for numerical features and mode imputation for categorical features. Categorical features are then converted into numerical form using the one-hot encoding technique and numerical features are normalized using the min-max scaling method to prevent the model from being biased towards

features with large scales. After that, the data is divided into two subsets, namely the training and testing data, generally with an 80:20 ratio. At the classification stage, various machine learning algorithms are used, such as Logistic Regression which measures the linear relationship between features and labels, Support Vector Machine (SVM) which maximizes the margin between classes, Naive Bayes which assumes independence between features, and Extreme Gradient Boosting (XGBoost) which combines multiple decision trees incrementally to improve accuracy [21],[22],[23]. Each algorithm has its characteristics and is chosen based on the complexity and distribution of the data. Model evaluation uses metrics such as accuracy, precision, recall, and F1-score, which are very important in the medical context because classification errors (especially false negatives and false positives) can seriously impact diagnosis and clinical decision-making [24].

2.3 Classification on Serialized Dataset using Classical NLP Method

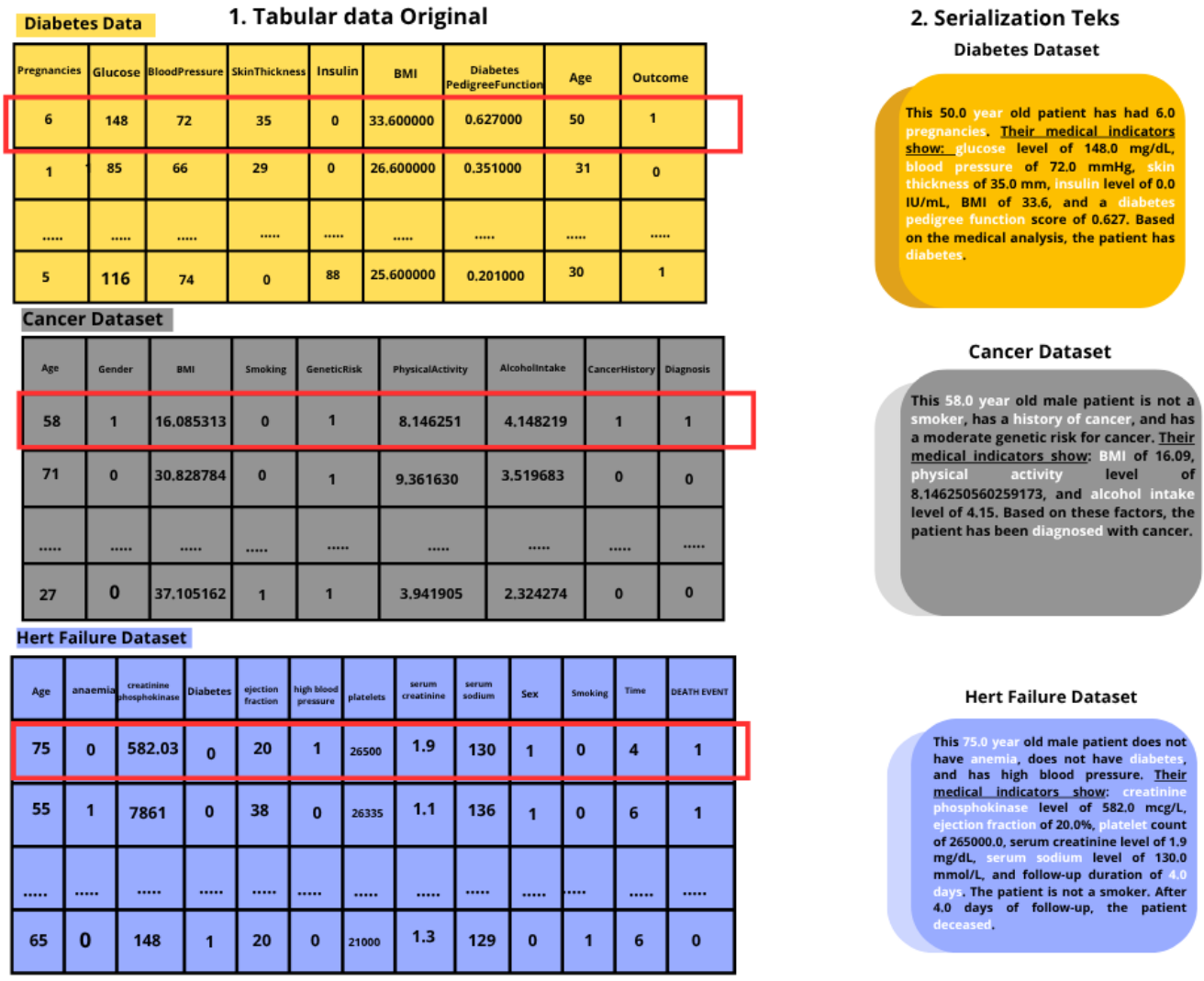

Figure 2. Preview of Serialization Result on Original Tabular Dataset

This paper also takes an alternative approach by re-representing medical tabular data in a natural text-based format. This transformation aims to leverage the semantic representation capabilities of language processing models, both for classical NLP methods and large language models. This design choice enables a systematic comparison between numerical-based machine learning models and language-based approaches under a unified narrative representation. Instead of processing numerical data directly, each data instance is transformed into a narrative description that comprehensively depicts the patient's clinical condition. This serialization process organizes patient information, such as age, gender, medical history, and key medical indicators, into informative and structured sentences. For example, a tabular entry can be transformed into a narrative such as: "This 75-year-old male patient does not suffer from anemia, does not suffer from diabetes, and has high blood pressure…", which is further expanded

with additional disease-related attributes and relevant clinical indicators [19]. After the serialization process, a text representation stage is applied using Natural Language Processing (NLP) techniques to convert narrative descriptions into numerical vectors suitable for machine learning models. Several text representation methods are employed, including Term Frequency–Inverse Document Frequency (TF-IDF), which captures the importance of terms within a document corpus, and Bag-of-Words (BoW), which represents word occurrence frequencies. To capture local linguistic patterns and contextual relationships, n-gram features (unigrams, bigrams, and trigrams) are also utilized. Additionally, semantic embedding-based approaches such as Word2Vec are applied to generate dense vector representations that encode semantic meaning and contextual relationships based on word co-occurrence across the corpus. These techniques are selected to represent classical NLP pipelines with different levels of lexical, contextual, and semantic granularity, enabling a fair and systematic comparison with large language model–based approaches.

All representation methods are trained using the training dataset and subsequently applied to the test dataset to ensure consistency in feature distribution. The resulting feature sets are then used for binary classification tasks, where labels indicate whether a patient is diagnosed with the target disease (positive) or not (negative). Various machine learning classifiers are applied, including Logistic Regression (LR), Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, to perform disease risk classification on serialized medical narratives. This stage evaluates the effectiveness of classical machine learning models in classifying serialized clinical text and provides a baseline for comparison with language-driven LLM-based classification approaches [24]. Figure 2 shows a snippet of the serialization results from each dataset's first column to illustrate the generated narrative structure. With this approach, tabular data that was originally only analyzable through numerical methods can now be integrated into the natural language processing ecosystem. This not only expands the range of classification techniques available, but more importantly, it enhances the interpretability and generalization ability of the model when processing complex clinical information.

2.4 Classification on Serialized using Large Language Model

This paper also utilizes the capabilities of large language models (LLM) to classify patients based on transformed structured medical data into text format through serialization techniques. In this context, LLM not only serves as a text processing tool but also as an intelligent system that can deeply understand and interpret clinical contexts [25]. LLM is capable of providing classifications in the form of yes/no related to the patient's health status, as well as offering recommendations or follow-up suggestions based on the classification results. We designed this LLM-based classification process through three main integrated stages to ensure accuracy, relevance, and effectiveness in handling complex medical data [25]. The first stage is data preparation and prompt creation. Patient medical data, such as medical history, test results, and risk factors, are transformed into a text narrative structured in such a way as to provide clear context to the LLM. this prompt is designed so that the LLM can provide a yes or no classification regarding certain medical conditions, such as whether a patient is at high risk for heart disease or not. Additionally, the LLM provides classification results and generates clinical suggestions or recommendations after classification, such as recommending further actions like advanced tests or initial treatments, which add more value to the medical decision-making process [26]. The second stage is the selection and implementation of LLM models based on advantages relevant to medical applications. We chose three models: LLaMA, Gemini, and Qwen [27],[28],[29]. We chose these models based on their respective advantages discussed in the previous literature. LLaMA was chosen for its capability in few-shot learning, which allows adaptation with limited data. Gemini excels in multidimensional reasoning, suitable for complex medical data. We chose Qwen because of its computational efficiency, which makes it ideal for quickly processing large datasets. We evaluated these three models based on their bias and adaptability to various medical data.

Next, the third stage involves the application of three main learning strategies, namely fine-tuning, zero-shot learning, and few-shot learning. Refinement is carried out by retraining the model using a medical dataset rich in clinical terminology and specific medical patterns, allowing the model to better adapt to certain medical contexts [30]. On the other hand, zero-shot learning allows the model to perform classification without explicit examples, relying solely on the given instructions or questions, thus enabling the model to handle previously unseen data[34]. Meanwhile, few-shot learning provides the model with a small number of examples to help understand patterns in a limited context, enhancing the model's ability to recognize relationships in sparse data [31]. A comprehensive evaluation of these three strategies aims to determine which approach is most effective in medical applications while also identifying areas that need improvement to enhance the performance of LLM in more complex medical classification tasks.

2.4.1 Fine-Tuning Learning Scenario Classification

In this section, the focus is on how the fine-tuning process is applied to LLM for medical data classification. The dataset is divided into training, validation, and test sets to monitor key evaluation metrics such as accuracy and F1-score, ensuring an optimal learning process and preventing overfitting. Also, cross-validation checks how well the model handles changes in the data, and low-rank adaptation (LoRA) adjusts only a small part of the parameters, which lowers computing costs while keeping accuracy high, making it suitable for large and complicated medical datasets [30].

2.4.2 Few-shot Learning Scenario Classification

In this paper, we use an approach that focuses on relevant questions for each task tested with the Fewshot scenario. This model aims to be efficient in understanding context and completing tasks to achieve this, the model strives to obtain optimal results in each task by following standard procedures. Most of the LLM models tested in this paper were adjusted according to the instructions. Therefore, the specific input prompt format used needs to consider the specifications that align with the model's characteristics and requirements. This step is developed in such a way that it enhances model compatibility and maximizes pattern recognition capabilities. As a result, in this approach, the model is trained with few-shot training where the few-shots used are several training stages of 4 shots, 20 shots, 50 shots, 100 shots, and 200 shots. This is done in order to identify trends in the data. The model performs better on tasks that are comparable to those it has previously learnt thanks to this technique, which also helps to better comprehend the intrinsic complexity of the training problems. Additionally, this method provides the option to change the model's setup so that it may manage different kinds of jobs with varying degrees of abstraction [31].

2.4.3 Zero-shot Learning Scenario Classification

In zero-shot Scenario, the answers are generated using the instructions that were provided previously. There is no need for any extra training data. Those stages contrast with other methods like few-shot learning or fine-tuning, which necessitate the use of training data. With zero-shot learning, the model is able to tap into the existing contextual intelligence and knowledge to create relevant responses. The main benefit of the zero-shot approach is its efficacy and flexibility. The model is capable of quickly changing over to new tasks without the use of a training data set and other costly processes [32].

3. Results and Discussion

This section presents the results of our experiments on classifying medical tabular data and predicting patient diagnoses, divided into four parts: (1) Performance of classical machine learning models (e.g., Random Forest, XGBoost) on tabular data across three datasets, Heart Failure, Diabetes, and Cancer, evaluated using accuracy, precision, recall, and F1-score. (2) Comparison of classical models using an NLP-based approach, where tabular data is transformed into text and classified using models like Logistic Regression, SVM, and Naive Bayes. (3) Evaluation of large language models (LLMs), such as LLaMA, Qwen, and Gemini, through fine-tuning, few-shot, and zero-shot learning to assess their ability to classify medical data. (4) An overall comparison of all approaches to highlight their strengths, limitations, and adaptability across different clinical contexts.

3.1 Classification Result on Original Tabular Dataset

Table 4 shows how successfully the different classification models worked on three datasets: Heart Failure, Diabetes, and Cancer, using metrics like accuracy, precision, recall, and F1-score. In the Heart Failure dataset, the logistic regression (LR) and Naive Bayes models showed the best performance with accuracies of 0.86 and the highest F1-scores of 0.78, respectively. Meanwhile, KNN performed the worst, especially regarding recall (0.41). For the diabetes dataset, the performance among models is relatively similar but tends to be lower compared to other datasets. The LR and Naive Bayes models recorded the highest F1 scores of 0.57 and 0.63, respectively. However, generally, the recall values of all models were below 0.70. In the Cancer dataset, the XGBoost model excelled overall with the highest accuracy of 0.92 and an F1-score of 0.88. The SVM and KNN models also show competitive performance with an F1 score above 0.80. Based on these results, XGBoost consistently delivers the best performance, especially on the Cancer dataset. At the same time, classic models like Naive Bayes and KNN show performance variations depending on the dataset's characteristics.

Table 4. Percentage Performance on Tabular Dataset

Dataset	Model	Accuracy	Precesion	Recall	F1-Score
Heart Failure	LR	**0.86**	0.85	0.70	0.77
	SVM	0.77	0.68	0.54	0.60
	KNN	0.74	0.66	0.41	0.51
	XGBoost	0.82	0.72	0.75	0.73
	Naive Bayes	**0.86**	0.81	0.75	0.78
Diabetes	LR	0.73	0.64	0.52	0.57
	SVM	0.75	0.66	0.56	0.61
	KNN	**0.78**	0.59	0.50	0.54
	XGBoost	0.73	0.62	0.59	0.61
	Naive Bayes	0.72	0.60	0.67	0.63
Cancer	LR	0.84	0.79	0.76	0.77
	SVM	0.86	0.85	0.77	0.81

KNN	0.88	0.89	0.79	0.83
XGBoost	**0.92**	0.92	0.85	0.88
NaiveBayes	0.70	0.65	0.46	0.54

3.2 Classification Result on Serialized Dataset using Classical NLP Method

Table 5 compares the performance of five classic NLP-based models on three medical datasets, namely Heart Failure, Diabetes, and Cancer. In the Heart Failure dataset, Logistic Regression (LR) and KNN recorded the highest accuracy of 0.80 with F1-scores of 0.77 and 0.78, respectively, indicating a balanced classification performance. Meanwhile, despite having high precision (0.79), SVM has a low F1-score (0.62) due to the imbalance between precision and recall. For the Diabetes case, the performance of all models tends to be lower, where LR, SVM, and Naive Bayes achieve F1-scores close to 0.69–0.70, while KNN performs poorly with an F1-score of 0.36. On the other hand, in the Cancer dataset, all the main models showed excellent performance, with KNN recording the highest F1-score of 0.83, followed by XGBoost and SVM with 0.82 and 0.79, respectively. Naive Bayes again became the model with the lowest performance in almost all cases, especially on the Cancer dataset, with an F1-score of only 0.48. In general, LR and KNN performed the most consistently, while the effectiveness of the models highly depended on the complexity and structure of the text features in each dataset.

Table 5. Percentage Performance on Serialized Dataset using NLP Method

Dataset	Model	Accuracy	Precesion	Recall	F1-Score
Heart Failure	LR	**0.80**	0.79	0.70	0.77
	SVM	0.71	0.79	0.71	0.62
	KNN	**0.80**	0.79	0.80	0.78
	XGBoost	0.75	0.67	0.64	0.65
	Naive Bayes	0.71	0.72	0.71	0.64
Diabetes	LR	0.69	0.70	0.69	0.69
	SVM	0.68	0.70	0.68	0.68
	KNN	0.44	0.71	0.44	0.36
	XGBoost	0.63	0.60	0.63	0.61
	Naive Bayes	**0.70**	0.71	0.70	0.70
Cancer	LR	0.84	0.79	0.76	0.77
	SVM	0.80	0.80	0.80	0.79
	KNN	**0.88**	0.89	0.79	0.83
	XGBoost	0.82	0.82	0.82	0.82
	NaiveBayes	0.63	0.39	0.63	0.48

3.3 Classification Result on Serialized Dataset Using Large Language Model: Fine-Tuning

The experimental results in Table 6 show performance variations among the three large language models (LLMs) tested: Llama, Qwen, and Gemini. In the Heart Failure dataset, the Gemini model performed best with the highest accuracy (0.683) and balanced other metrics, while Qwen excelled in recall (0.960). For the diabetes dataset, Qwen gave the best accuracy (0.669) and F1 score (0.967), surpassing Gemini. However, the Llama model's performance on this dataset was lower. On the Cancer dataset, the Qwen model once again took first place with accuracy (0.766), F1 Score (0.944), precision (0.927), and recall (0.962), demonstrating its ability to handle complex datasets. Llama was more accurate than Gemini (0.613 vs. 0.589), although both models did poorly.

Table 6. Model Performance on LLM Fine-Tuning scenario

Dataset	LLM	Accuracy	F1 Score	Precesion	Recall
Heart Failure	Llama	0.58	0.42	0.34	0.58
	Qwen	0.61	0.94	0.92	0.96
	Gemini	**0.68**	0.71	0.68	0.70
Diabetes	Llama	0.64	0.50	0.41	0.64
	Qwen	0.64	0.96	0.95	0.97
	Gemini	**0.66**	0.52	0.43	0.66
Cancer	Llama	0.61	0.46	0.37	0.61
	Qwen	**0.76**	0.94	0.92	0.96
	Gemini	0.58	0.42	0.33	0.58

Overall, Gemini is more suitable for tasks that require metric balance, while Qwen consistently detects positive cases. Although Llama has a lower overall performance, it still performs well on specific datasets like Cancer. The best model selection depends on the dataset's characteristics and the classification task's specific needs.

3.4 Classification Result on Serialized Dataset Using Large Language Model: Few-Shot and Zero-shot

Table 7 shows the performance of three large language models (LLaMA, Qwen, and Gemini) in zero-shot and few-shot classification scenarios on three medical datasets: Heart Failure, Diabetes, and Cancer. Overall, all three models perform better as they are given more examples, showing that the few-shot method helps LLMs grasp the classification task better. In the zero-shot scenario, the performance of the models varies, with Gemini performing best on the Diabetes dataset (0.68). At the same time, LLaMA does better on Cancer (0.62), showing that even without specific examples, how well LLMs do is still affected by how the prompts are structured and the nature of the data. When the number of examples increases to medium-shot (for example, 31 and 60), a noticeable spike in accuracy begins to appear, especially in Qwen on the Cancer dataset, which increased from 0.61 (2-shot) to 0.85 (31-shot). Gemini consistently demonstrates performance dominance, especially at high-shot (250), where this model achieves the highest accuracy on almost all datasets, including a perfect score of 1.00 on Cancer. The result indicates that Gemini is more efficient in absorbing classification patterns when adequate examples are provided. This table illustrates that employing a few-shot strategy significantly enhances the classification performance of LLMs, with Gemini demonstrating the best adaptability when provided with more examples. In contrast, zero-shot methods still encounter accuracy issues across various medical domains.

Table 7. LLM Accuracy in Few-shot and Zero-shot Scenarios

			Number of Shoots				
Dataset	Method	**Zero-Shot**	**2**	**4**	**31**	**60**	**250**
Heart Failure	Llama	**0.60**	**0.68**	**0.76**	**0.83**	0.90	0.92
	Qwen	0.58	0.58	0.50	0.80	0.91	0.98
	Gemini	0.58	0.60	0.68	0.81	**0.93**	**0.99**
Diabetes	Llama	0.56	0.60	0.61	0.70	0.82	0.83
	Qwen	0.64	0.68	**0.70**	**0.78**	0.84	**0.99**
	Gemini	**0.68**	**0.70**	0.65	0.77	**0.89**	0.98
Cancer	Llama	**0.62**	0.68	0.69	0.83	0.89	0.98
	Qwen	0.61	0.61	0.50	**0.85**	0.89	0.95
	Gemini	0.61	**0.70**	**0.74**	0.80	**0.92**	**1.00**

3.5 Comparative Analysis of the Best Model's Classification Results

Table 8 presents a comparative analysis of the best-performing models across different classification paradigms for Heart Failure, Diabetes, and Cancer, including direct classification on original tabular data, classification on serialized data using classical NLP methods, and classification on serialized data using the Gemini large language model. On the original tabular datasets, conventional machine learning models such as XGBoost and Logistic Regression consistently achieve strong accuracy, serving as a reliable performance baseline due to their effectiveness in handling structured numerical features. However, when the same datasets are transformed into narrative text and processed using classical NLP-based classifiers, a general decline in accuracy is observed across all diseases. This performance degradation suggests that traditional NLP pipelines face limitations in preserving complex clinical relationships when tabular medical data are converted into sequential text representations.

In contrast, the serialized datasets classified using the Gemini large language model demonstrate substantially improved performance, particularly under the few-shot learning paradigm. While fine-tuning and zero-shot configurations initially yield lower accuracy, the few-shot approach enables Gemini to rapidly adapt to the classification task, achieving near-perfect accuracy of 0.99 for Heart Failure, 0.98 for Diabetes, and 0.98 for Cancer. These results indicate that large language models are more capable of capturing contextual and semantic relationships embedded in serialized clinical narratives, even with a limited number of labeled examples. Overall, this comparison highlights the superior adaptability and generalization ability of Gemini in few-shot settings, positioning LLM-based approaches as a highly effective solution for medical risk classification tasks involving complex and heterogeneous clinical data.

Table 8.Comparison of Model Accuracy on Medical Classification Across Tabular, NLP, and LLM-Based Approaches

Classification Scenario	Model	HertFailure	Diabetes	Cancer
Result on Original Tabular Dataset	LR	0.86	0.73	0.84
	KNN	0.74	0.78	0.88

	XGBoost	0.82	0.73	**0.92**
Result on Serialized Dataset using Classical NLP Method	LR	**0.84**	0.69	**0.84**
	KNN	0.80	0.50	0.63
	XGBoost	0.75	0.63	0.82
Results on the Serialized Dataset Using the Gemini Large Language Model	Fine-tuning	0.68	0.66	0.76
	Zero-shot	0.58	0.68	0.61
	Few-shot	0.99	**0.98**	0.98

4 Conclusion

This paper addresses the challenge of improving patient diagnosis predictions for chronic diseases like heart failure, diabetes, and cancer, given the limitations of current Electronic Health Record (EHR) systems. It aims to develop and compare the performance of various classical machine learning (ML) models (such as XGBoost, SVM, and Logistic Regression) with Large Language Models (LLMs) like Gemini, LLaMA, and Qwen, particularly by transforming medical tabular data into narrative text descriptions through a process called serialization. The study evaluates these models across original tabular datasets and serialized text datasets using different approaches, including classical NLP methods for serialized data, and fine-tuning, zero-shot, and few-shot learning scenarios for LLMs. The most significant finding is that the Gemini LLM, when combined with a few-shot learning approach (specifically 250 shots), achieved the highest accuracy of up to 99.8% in predicting heart failure risk using these narrative representations. The results show that transforming structured tabular data into narrative text significantly improves the ability of Large Language Models (LLMs) to understand the clinical context and make more accurate classifications. This finding emphasizes the high potential of this approach to be applied in AI-based clinical decision support systems.

References

[1] N. R. F. Collaboration, "Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants," *Lancet*, vol. 404, no. 10467, pp. 2077–2093, 2024. https://doi.org/10.1016/S0140-6736(24)02317-1

[2] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.," *CA. Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021. https://doi.org/10.3322/caac.21660

[3] B. Shahim, C. J. Kapelios, G. Savarese, and L. H. Lund, "Global Public Health Burden of Heart Failure: An Updated Review.," *Card. Fail. Rev.*, vol. 9, p. e11, 2023. https://doi.org/10.15420/cfr.2023.05

[4] T. N. Bogale *et al.*, "Effect of electronic records on mortality among patients in hospital and primary healthcare settings: a systematic review and meta-analyses," *Front. Digit. Heal.*, vol. 6, no. June, pp. 1–14, 2024. https://doi.org/10.3389/fdgth.2024.1377826

[5] T. S. Hwang, M. Thomas, M. Hribar, A. Chen, and E. White, "The Impact of Documentation Workflow on the Accuracy of the Coded Diagnoses in the Electronic Health Record," *Ophthalmol. Sci.*, vol. 4, no. 1, p. 100409, 2024. https://doi.org/10.1016/j.xops.2023.100409

[6] R. A. Dixit, C. L. Boxley, S. Samuel, V. Mohan, R. M. Ratwani, and J. A. Gold, "Electronic Health Record Use Issues and Diagnostic Error: A Scoping Review and Framework," *J. Patient Saf.*, vol. 19, no. 1, pp. E25–E30, 2023. https://doi.org/10.1097/PTS.0000000000001081

[7] M. R. Kale, A. H. Mutlag, S. P, N. H. Al-Muraad, H. S. Mahdi, and S. Muthuperumal, "AI Powered Decision Support Systems for Healthcare Enhancing Diagnosis and Treatment with Deep Learning," in *2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE)*, 2025, pp. 1–5. https://doi.org/10.1109/ICICKE65317.2025.11136681

[8] E. Hassan and C. E. Omenogor, "AI powered predictive healthcare: Deep learning for early diagnosis, personalized treatment, and disease prevention," *Int. J. Sci. Res. Arch.*, vol. 14, no. 3, pp. 806–823, 2025. https://doi.org/10.30574/ijsra.2025.14.3.0731

[9] A. Jafar, N. Bibi, and R. A. Naqvi, "Revolutionizing agriculture with arti fi cial intelligence : plant disease detection methods , applications , and their limitations," no. March, pp. 1–20, 2024. https://doi.org/10.3389/fpls.2024.1356260

[10] R. M. Shohel and S. Jeff, "AI in Healthcare: Transforming Patient Care through Predictive Analytics and Decision Support Systems," *J. Artif. Intell. Gen. Sci. ISSN3006-4023*, vol. 1, no. 1, 2024. https://doi.org/10.60087/jaigs.v1i1.30

[11] M. A. Islam *et al.*, "Harnessing Predictive Analytics: The Role of Machine Learning in Early Disease Detection and Healthcare Optimization," *J. Ecohumanism*, vol. 4, no. 3, pp. 312–321, 2025. https://doi.org/10.62754/joe.v4i3.6642

[12] S E. Z. Snigdha, M. R. Hossain, and S. Mahabub, "AI-Powered Healthcare Tracker Development: Advancing Real-Time Patient Monitoring and Predictive Analytics Through Data-Driven Intelligence"," *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 4, pp. 229–239, 2023. https://doi.org/10.32996/jcsts.2023.5.4.24

[13] V Q. Niu, K. Chen, M. Li, P. Feng, Z. Bi, L. K. Q. Yan, Y. Zhang, C. H. Yin, C. Fei, J. Liu, T. Wang, Y. Wang, S. Chen, and B. Peng, "From text to multimodality: Exploring the evolution and impact of large language models in medical practice," arXiv preprint arXiv:2410.01812, 2024. https://doi.org/10.48550/arXiv.2410.01812

[14] A S. Maity and M. J. Saikia, "Large language models in healthcare and medical applications: A review," Bioengineering, vol. 12, no. 6, p. 631, 2025. https://doi.org/10.3390/bioengineering12060631

[15] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023. https://doi.org/10.1002/hcs2.61

[16] G. Huang, Y. Li, S. Jameel, Y. Long, and G. Papanastasiou, "From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality?," *Comput. Struct. Biotechnol. J.*, vol. 24, no. November 2023, pp. 362–373, 2024. https://doi.org/10.1016/j.csbj.2024.05.004

[17] F. Markowetz, "All models are wrong and yours are useless: making clinical prediction models impactful for patients," *npj Precis. Oncol.*, vol. 8, no. 1, pp. 6–8, 2024. https://doi.org/10.1038/s41698-024-00553-6

[18] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, "Bias in Machine Learning: A Literature Review," *Appl. Sci.*, vol. 14, no. 19, 2024. https://doi.org/10.3390/ app14198860

[19] K. Ono and S. A. Lee, "Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning," 2024. https://doi.org/10.48550/arXiv.2406.13846
[20] World Health Organization, "The top 10 causes of death," 8 August 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
[21] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Inf.*, vol. 15, no. 4, 2024. https://doi.org/10.3390/info15040235
[22] K. Gohari *et al.*, "A Bayesian latent class extension of naive Bayesian classifier and its application to the classification of gastric cancer patients," *BMC Med. Res. Methodol.*, vol. 23, no. 1, pp. 1–15, 2023. https://doi.org/10.1186/s12874-023-02013-4
[23] X. Tang *et al.*, "A clinical diagnostic model based on an eXtreme Gradient Boosting algorithm to distinguish type 1 diabetes," *Ann. Transl. Med.*, vol. 9, no. 5, pp. 409–409, 2021. https://doi.org/10.21037/atm-20-7115
[24] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022. https://doi.org/10.1038/s41598-022-09954-8
[25] S. Maity and M. J. Saikia, "Large Language Models in Healthcare and Medical Domain: A Review," *Informatics*, vol. 11, no. 3, pp. 1–25, 2024, doi: https://doi.org/10.3390/informatics11030057.
[26] M. J. Schuemie *et al.*, "Standardized patient profile review using large language models for case adjudication in observational research," *npj Digit. Med.*, vol. 8, no. 1, pp. 1–7, 2025. http://dx.doi.org/10.1038/s41746-025-01433-4
[27] A. Q. Xie, Q. Chen, A. Chen, and C. Peng, "Me-LLaMA : Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond," pp. 1–21. https://doi.org/10.21203/rs.3.rs-5456223/v1
[28] K. Saab *et al.*, "Capabilities of Gemini Models in Medicine," pp. 1–58, 2024. https://doi.org/10.48550/arXiv.2404.18416
[29] S. Zhu, W. Hu, Z. Yang, J. Yan, and F. Zhang, "Qwen-2.5 Outperforms Other Large Language Models in the Chinese National Nursing Licensing Examination: Retrospective Cross-Sectional Comparative Study.," *JMIR Med. informatics*, vol. 13, p. e63731, Jan. 2025. https://doi.org/10.2196/63731
[30] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning Large Language Models for Specialized Use Cases," *Mayo Clin. Proc. Digit. Heal.*, vol. 3, no. 1, p. 100184, 2025. https://doi.org/10.1016/j.mcpdig.2024.11.005
[31] Y. Ge, Y. Guo, Y.-C. Yang, M. A. Al-Garadi, and A. Sarker, "Few-shot learning for medical text: A systematic review," 2022. https://doi.org/10.48550/arXiv.2204.14081
[32] B. Neves *et al.*, "Zero-shot learning for clinical phenotyping: Comparing LLMs and rule-based methods," *Comput. Biol. Med.*, vol. 192, no. PA, p. 110181, 2025. https://doi.org/10.1016/j.compbiomed.2025.110181