

The Role of Advanced Language Models in Medical Diagnostics: A Case Study on Breast Cancer Prediction

Habibe Karayığit^{1*} , Filiz Kalelioğlu² 

¹ General Directorate of Secondary Education, Ministry of National Education, Adana (Türkiye)

² Department of Computer Education and Instructional Technology, Başkent University, Ankara (Türkiye)

* Corresponding author: habibekarayıgit@aof.anadolu.edu.tr

Received 21 November 2024 | Accepted 10 June 2026 | Published 28 August 2026



ABSTRACT

Breast cancer remains one of the most common and life threatening cancers worldwide, and early detection is strongly associated with improved survival and reduced treatment burden. This study investigates the ability of Large Language Models to perform diagnostic prediction from structured breast cancer related data. We systematically evaluated 12 LLMs across three public datasets with different clinical characteristics: the Wisconsin Breast Cancer Dataset (WBCD) based on cytological features, the Breast Cancer Coimbra Dataset (BCCD) based on metabolic biomarkers, and the Mammographic Mass Dataset (MMD) based on mammographic attributes. The evaluation covered 13 prompting strategies, including three zero-shot variants, few-shot prompting, and multiple Chain-of-Thought (CoT) and knowledge-enhanced reasoning settings. Performance was assessed using confusion-matrix-based metrics, including accuracy, precision, recall, F1-score, specificity, and Matthews Correlation Coefficient. The results showed that performance was strongly dependent on both dataset type and prompting design, and no single model dominated all tasks. The best model-strategy pair varied by dataset: Cogito-v1-preview-qwen-32B achieved the highest F1-score on WBCD with an F1-score of 92.00%, GPT 4.1 and GPT 4o on BCCD with an F1-score of 85.39%, and Gemini 2.5 Flash Lite on MMD with an F1-score of 82.91%. Prompt engineering had a substantial effect on outcomes, but its benefit varied across models, with some systems improving under knowledge-enhanced few shot prompting and others performing best under simpler strategies. Comparison with state-of-the-art traditional ML baselines showed that, while LLMs do not yet surpass supervised methods, the performance gap has narrowed substantially, particularly on MMD, where the best single-run gap was 2.04 percentage points in F1 and the mean gap under robustness analysis was approximately 5.3 points. Robustness analysis across multiple few-shot example sets confirmed stable performance on WBCD ($F1 = 91.37 \pm 0.71\%$) and BCCD ($F1 = 87.46 \pm 1.91\%$), while revealing moderate sensitivity on MMD ($F1 = 79.62 \pm 2.87\%$). Although the evaluated LLMs did not outperform traditional supervised models, the study provides a clear performance baseline for future research on structured clinical prediction with language models. The results show that LLMs may offer value as complementary exploratory tools, but their outputs should be interpreted only with expert oversight because clinically significant errors remain.

KEYWORDS

AI-based Medical Diagnosis, AI-based Prediction, Breast Cancer Diagnosis, Comparative Analysis, Early Detection, Large Language Models.

DOI: 10.9781/ijimai.2026.2221
eLocator: e2221

I. INTRODUCTION

BREAST cancer (BrC) remains one of the most common and clinically significant cancers worldwide and is a major cause of mortality among women, although it can also affect men [1], [2]. It arises from the uncontrolled growth of breast cells and is influenced by multiple risk factors, including age, genetic mutations, family history, radiation exposure, and hormonal changes [3]. Diagnosis typically relies on self-examination, clinical assessment, mammography, ultrasonography, and magnetic resonance imaging [4]. Numerous studies have shown that early detection is strongly associated with improved survival,

with five-year survival rates exceeding 90% for early-stage diagnosis and dropping below 30% for late-stage cases [2], [5].

The consequences of delayed diagnosis extend well beyond reduced survival. Advanced-stage BrC and its treatment may lead to serious complications, including febrile neutropenia, organ dysfunction, and intensive care admission, all of which are associated with substantial mortality and clinical burden [6], [7], [8]. In this context, the need for rapid, reliable, and sensitive diagnostic support tools becomes even more important. Recent work has also shown that the analysis of unstructured clinical text can improve the early identification of severe conditions such as sepsis, further underscoring the broader potential of AI-based systems in healthcare [9].

Please cite this article as:

H. Karayığit, F. Kalelioğlu. The Role of Advanced Language Models in Medical Diagnostics: A Case Study on Breast Cancer Prediction, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 10, no. 1, e2221, 2026, <http://doi.org/10.9781/ijimai.2026.2221>

At the same time, conventional diagnostic workflows may still suffer from interpretation variability, dependence on clinician experience, and delays in medical decision-making. These limitations have motivated growing interest in Artificial Intelligence, and particularly in Large Language Models, as supplementary tools for healthcare decision support [10]. LLMs have demonstrated strong performance in a wide range of natural language processing tasks and have raised increasing interest in diagnostic classification settings because of their ability to understand instructions, reason over inputs, and generate structured outputs [11]. Within this context, the present study investigates the use of LLMs for breast-cancer-related prediction from structured clinical data.

This study presents a systematic evaluation of 12 AI-based large language models for breast-cancer-related classification across three datasets: the Wisconsin Breast Cancer Dataset (WBCD), the Breast Cancer Coimbra Dataset (BCCD), and the Mammographic Mass Dataset (MMD). The evaluated models include GPT-4.1, GPT-4o, Qwen 2.5 7B Instruct, Gemma-3-4B-IT, Mistral 7B Instruct v0.3, Phi-3 Medium 14B, DeepSeek-R1-Distill-Llama-8B, GPT-OSS-20B, Qwen3-4B-Thinking-Gemini-VIBE, Llama 3.1 8B Instruct, Gemini 2.5 Flash Lite, and Cogito-v1-preview-qwen-32B. Their performance under 13 distinct prompting strategies, including three zero-shot variants, few-shot, and various Chain-of-Thought (CoT) reasoning methods, was examined. The results revealed that the best-performing model-strategy pair varied by dataset: Cogito-v1-preview-qwen-32B achieved the highest F1-Score on WBCD (92.00% with 3-Shot), GPT-4.1 on BCCD (85.39% with 5-Shot KE-CoT), and Gemini 2.5 Flash Lite on MMD (82.91% with 3-Shot KE-CoT-ExR). These findings represent a significant step in understanding how to effectively utilize LLMs for diagnostic tasks involving structured data in the healthcare sector.

The main contributions of this study can be summarized as follows: 1. To the best of our knowledge, this is the first study to systematically compare 12 different language models - spanning both proprietary API-based and locally-deployed open-source architectures - specifically applied to BrC diagnosis under an extensive set of 13 distinct prompting strategies and across three distinct datasets. 2. Exploring the impact of prompt engineering on the prediction performance of LLMs working with structured data, demonstrating that the optimal model-strategy pair is dataset-dependent. 3. Providing formal baselines by comparing LLM performance against published state-of-the-art traditional ML/DL methods, and highlighting clinically significant error rates to underscore that results must be evaluated by an expert. 4. Conducting a robustness analysis across three different few-shot example sets to assess the stability of prompt-driven results and quantify sensitivity to example selection.

The remaining part of this paper is organized as follows: Section II presents previous studies on the use of LLMs in the diagnosis of diseases. Section III outlines the materials and methods employed in this study. Section IV presents the experimental results, and Section V provides a discussion for future research directions based on the findings, along with various insights.

II. LITERATURE REVIEW

The use of transformer-based language models (TLMs) has revolutionized the ability to capture semantic relationships within text data, allowing for a more nuanced understanding of specific datasets [12]. There is a growing body of work exploring the efficacy of AI-driven intelligent language models in the diagnosis of diseases, particularly those of low and medium risk [13], [14]. The application of TLMs in medical diagnosis and treatment has yielded positive outcomes for advanced tasks, highlighting the potential of these models to enhance healthcare delivery [15], [16]. Recent advancements

in AI-based language models, particularly in the healthcare sector, have sparked interest in their potential to support medical education and clinical decision-making. For instance, ChatGPT's performance on the United States Medical Licensing Examination (USMLE), which consists of three exams, was assessed without any specialized training or additional information, demonstrating high levels of compatibility and comprehension. Consequently, the potential of LLMs to assist in medical education and potentially in clinical decision-making has been reported [17]. In another study utilizing ChatGPT, the model was employed to validate medical alerts and clinical recommendations developed by clinicians at the Vanderbilt University Medical Center, yielding promising results [18]. In a study assessing the potential of ChatGPT-3.5 and GPT-4 for radiological diagnosis in BrC screening, a need for broader usage scenarios and higher accuracy standards was reported [19]. ChatGPT's success in standard clinical cases has been evaluated, showing high accuracy rates in clinical decision-making, with particular strengths emerging due to the model's access to extensive clinical data [20].

Recent studies have shown that language models can be useful in a variety of medical diagnosis and biomedical text analysis tasks. For example, BioBERT, which was pre-trained on large-scale biomedical corpora, has outperformed the original BERT model in several biomedical text mining applications [21]. Similarly, BioGPT has been evaluated across multiple biomedical NLP tasks and achieved an accuracy of 78.2% on PubMedQA [22]. In addition, BERT-based diagnostic models that learn contextual dependency patterns from patient-related text have also reported strong results, with accuracy reaching as high as 96.63% in some settings [23].

Although LLMs have shown potential applications in medical fields, especially in administrative tasks, clinical knowledge support, medical education, and research, concerns regarding biases in datasets, data quality, unpredictability of outputs, and patient privacy pose challenges to their adoption in healthcare [11]. A study investigating the accuracy of ChatGPT's tuberculosis diagnoses based on symptoms under medical expert supervision reported improved outcomes, suggesting economic benefits to the health system. Furthermore, a new evaluation method for the use of LLMs in medical interactions has been proposed [24]. ChatGPT-4 has been found to consistently provide accurate answers to complex vascular surgery questions at a clinician level, though limitations persist, including hallucinations, data privacy issues, and the black box problem [25].

The Ethical, Legal, and Social Implications (ELSI) within the field of AI-based medical decision support tools have been scrutinized, with studies drawing inferences on how AI decision-making tools may shape the future of healthcare services. While these intelligent systems have shown potential for improvement in healthcare systems, various disadvantages, including ethical concerns, have been highlighted [26]. In an attempt to collate expert opinions on the early clinical evaluation reporting of AI systems, a two-phase modified Delphi process was conducted, culminating in the DECIDE-AI reporting guideline, which includes positive outcomes related to the evaluation of early-stage and small-scale applications of AI-based decision support systems [27]. A comprehensive overview of AI intelligent systems in clinical decision support systems has presented proven current use cases, common pitfalls, and potential harms, emphasizing the need for further advancements in usability, speed, deployment ease, and affordability [28]. Recent advancements in AI-based LLMs have raised significant interest in their potential impact on healthcare, particularly regarding their effectiveness in interpreting patient-reported symptoms of BrC. Therefore, in this paper, we compare the performance of 12 popular LLMs on structured breast-cancer-related classification using cytological, biomarker-based, and mammographic features. To the best of our knowledge, this is among the first studies to systematically

TABLE I. COMPARATIVE ANALYSIS OF EXISTING STUDIES ON AI AND LLMs IN MEDICAL DIAGNOSTICS

Study/Method	Year	Approach/Model	Dataset	Key Features	Performance
Street et al. [29]	1993	Feature extraction	WBCD (Diagnostic)	30 digitalized nuclear features	97% Acc
Wolberg et al. [30]	1995	Linear Programming	WBC (Original)	9 FNA cytology features	97.5% Acc
Chougrad et al. [31]	2018	Convolutional Neural Network (CNN)	DDSM, IN breast, and BCDR	Automated feature learning	97.35%, 95.50%, 96.67% Acc for three datasets respectively
Karthik [32]	2018	Deep neural network (DNN)	WBCD	9 attributes with 569 instances	98.62% Acc
Kung et al. [17]	2023	ChatGPT	USMLE Step 1/2CK/3 question sets	Zero-shot learning	~60–70% Acc (step-dependent)
Rao et al. [19]	2023	GPT-3.5/GPT-4	Breast imaging pilot cases	Radiology decision support (text/vignette interpretation)	72–83% Acc (task-dependent; GPT-4 > GPT-3.5)
Lee et al. [21]	2020	BioBERT	Biomedical corpora (PubMed + PMC)	Domain-specific pretraining on PubMed + PMC	F1 -89–94% (task-dependent; SOTA/near-SOTA on BioNLP tasks)
Luo et al. [22]	2022	BioGPT	PubMedQA	Domain-specific generative pretraining	78.2% Acc
Our Proposed Approach	2026	12 LLMs (GPT-4.1, GPT-4o, Qwen 2.5 7B, Gemma-3-4B-IT, Mistral 7B v0.3, Phi-3 14B, DeepSeek-R1 8B, GPT-OSS-20B, Qwen3-4B-Think., Llama 3.1 8B, Gemini 2.5 FL, Cogito-v1 32B)	WBCD, BCCD, MMD	13 prompting strategies; Zero/Few-shot; CoT; KE-CoT; Robustness analysis; ML baselines	Best F1: WBCD 92.00%, BCCD 85.39%, MMD 82.91%

compare a broad set of LLMs across multiple prompting strategies for structured breast-cancer-related classification. To provide a clearer context of how the proposed study compares with previous research, Table I summarizes representative studies on BrC diagnosis using AI and language-model-based methods.

To systematically contextualize our research within the existing body of work, Table I provides a comparative summary of key studies employing AI and LLMs in medical diagnostics. This comparison highlights the diversity of approaches, from domain-specific model fine-tuning to the use of general-purpose chatbots. While previous studies have demonstrated the potential of LLMs in medical question-answering and text mining, our study is distinguished by its systematic, multi-model evaluation focused specifically on a structured clinical prediction task (BrC diagnosis) and the extensive investigation of prompt engineering strategies, which remains an underexplored area in clinical applications.

III. MATERIALS AND METHODS

A. Materials

Three publicly available datasets from the UCI Machine Learning Repository were used in this study. These datasets represent three distinct diagnostic modalities: cytological features, metabolic/biochemical biomarkers, and mammographic descriptors.

1. Wisconsin Breast Cancer Dataset (WBCD)

The Wisconsin Breast Cancer Dataset (WBCD) contains 569 samples, including 357 benign and 212 malignant cases, described by 30 cytological and morphometric features extracted from fine-needle aspiration samples [33]. These variables include measurements related to radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset is widely used as a benchmark for breast cancer classification and provides a relatively high-dimensional structured feature space.

2. Breast Cancer Coimbra Dataset (BCCD)

The Breast Cancer Coimbra Dataset (BCCD) contains 116 participants, including 52 healthy controls and 64 patients, represented

by 9 metabolic and anthropometric features [34]. These features are Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1. Unlike WBCD, which is based on direct tumor-related cytological measurements, BCCD reflects a more indirect physiological characterization of breast cancer status through systemic biomarkers.

3. Mammographic Mass Dataset (MMD)

The Mammographic Mass Dataset (MMD) contains 961 observations, including 516 benign and 445 malignant cases, described by 5 BI-RADS-based mammographic attributes [35]. These are BI-RADS assessment, Age, Shape, Margin, and Density. Compared with WBCD and BCCD, MMD contains fewer input variables but encodes clinically meaningful radiological descriptors and ordinal relationships. Feature relevance analysis using Pearson correlation and Mutual Information (MI) was conducted for each dataset. The top-5 most informative features are summarized in Table II. Despite varying relevance levels, all available features were retained in the LLM prompts in order to evaluate model robustness under realistic structured-input settings and to avoid introducing feature-selection bias into the prompt design.

TABLE II. TOP-5 MOST INFORMATIVE FEATURES BY DATASET BASED ON PEARSON CORRELATION AND MUTUAL INFORMATION

Dataset	Rank	Feature	r	MI Score
WBCD	1	worst concave points	0.79	0.42
	2	worst perimeter	0.78	0.45
	3	mean concave points	0.78	0.41
	4	worst radius	0.78	0.44
	5	mean perimeter	0.74	0.38
BCCD	1	Glucose	0.35	0.12
	2	HOMA	0.27	0.10
	3	Insulin	0.22	0.08
	4	Resistin	0.18	0.06
	5	BMI	0.14	0.04
MMD	1	Margin	0.59	0.19
	2	Shape	0.56	0.20
	3	Age	0.47	0.13
	4	BI-RADS	0.23	0.25
	5	Density	0.06	0.08

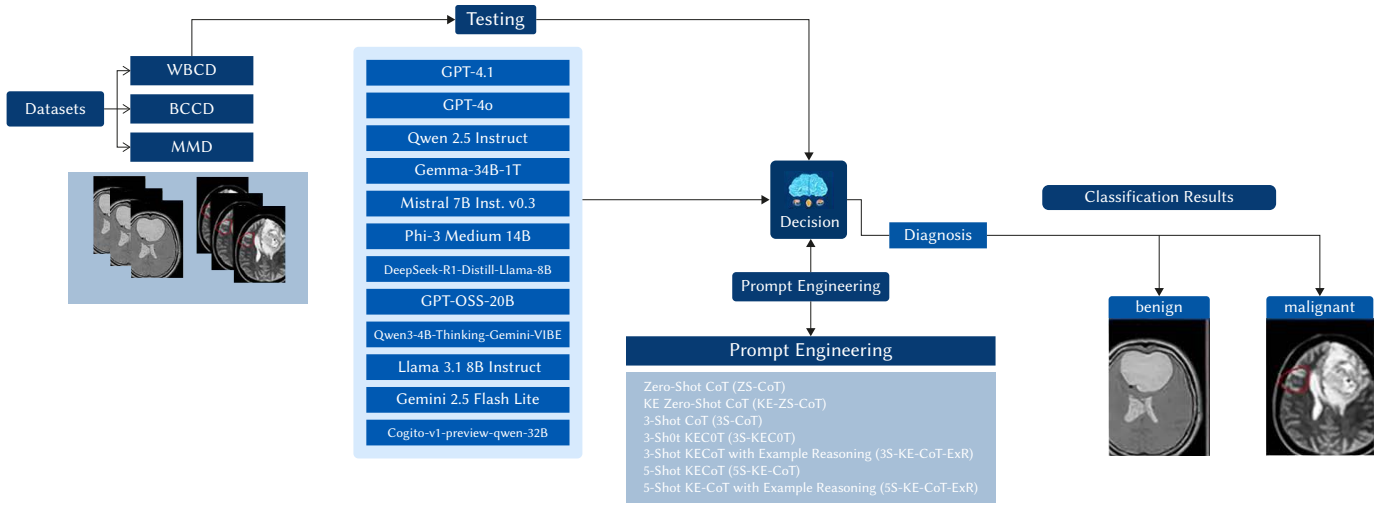


Fig. 1. Pipeline of the process of identifying BrC from symptoms using LLMs.

B. Methods

In this study, LLMs were chosen as a specific mechanism to address the limitations of traditional machine learning approaches used in BrC diagnosis, such as the need for extensive labeled data and complex model training. LLMs offer the potential for a more flexible and less data-dependent diagnostic support system, thanks to their ability to reason on structured data without retraining (zero-shot) or with very few labeled examples (few-shot). This section details the multi-phase experimental process and the model configurations used to systematically evaluate the performance of this innovative approach across 12 different LLMs. The overall workflow is summarized in Fig. 1. The process began with the selection of the Wisconsin Breast Cancer Dataset (WBCD), Breast Cancer Coimbra Dataset (BCCD), and Mammographic Mass Dataset (MMD), followed by data preprocessing. These three datasets represent distinct diagnostic domains: WBCD contains cytological and morphometric features derived from fine-needle aspirates; BCCD focuses on metabolic and anthropometric biomarkers obtained from blood samples; and MMD includes imaging-derived features based on BI-RADS criteria from mammography reports. A feature relevance analysis was conducted using both correlation and MI scores, applying the CBFS approach for each dataset. However, to preserve contextual richness and evaluate model robustness under high-dimensional input conditions, all available features from WBCD (30 features), BCCD (9 features), and MMD (5 features) were ultimately retained in the prompts. Three learning strategies were employed: zero-shot, few-shot, and CoT prompting. For each test instance, the LLM was prompted with structured clinical, metabolic, or imaging data depending on the dataset and was expected to classify the case as malignant (1) or benign (0). For BCCD, the output labels were adapted to represent patient (1) and healthy (0) classes. The models were not fine-tuned; instead, dynamically generated prompts containing a small number of labeled examples (e.g., 3, 5, or 10) were used to simulate an in-context learning scenario, effectively reframing the classification task into a question-answering problem. Prompt engineering was carefully implemented to explore how context size and structure influence model performance across the three datasets. API configurations were customized for each LLM, ensuring consistent and reliable output collection.

1. Prompt Engineering

Prompt engineering was applied within an in-context learning framework, transforming the classification task into a question-answering problem [36]. Thirteen prompting strategies were designed across three categories:

Zero-shot (3 variants): (i) Direct—simulating a non-expert query; (ii) Technical—formal, data-centric framing; (iii) Clinical—adopting professional terminology. Example (Direct): “I have the following measurement values from a breast mass: [FEATURE_LIST]. Based on these values, is the diagnosis more likely to be 0 (Benign) or 1 (Malignant)? Please provide the answer as a single digit, 0 or 1.”

Few-shot (3 variants): $k=3, 5$, and 10 class-balanced examples were prepended to the technical query template.

Chain-of-Thought (7 variants): (i) ZS-CoT—appending “Let’s think step by step”; (ii) KE-ZS-CoT—adding domain-specific role framing and feature cues; (iii) 3S-CoT—combining 3-shot examples with basic CoT instruction; (iv) 3S-KE-CoT—combining 3-shot examples with KE framing and CoT; (v) 3S-KE-CoT-ExR—including human-written reasoning in each of 3 examples; (vi) 5S-KE-CoT—combining 5-shot examples with KE framing and CoT; (vii) 5S-KE-CoT-ExR—including human-written reasoning in each of 5 examples.

2. Technical Algorithm for LLM-based Diagnostic Classification

This study employs a single-pass inference methodology using LLMs with in-context prompts for the binary classification of structured clinical cases. For label harmonization, a consistent binary outcome was enforced across all datasets: 0 = Benign/Healthy and 1 = Malignant/Patient. For the BCCD, original labels $\{1, 2\}$ were remapped to $\{0, 1\}$. To ensure prompt compliance, all templates strictly instruct the model to output only a single numerical label (0 or 1). The API settings follow model and strategy-specific parameters from Table III, and for Chain-of-Thought (CoT) variants, the `max_tokens` parameter is increased to accommodate the generated reasoning text. The dataset D is defined as $D = \{x_i, y_i\}_{i=1}^N$, where x_i is the feature vector and $y_i \in \{0, 1\}$ is the ground-truth label. The strategy S belongs to the set $\{ZS - Direct, ZS - Technical, ZS - Clinical, k - Shot, ZS - CoT, KE - ZS - CoT, k - Shot - CoT, k - Shot - KE - CoT, k - Shot - KE - CoT - ExR\}$, with $k \in \{0, 3, 5, 10\}$ where $k = 0$ denotes the zero-shot approach, yielding a total of 13 distinct strategy configurations.

The process utilizes a single-pass LLM inference loop. For each test instance, a strategy-specific prompt (optionally including k labeled examples and/or reasoning cues) is constructed. The target LLM is queried, the single-digit label is extracted, and predictions are aggregated to compute evaluation metrics.

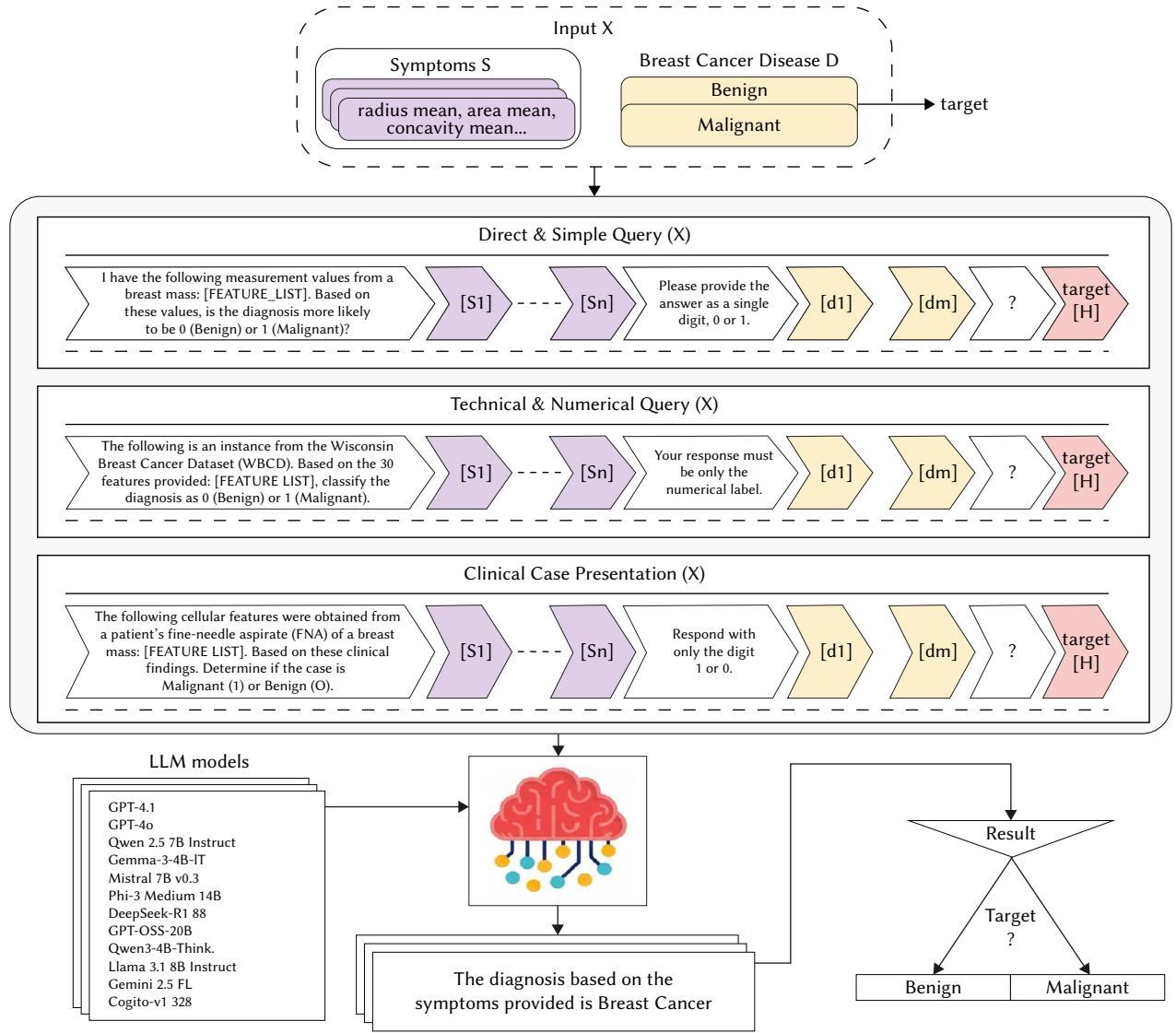


Fig. 2. Zero-shot prompt engineering methodology of BrC prediction.

Algorithm 1: Single-pass LLM Inference with In-Context Prompting

Input: Dataset D , LLM L , strategy S , number of shots $k \in \{0, 3, 5, 10\}$

Output: $Y_{pred} = \text{list of } (\hat{y}_{pred}, y_{true})$

```

1:  $Y_{pred} \leftarrow []$ 
2: for each  $(x_{test}, y_{true})$  in  $D$  do
3:    $E_k \leftarrow \emptyset$ 
4:   if  $k > 0$  then
5:      $E_k \leftarrow \text{SelectExamples}(D, k)$ 
6:   end if
7:    $\text{prompt} \leftarrow \text{GeneratePrompt}(x_{test}, S, E_k)$ 
8:    $\text{params} \leftarrow \text{GetAPIParams}(L, S)$ 
9:    $\text{response} \leftarrow L.\text{query}(\text{prompt}, \text{params})$ 
10:   $\hat{y}_{pred} \leftarrow \text{ExtractLabel}(\text{response})$ 
11:   $\text{append}(\hat{y}_{pred}, y_{true})$  to  $Y_{pred}$ 
12: end for
13: return  $Y_{pred}$ 
    
```

From the aggregated predictions Y_{pred} , True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN) are computed. Standard metrics, including Accuracy, Precision, Recall, F1, Specificity, and Matthews Correlation Coefficient (MCC), are derived. Statistical comparisons utilize a chi-squared test on correct vs. incorrect counts. The algorithm is a faithful reflection of the study's design, employing no model updates, per-instance dynamic prompting, class-balanced few-shot context, explicit CoT and KE variants, and crucially, single-pass inference without retries or auxiliary parsing to assess the baseline LLM performance.

3. Learning Strategy Selection

Different experimental scenarios were designed based on the learning strategies employed in this study. These scenarios include zero-shot, few-shot learning, and CoT reasoning. Each scenario was tested to evaluate its impact on the BrC prediction performance of the 12 different LLMs across all three datasets—WBCD, BCCD, and MMD. In the zero-shot learning scenario, the model performs a task without any provided examples or prior training (Fig. 2). To establish a baseline, we designed three distinct zero-shot prompt templates that simulate different user personas, following the Manual Template Engineering methodology [37]. These prompting strategies were

applied consistently to each dataset: WBCD (cytological features), BCCD (metabolic and anthropometric biomarkers), and MMD (imaging-derived attributes) to assess the adaptability of LLMs across diverse medical data modalities.

For a given input instance x containing a set of features $S = \{s_1, s_2, \dots, s_d\}$, $d \in \{30, 9, 5\}$ we defined the following prompting functions:

- **Direct & Simple Query ($f_{direct}(x)$):** Simulates a non-expert user.

Template: “I have the following measurement values from a breast mass: [FEATURE_LIST]. Based on these values, is the diagnosis more likely to be 0 (Benign) or 1 (Malignant)? Please provide the answer as a single digit, 0 or 1.”

- **Technical & Neutral Query ($f_{technical}(x)$):** Frames the task in a more formal, data-centric manner.

Template: “The following is an instance from {DATASET_NAME} $\in \{WBCD, BCCD, MMD\}$, $\{d\} \in \{30, 9, 5\}$. Based on the $\{d\}$ features provided: [FEATURE_LIST], classify the case as 0 (Benign/Healthy) or 1 (Malignant/Patient). Your response must be only the numerical label.”

Template: “The following cellular features were obtained from a patient evaluation, such as a fine-needle aspirate (for WBCD), a biochemical assessment (for BCCD), or a mammography report (for MMD): [FEATURE_LIST]. Based on these findings, determine whether the case is Malignant (1) or Benign (0). Respond with only the digit 1 or 0.”

Few-shot learning involves providing the model with k complete examples to help it generalize to new cases [27]. The objective is to enable the model to generate accurate predictions by learning from the patterns within a small number of examples. We evaluated scenarios for $k=3$, $k=5$, and $k=10$. During the process of example selection, to prevent bias, we ensured a balanced number of examples for each outcome (e.g., three malignant and two benign cases for a 5-shot scenario). For each test instance x_{test} , a set of k examples, E_k , was randomly selected, and the final prompt x' was constructed by concatenating E_k with one of the zero-shot query templates.

- **Clinical Case Presentation ($f_{clinical}(x)$):** Adopts clinical terminology to simulate a professional query.

Formal Representation:

$$Prompt_{k-shot} = E_k \oplus f_{technical}(x_{test})$$

The template for this strategy concatenates k examples with a zero-shot query.

Template ($k=3$ example):

“The following are examples of BrC classification:

Example 1:

Features: [FEATURE_LIST_1]

Diagnosis: 1

Example 2:

Features: [FEATURE_LIST_2]

Diagnosis: 0

Example 3:

Features: [FEATURE_LIST_3]

Diagnosis: 1

Based on the examples, classify the following instance:

The following instance originates from the selected dataset (e.g., WBCD – 30 features, BCCD – 9 features, MMD – 5 features). Based on the features provided [FEATURE_LIST_TEST], classify the case as 0 (Benign/Healthy) or 1 (Malignant/Patient). Your response must be only the numerical label (0 or 1).

To investigate the model’s reasoning abilities, several Chain-of-Thought (CoT) prompting strategies were implemented. CoT encourages the model to generate a step-by-step rationale before providing a final answer, which can improve accuracy on complex tasks [38]. We explored the following seven variations:

- **Zero-Shot CoT (ZS-CoT):** This is the simplest form, where a reasoning trigger is appended to a standard query. The prompt is modified by adding the phrase “Let’s think step by step.”

Template: “The following is an instance from the selected dataset (e.g., WBCD – 30 features, BCCD – 9 features, MMD – 5 features). Based on the provided feature values [FEATURE_LIST], classify the diagnosis as 0 (Benign/Healthy) or 1 (Malignant/Patient). Let’s think step by step. Your final answer must be only the numerical label (0 or 1).”

- **KE Zero-Shot CoT (KE-ZS-CoT):** To improve upon the standard CoT, we used a “Knowledge-Enhanced” approach that integrates domain-specific knowledge to help the model make more context-aware predictions. The prompt message consists of three parts: first, the model’s role is defined (Role Assignment and Context Definition); second, it is encouraged to reason sequentially (Step-by-Step Reasoning Instruction); finally, the structure and domain-specific instructions are provided to guide the reasoning process (Prediction Steps with Domain-Specific Knowledge).

Template: “Think of yourself as a BrC classification model familiar with the selected dataset (e.g., WBCD – cytological features, BCCD – metabolic biomarkers, MMD – mammography attributes). The following is an instance from that dataset. Based on the provided feature values [FEATURE_LIST], classify the diagnosis as 0 (Benign/Healthy) or 1 (Malignant/Patient). Let’s think step by step, considering that key indicators (e.g., ‘worst concave points’ and ‘worst perimeter’ for WBCD) are often highly indicative of malignancy. Your final answer must be only the numerical label (0 or 1).”

- **3-Shot CoT (3S-CoT):** This variant combines three in-context labeled examples with the basic CoT instruction “Let’s think step by step,” but without knowledge-enhanced role framing. It tests whether the mere presence of examples combined with a generic reasoning trigger is sufficient to improve performance over plain few-shot prompting.

- **3-Shot KE-CoT (3S-KE-CoT):** This variant extends 3S-CoT by incorporating the knowledge-enhanced role framing from KE-ZS-CoT. Three labeled examples are prepended to the KE-enriched query with CoT instruction, combining in-context learning with domain-aware reasoning.

- **3-Shot KE-CoT with Example Reasoning (3S-KE-CoT-ExR):** This represents one of the most sophisticated prompting strategies in the study. For $k=3$, the prompt includes 3 examples each containing not only the features and the target label but also a human-written rationale explaining why the diagnosis is classified as such. Through these illustrations, the model learns both how to formulate predictions and the logical progression of reasoning behind them, thereby enhancing the reliability and interpretability of its outputs. This strategy is explicitly designed to teach the desired step-by-step clinical reasoning process.

Template (k=3 example):

“Think of yourself as a breast cancer classification model familiar with the WBCD. Below are examples with reasoning.

Example 1:

Features: [FEATURE_LIST_1]

Reasoning: The ‘worst radius’ and ‘worst concave points’ values are high, which strongly suggests a malignant tumor. Therefore, the diagnosis is malignant.

Diagnosis: 1

Example 2:

Features: [FEATURE_LIST_2]

Reasoning: The feature values, such as ‘mean radius’ and ‘worst compactness’, are within the normal range, indicating a benign case.

Diagnosis: 0

Example 3:

Features: [FEATURE_LIST_3]

Reasoning: Several key indicators like ‘worst perimeter’ and ‘mean concavity’ are significantly elevated. This pattern is consistent with a malignant diagnosis.

Diagnosis: 1

Now, analyze the following case:

Based on the 30 features provided: [FEATURE_LIST_TEST], classify the diagnosis as 0 (Benign) or 1 (Malignant). Let’s think step by step, considering that features like ‘worst concave points’ and ‘worst perimeter’ are highly indicative of malignancy. Your final answer must be only the numerical label.”

Formal Representation (for KE-CoT-ExR):

$$Prompt_{KE-CoT-ExR} = R_k \oplus f_{KE-CoT}(x_{test})$$

Where R_k represents k examples that include reasoning steps.

- **5-Shot KE-CoT (5S-KE-CoT):** This variant extends the KE-CoT framework to five in-context examples. Five class-balanced labeled examples are prepended to the knowledge-enhanced query with CoT instruction. The increased number of exemplars provides the model with a richer pattern base for in-context learning while maintaining the domain-aware reasoning framework.
- **5-Shot KE-CoT with Example Reasoning (5S-KE-CoT-ExR):** This represents the most comprehensive prompting strategy, combining five examples—each accompanied by human-written rationales—with knowledge-enhanced framing and CoT instruction. The template structure mirrors that of 3S-KE-CoT-ExR but with two additional exemplars to provide broader coverage of diagnostic patterns.

4. Large Language Models

Twelve LLMs spanning two deployment categories were evaluated. Three proprietary API-based models were accessed through commercial services: GPT-4.1 and GPT-4o via the OpenAI API, and Gemini 2.5 Flash Lite via the Google Gemini API. The remaining nine models were deployed locally using Hugging Face Transformers or llama.cpp on NVIDIA GPU infrastructure. Two proprietary API-based models were accessed via the OpenAI API: GPT-4.1 and GPT-4o, both belonging to OpenAI’s latest GPT-4 series with approximately 1.8 trillion parameters and large context windows (128K–1M tokens) [39]. Qwen 2.5 7B Instruct, developed by Alibaba Cloud, is an instruction-tuned model with 7 billion parameters and a 128K-token context window, designed for multilingual understanding and reasoning tasks [40]. Gemma-3-4B-IT, developed by Google, is a 4 billion parameter instruction-tuned model from the Gemma 3 family, optimized for

efficient inference on resource-constrained hardware [41]. Mistral 7B Instruct v0.3, developed by Mistral AI, is a 7 billion parameter instruction-tuned model with a 32K-token context window, known for its efficiency and strong performance on code generation and multi-language tasks [42]. Phi-3 Medium 14B, developed by Microsoft, is a 14 billion parameter model from the Phi-3 family with a 128K-token context window, designed to deliver strong reasoning capabilities relative to its compact size [43]. DeepSeek-R1-Distill-Llama-8B, developed by DeepSeek, is an 8 billion parameter model that distills the reasoning capabilities of larger DeepSeek-R1 models into a Llama-based architecture, specifically designed for enhanced step-by-step reasoning [44]. GPT-OSS-20B is a 20 billion parameter open-source model built upon the GPT architecture, providing a mid-scale alternative for evaluating generative reasoning on structured data [45]. Qwen3-4B-Thinking-Gemini-VIBE is a 4 billion parameter community fine-tuned variant based on Qwen3 architecture, incorporating enhanced reasoning and instruction-following capabilities through specialized training [46]. Llama 3.1 8B Instruct, developed by Meta, is an open-source transformer-based language model with 8 billion parameters and a 128K-token context window, designed for generating coherent, human-like responses and applicable to a wide range of NLP tasks including translation and summarization [47], [48]. Gemini 2.5 Flash Lite, developed by Google, is a lightweight variant of the Gemini 2.5 Flash model optimized for speed and efficiency, supporting multimodal processing capabilities [49]. Cogito-v1-preview-qwen-32B, developed by Deep Cogito, is a 32 billion parameter model based on the Qwen architecture, designed for enhanced reasoning and analytical capabilities with a focus on complex multi-step tasks [50]. Detailed hyperparameter configurations are provided in Table III.

TABLE III. HYPERPARAMETER CONFIGURATION FOR EVALUATED LLMs

LLM	Temp	Top_p	API / Service
GPT-4.1	0.6	0.95	OpenAI
GPT-4o	0.6	0.95	OpenAI
Qwen 2.5 7B Instruct	0.5	0.9	HF Transformers (local A100)
Gemma-3-4B-IT	0.0	0.9	HF Transformers (local L4 GPU)
Mistral 7B v0.3	0.0	0.9	HF Transformers (local A100)
Phi-3 Medium 14B	0.0	0.9	HF Transformers (local A100)
DeepSeek-R1 8B	0.6	0.95	HF Transformers (local L4)
GPT-OSS-20B	0.6	0.95	HF Transformers (local RTX PRO 6000)
Qwen3-4B-Think.	0.6	0.95	HF Transformers (local A100)
Llama 3.1 8B Instruct	0.6	0.95	Local llama.cpp (CUDA, A100 80GB)
Gemini 2.5 FL	0.0	1.0	Google Gemini API
Cogito-v1 32B	0.6	0.95	Local llama.cpp (CUDA, A100 40GB)

5. LLMs API Configuration and Parameter Optimization

In this study, we evaluated a diverse set of 12 LLMs to assess their performance on BrC classification using structured patient data. The models were accessed via two deployment approaches: (i) proprietary API-based access through the OpenAI platform for GPT-4.1 and GPT-4o, and (ii) local deployment of open-source models using Hugging Face Transformers and llama.cpp inference frameworks on dedicated GPU infrastructure. The selection of hyperparameters for each model was a critical step, guided by a combination of established best practices for factual tasks, preliminary testing, and the default configurations recommended by each framework to ensure both reproducibility and optimal performance [51]. A comprehensive summary of the specific configurations for each model is provided in Table III. The temperature parameter, which controls the randomness of the output, was carefully chosen based on the task’s requirements [52]. For models

TABLE IV. F1-score (%) COMPARISON ACROSS PROMPTING STRATEGIES — WBCD (N=569)

Model	ZS-D	ZS-T	ZS-C	3S	5S	10S	ZS-CoT	KE-ZS-CoT	3S-CoT	3S-KE-CoT	3S-KE-CoT-ExR	5S-KE-CoT	5S-KE-CoT-ExR
GPT-4.1	64.30	60.02	60.12	67.41	69.96	70.40	61.37	65.19	68.15	70.38	70.09	72.32	71.23
GPT-4o	63.72	60.17	60.45	68.01	69.35	69.80	61.95	65.78	67.56	69.78	69.49	71.71	70.6
Qwen 2.5 7B	25.93	76.52	77.11	77.09	77.80	77.91	77.11	77.11	76.69	79.37	83.68	78.18	78.35
Gemma-3 4B	54.29	54.29	54.29	57.14	68.30	79.92	14.65	32.77	23.19	0.00	76.89	70.95	81.76
Mistral 7B	15.85	0.00	13.79	77.83	84.49	83.53	30.99	36.85	59.83	70.97	63.10	76.73	84.63
Phi-3 14B	46.08	59.27	55.73	48.38	70.73	74.60	43.86	54.33	61.17	69.20	70.06	61.08	82.55
DeepSeek-R1 8B	15.94	14.09	44.40	16.08	19.26	20.42	23.53	54.12	53.18	23.93	51.36	38.86	40.51
GPT-OSS-20B	54.05	0.00	43.60	54.12	7.11	54.88	33.78	60.69	53.70	48.14	55.00	40.00	46.80
Qwen3-4B-Think.	54.05	13.79	54.17	20.77	18.08	31.97	32.99	0.00	43.81	3.72	24.83	0.00	16.73
Gemini 2.5 FL	55.60	83.87	56.49	87.89	88.05	82.57	60.06	60.37	35.16	79.72	85.18	59.47	75.18
Cogito-v1 32B	48.71	55.24	53.87	92.00	89.47	81.67	15.56	46.18	33.77	40.23	68.90	43.60	75.50
Llama 3.1 8B	39.32	52.70	46.68	59.05	59.62	48.63	54.43	54.15	55.61	66.42	69.05	62.26	74.16

TABLE V. F1-score (%) COMPARISON ACROSS PROMPTING STRATEGIES — BCCD (N=116)

Model	ZS-D	ZS-T	ZS-C	3S	5S	10S	ZS-CoT	KE-ZS-CoT	3S-CoT	3S-KE-CoT	3S-KE-CoT-ExR	5S-KE-CoT	5S-KE-CoT-ExR
GPT-4.1	75.00	76.11	74.03	76.56	81.25	81.25	69.77	76.40	80.90	83.15	83.15	85.39	80.03
GPT-4o	73.44	71.18	70.18	78.13	79.69	79.69	70.31	78.65	78.65	83.15	80.90	85.39	82.49
Qwen 2.5 7B	0.00	67.08	28.21	69.50	69.84	71.76	72.00	71.51	61.54	53.21	55.36	72.73	64.81
Gemma-3 4B	71.11	58.06	67.09	62.07	58.00	70.92	43.48	68.21	70.89	0.00	60.19	41.98	58.49
Mistral 7B	35.96	36.96	66.25	57.63	38.20	27.03	53.33	64.12	48.70	57.85	67.18	66.67	59.62
Phi-3 14B	59.09	51.38	67.11	61.29	63.06	62.96	65.75	67.69	66.67	65.45	62.50	58.93	60.00
DeepSeek-R1 8B	0.00	0.00	0.00	68.64	6.25	0.00	71.11	71.11	52.54	67.11	51.92	41.51	52.34
GPT-OSS-20B	37.78	0.00	66.67	70.86	14.08	6.56	69.36	55.81	44.86	64.56	60.56	61.43	63.64
Qwen3-4B-Think.	68.75	56.90	60.29	47.06	60.19	0.00	71.11	44.44	63.49	37.04	30.59	31.11	37.89
Gemini 2.5 FL	71.70	72.09	72.32	36.59	32.00	63.49	58.33	71.35	56.90	68.70	68.85	53.21	72.27
Cogito-v1 32B	61.19	47.52	45.10	50.00	61.68	61.95	65.84	63.64	66.23	66.25	69.23	64.97	65.08
Llama 3.1 8B	73.68	71.86	72.00	69.41	69.18	71.23	70.79	67.06	70.86	67.50	68.70	66.20	71.53

TABLE VI. F1-score (%) COMPARISON ACROSS PROMPTING STRATEGIES — MMD (N=961)

Model	ZS-D	ZS-T	ZS-C	3S	5S	10S	ZS-CoT	KE-ZS-CoT	3S-CoT	3S-KE-CoT	3S-KE-CoT-ExR	5S-KE-CoT	5S-KE-CoT-ExR
GPT-4.1	61.28	62.31	61.31	64.42	67.04	67.49	58.30	62.12	64.51	67.27	66.75	69.08	70.05
GPT-4o	60.69	62.06	62.00	65.10	65.99	66.59	58.90	62.81	63.82	66.50	65.98	68.30	70.16
Qwen 2.5 7B	21.43	78.78	83.69	73.08	82.94	83.35	75.58	65.58	83.43	71.26	82.82	80.86	82.54
Gemma-3 4B	65.20	65.47	73.24	76.42	82.83	82.21	65.72	64.24	65.28	0.00	79.23	18.51	81.44
Mistral 7B	4.69	1.47	39.07	74.32	76.55	82.93	69.45	67.94	60.61	60.72	73.87	68.79	81.20
Phi-3 14B	61.83	67.57	64.20	83.81	82.93	71.30	67.86	68.93	75.19	71.50	74.00	67.00	78.50
DeepSeek-R1 8B	0.00	65.37	0.00	0.00	63.68	0.00	66.18	65.15	33.81	70.68	65.34	48.25	65.89
GPT-OSS-20B	0.00	66.28	63.49	65.31	7.44	0.50	58.95	31.86	72.32	71.13	31.85	61.25	39.35
Qwen3-4B-Think.	47.67	0.00	0.00	34.67	0.00	0.00	62.78	36.42	48.66	63.09	40.05	57.74	48.34
Gemini 2.5 FL	65.95	66.23	66.06	76.83	80.77	80.39	66.12	68.14	79.84	75.41	82.91	75.23	81.83
Cogito-v1 32B	48.39	74.62	76.69	81.59	81.83	82.26	63.15	68.70	51.23	72.77	82.38	67.63	82.54
Llama 3.1 8B	62.58	64.54	66.50	65.46	67.63	80.27	65.15	65.69	65.42	65.41	71.05	66.61	75.20

where maximum determinism was desired - Gemma-3-4B-IT, Mistral 7B Instruct v0.3, Phi-3 Medium 14B, and Gemini 2.5 Flash Lite - a temperature of 0.0 was selected, aligning with methodologies that aim to produce consistent, deterministic, and factually relevant responses, which is crucial for a medical classification task where reliability is paramount. This setting minimizes the likelihood of the model generating creative or unexpected outputs, including hallucinations.

For the proprietary GPT models (GPT-4.1, GPT-4o), a moderate temperature of 0.6 was applied, balancing response consistency with nuanced interpretation capability. The same moderate temperature was used for several open-source models (DeepSeek-R1-Distill-Llama-8B, GPT-OSS-20B, Qwen3-4B-Thinking-Gemini-VIBE, Llama 3.1 8B Instruct, and Cogito-v1-preview-qwen-32B), established after preliminary exploration as a balanced approach that allows for nuanced interpretation while mitigating excessive output variability. For Qwen 2.5 7B Instruct, a temperature of 0.5 was used

as recommended by the model's documentation. The top_p (nucleus sampling) parameter was consistently set to a high value, typically 0.9 or 0.95, across most models to ensure that the model considers a wide yet probable distribution of possible next tokens, preventing overly restrictive or repetitive outputs. Gemini 2.5 Flash Lite was configured with top_p = 1.0 as per its default settings. The max_tokens (or max_output_tokens) parameter was configured dynamically to align with the requirements of each prompting strategy, balancing the need for sufficient detail with processing efficiency [53]. For standard zero-shot and few-shot strategies, where the expected output for the BrC diagnosis was a single binary classification (0 or 1), only a minimal number of tokens were needed to capture the response. In contrast, for CoT strategies, the output was expected to include reasoning steps prior to the final prediction. Consequently, for these scenarios, the max_tokens limit was set to a higher value to ensure the model had sufficient capacity to generate a detailed, step-by-step clarification of

TABLE VII. EVALUATION RESULTS UNDER BEST STRATEGY PER MODEL AND DATASET

Model	Best Strategy	Dataset	TP	FP	FN	TN	Acc%	Prec%	Rec%	F1%
GPT-4.1	5S-KE-CoT	WBCD	162	74	50	283	78.21	68.64	76.42	72.32
GPT-4.1	5S-KE-CoT	BCCD	38	8	5	65	88.79	82.61	88.37	85.39
GPT-4.1	5S-KE-CoT	MMD	267	148	91	455	75.13	64.34	74.58	69.08
GPT-4o	5S-KE-CoT	WBCD	161	76	51	281	77.68	67.93	75.94	71.71
GPT-4o	5S-KE-CoT	BCCD	38	8	5	65	88.79	82.61	88.37	85.39
GPT-4o	5S-KE-CoT	MMD	265	153	93	450	74.40	63.40	74.02	68.30
Qwen 2.5 7B	3S-KE-CoT-ExR	WBCD	364	141	1	63	75.04	72.08	99.73	83.68
Qwen 2.5 7B	5S-KE-CoT	BCCD	56	37	5	13	62.16	60.22	91.80	72.73
Qwen 2.5 7B	3S-CoT	MMD	390	141	14	282	81.26	73.45	96.53	83.42
Gemma-3 4B	5S-KE-CoT-ExR	WBCD	316	103	38	107	75.00	75.42	89.27	81.76
Gemma-3 4B	ZS-Direct	BCCD	32	16	10	32	71.11	66.67	76.19	71.11
Gemma-3 4B	5-Shot	MMD	384	141	18	282	80.73	73.14	95.52	82.85
Mistral 7B	5S-KE-CoT-ExR	WBCD	179	32	33	320	88.48	84.83	84.43	84.63
Mistral 7B	3S-KE-CoT-ExR	BCCD	44	22	21	44	67.18	66.67	67.69	67.18
Mistral 7B	10-Shot	MMD	380	143	13	284	80.98	72.66	96.70	82.93
Phi-3 14B	5S-KE-CoT-ExR	WBCD	175	37	37	315	86.88	82.55	82.55	82.55
Phi-3 14B	KE-ZS-CoT	BCCD	44	28	14	22	61.11	61.11	75.86	67.69
Phi-3 14B	3-Shot	MMD	389	136	14	288	81.86	74.10	96.53	83.84
DeepSeek-R1 8B	KE-ZS-CoT	WBCD	105	73	105	283	68.55	59.00	50.00	54.12
DeepSeek-R1 8B	ZS-CoT	BCCD	32	14	12	32	71.11	69.57	72.73	71.11
DeepSeek-R1 8B	3S-KE-CoT	MMD	381	303	12	121	61.44	55.70	96.95	70.68
GPT-OSS-20B	KE-ZS-CoT	WBCD	125	85	77	266	70.71	59.52	61.88	60.68
GPT-OSS-20B	3-Shot	BCCD	48	26	13	24	64.86	64.86	78.69	71.11
GPT-OSS-20B	3S-CoT	MMD	363	244	34	181	66.18	59.80	91.44	72.31
Qwen3-4B-Think.	ZS-Direct	WBCD	206	344	6	13	38.49	37.45	97.17	54.07
Qwen3-4B-Think.	ZS-CoT	BCCD	32	14	12	32	71.11	69.57	72.73	71.11
Qwen3-4B-Think.	3S-KE-CoT	MMD	316	291	79	132	54.77	52.06	80.00	63.07
Gemini 2.5 FL	5-Shot	WBCD	199	44	10	311	90.43	81.89	95.22	88.05
Gemini 2.5 FL	5S-KE-CoT-ExR	BCCD	43	15	18	35	70.27	74.14	70.49	72.27
Gemini 2.5 FL	3S-KE-CoT-ExR	MMD	359	106	42	320	82.10	77.20	89.53	82.91
Cogito-v1 32B	3-Shot	WBCD	207	31	5	326	93.67	86.97	97.64	92.00
Cogito-v1 32B	3S-KE-CoT-ExR	BCCD	45	20	20	45	69.23	69.23	69.23	69.23
Cogito-v1 32B	5S-KE-CoT-ExR	MMD	383	144	18	280	80.36	72.68	95.51	82.54
Llama 3.1 8B	5S-KE-CoT-ExR	WBCD	179	59	66	265	78.03	75.21	73.06	74.12
Llama 3.1 8B	ZS-Direct	BCCD	42	11	19	42	73.68	79.25	68.85	73.68
Llama 3.1 8B	10-Shot	MMD	385	162	27	239	76.75	70.38	93.45	80.29

TABLE VIII. SPECIFICITY AND MCC VALUES (BEST STRATEGY)

Model	WBCD Spec.	WBCD MCC	BCCD Spec.	BCCD MCC	MMD Spec.	MMD MCC
GPT-4.1	0.7927	0.5465	0.8904	0.7640	0.7541	0.4880
GPT-4o	0.7871	0.5360	0.8904	0.7640	0.7463	0.4740
Qwen 2.5 7B	0.3088	0.4132	0.2600	0.2403	0.6667	0.6558
Gemma-3 4B	0.7379	0.5401	0.6667	0.4286	0.6667	0.6483
Mistral 7B	0.9091	0.7584	0.6667	0.3435	0.6651	0.6543
Phi-3 14B	0.8949	0.7205	0.4400	0.2104	0.6729	0.6725
DeepSeek-R1 8B	0.7294	0.2974	0.6957	0.4229	0.2854	0.3061
GPT-OSS-20B	0.7578	0.3547	0.4800	0.2831	0.4259	0.3697
Qwen3-4B-Think.	0.0364	0.0315	0.6957	0.4229	0.3121	0.1264
Gemini 2.5 FL	0.8761	0.8077	0.7000	0.4033	0.7512	0.6512
Cogito-v1 32B	0.9132	0.8654	0.6923	0.3846	0.6604	0.6473
Llama 3.1 8B	0.8179	0.5323	0.7925	0.4810	0.5960	0.5692

its diagnostic reasoning without being prematurely cut off. All API-based evaluations were conducted programmatically using Python clients. Open-source models were deployed locally using Hugging Face Transformers (with AutoModelForCausalLM and pipeline APIs) or llama.cpp (via llama-cpp-python with CUDA acceleration). Inference processes were executed on cloud server infrastructure equipped with NVIDIA A100 80GB, A100 40GB, L4, and RTX PRO 6000 Blackwell GPUs to ensure efficient and timely results across all models.

C. Performance Evaluation Metrics

Classification performance was assessed using Accuracy, Precision, Recall, F1-Score, Specificity, and Matthews Correlation Coefficient (MCC) derived from the confusion matrix [54] [55]. MCC was included as it provides a balanced score even with imbalanced classes. Statistical significance was determined using chi-squared tests of independence ($\alpha=0.05$) on 2×2 contingency tables of correct vs. incorrect predictions.

IV. RESULTS AND DISCUSSION

This section presents a detailed analysis of the performance of 12 LLMs across three distinct BrC-related datasets: WBCD, BCCD, and MMD. The evaluation was conducted using 13 prompting strategies, including three zero-shot variants, few-shot prompting, and multiple Chain-of-Thought (CoT) and knowledge-enhanced prompting schemes. The results are presented in Tables IV–XI. The analysis emphasizes four aspects: (i) prompt sensitivity, (ii) best-strategy performance per model and dataset, (iii) reliability through specificity and MCC, and (iv) statistical, robustness, and baseline comparisons.

A. Effect of Prompting Strategies

Tables IV–VI present the F1-score results for all models across the evaluated prompting strategies on WBCD, BCCD, and MMD. A central finding of this study is that no single prompting strategy is universally optimal across all models and datasets. Instead, the effect of prompting is strongly model-dependent and dataset-dependent. For the GPT family, a relatively consistent progression is observed from zero-shot baselines toward few-shot and knowledge-enhanced prompting. For example, on WBCD, GPT-4.1 improved from 64.30% under zero-shot direct prompting to 69.96% under 5-shot prompting and to 72.32% under 5-shot KE-CoT. A similar pattern was observed on BCCD and MMD, confirming that in-context examples and domain-aware prompting can substantially improve proprietary models. However, this trend does not generalize to all models. Several open-source and locally deployed models exhibited highly non-monotonic behavior across prompting strategies. For example, Cogito-v1-preview-qwen-32B achieved its best WBCD result under a simple 3-shot strategy (F1 = 92.00%), whereas more elaborate CoT-based variants did not improve its performance. Likewise, Gemini 2.5 Flash Lite achieved its strongest WBCD performance under standard 5-shot prompting (F1 = 88.05%), while its 3-shot CoT result dropped sharply to 35.16%. These observations suggest that prompt complexity does not automatically translate into better performance.

Another important result is that unguided CoT prompting often degraded performance, especially for stronger closed models. For example, GPT-4.1 on WBCD dropped from 64.30% under zero-shot direct prompting to 61.37% under zero-shot CoT. This pattern underscores the importance of sensitivity analysis in prompt-driven evaluations, as naive reasoning prompts may destabilize outputs rather than improve them. Finally, increasing the number of examples from 5-shot to 10-shot did not consistently improve performance. In many settings, gains were marginal or absent, and in some cases performance declined. This indicates diminishing returns from longer prompts and suggests that carefully selected compact few-shot contexts may be more effective than simply increasing context length.

B. Best-Strategy Evaluation Results

To provide a clearer comparison across models, Table VII reports the best-performing strategy for each model and dataset together with the corresponding confusion-matrix metrics. A major result of this study is that the best-performing model differs by dataset. On WBCD, Cogito-v1-preview-qwen-32B achieved the highest overall performance, reaching 93.67% accuracy and 92.00% F1 under a 3-shot strategy. This performance is substantially above GPT-4.1 on the same dataset (78.21% accuracy; 72.32% F1). On BCCD, GPT-4.1 and GPT-4o jointly achieved the strongest results, each reaching 88.79% accuracy and 85.39% F1 under 5-shot knowledge-enhanced prompting. On MMD, Gemini 2.5 Flash Lite achieved the highest accuracy (82.10%) with an F1-score of 82.91% under 3-shot KE-CoT-ExR. These findings highlight a nuanced picture: performance is data-modality-specific, and some local open-source models outperform proprietary models on certain structured datasets.

C. Specificity and MCC

Table VIII reports specificity and Matthews Correlation Coefficient (MCC) for each model under its best-performing strategy. These metrics are essential because several models achieved high recall by overpredicting the positive class, which would not be acceptable in a medical setting. On WBCD, Cogito-v1-preview-qwen-32B achieved the highest specificity (0.9132) and MCC (0.8654), confirming that its strong F1-score reflects balanced classification rather than class bias. Mistral 7B v0.3 and Gemini 2.5 Flash Lite also achieved high WBCD specificity and MCC. In contrast, Qwen3-4B-Thinking-Gemini-VIBE showed extremely poor specificity (0.0364), indicating that it classified almost all cases as malignant. On BCCD, GPT-4.1 and GPT-4o achieved the highest specificity (0.8904) and MCC (0.7640), reinforcing their strong performance on biomarker-based features. On MMD, the highest specificity was obtained by GPT-4.1 (0.7541), while Gemini 2.5 Flash Lite and several open-source models achieved competitive MCC values. These results show that F1 alone is insufficient for interpreting model quality in structured clinical classification.

D. Statistical Significance

To determine whether observed differences between models were statistically meaningful, chi-squared tests were performed against the best model for each dataset. Table IX summarizes the p-values. On WBCD, Cogito-v1-preview-qwen-32B significantly outperformed all other models, including GPT-4.1 and GPT-4o. On BCCD, GPT-4.1 and GPT-4o were statistically indistinguishable, while most other models performed significantly worse. On MMD, Gemini 2.5 Flash Lite served as the reference model, but several models—including Qwen3-4B-Thinking-Gemini-VIBE, Gemma-3-4B-IT, Mistral 7B Instruct v0.3, Phi-3 Medium 14B, and Cogito-v1-preview-qwen-32B—were not significantly different from it. This confirms that top performance on MMD is shared by a broader cluster of models than on WBCD or BCCD.

TABLE IX. CHI-SQUARED SIGNIFICANCE TESTS (VS. BEST MODEL PER DATASET)

Model	WBCD (Ref: Cogito-v1) p-value	BCCD (Ref: GPT-4.1) p-value	MMD (Ref: Gemini 2.5 FL) p-value
GPT-4.1	<0.0001 *	[Baseline / Ref]	0.0007 *
GPT-4o	<0.0001 *	1.0000 ns	<0.0001 *
Qwen 2.5 7B Inst.	<0.0001 *	<0.0001 *	0.6903 ns
Gemma-3-4B-IT	<0.0001 *	0.0021 *	0.5012 ns
Mistral 7B v0.3	0.0034 *	<0.0001 *	0.5934 ns
Phi-3 Medium 14B	0.0001 *	<0.0001 *	0.9412 ns
DeepSeek-R1-Distill-Llama-8B	<0.0001 *	0.0021 *	<0.0001 *
GPT-OSS-20B	<0.0001 *	<0.0001 *	<0.0001 *
Qwen3-4B-Think.	<0.0001 *	0.0021 *	<0.0001 *
Gemini 2.5 FL	0.0452 *	0.0011 *	[Baseline / Ref]
Cogito-v1 32B	[Baseline / Ref]	<0.0001 *	0.3855 ns
Llama 3.1 8B	<0.0001 *	0.0041 *	0.0092 *

* p < 0.05; ns = not significant

E. Comparison With Traditional ML Baselines

Table X compares the strongest LLM result on each dataset against published state-of-the-art ML/DL baselines. The results show that LLMs still do not outperform specialized supervised methods. On WBCD, the gap remains large, with the shallow ANN baseline achieving 99.84% F1 versus 92.00% for the best LLM configuration. On BCCD, the gap is smaller but still clear (92.31% vs. 85.39%). On MMD, however, the difference narrows considerably (84.95% vs. 82.91%), suggesting that LLMs can become competitive on some structured radiology-style datasets even without task-specific training. Nevertheless, this

comparison remains asymmetric. The traditional baselines are fully supervised systems trained and tuned on the target task, whereas the LLMs in this study are used without fine-tuning. Furthermore, the robustness analysis (Table 11) revealed that the mean F1 on MMD across multiple example sets was 79.62%, widening the gap with the ML baseline to approximately 5.3 percentage points and indicating that single-run comparisons should be interpreted cautiously.

TABLE X. COMPARISON WITH STATE-OF-THE-ART ML/DL BASELINES

Dataset	Study	Year	Method	Acc%	F1%
WBCD	Alshayegi et al. [56]	2022	Shallow ANN	99.85	99.84
WBCD	Our Study	2026	Cogito-v1 32B (3-Shot)	93.67	92.00
BCCD	Alsabry et al. [57]	2025	Subspace KNN (ADASYN)	91.70	92.31
BCCD	Our Study	2026	GPT-4.1 (5S-KE-CoT)	88.79	85.39
MMD	Bacha et al. [58]	2025	RF-RE-SMOTEBoost	84.97	84.95
MMD	Our Study	2026	Gemini 2.5 FL (3S-KE-CoT-ExR)	82.10	82.91

F. Robustness Analysis

Because the main findings of this study depend on prompt engineering and few-shot example selection, it is essential to assess whether performance is stable across different demonstration sets or represents an artifact of a single favorable configuration. To this end, the best model-strategy pair for each dataset was re-evaluated using three independent class-balanced example sets (Sets A, B, and C), each constructed through random sampling with a different seed. Table XI reports the complete results. On WBCD, Cogito-v1-preview-qwen-32B with 3-Shot prompting demonstrated highly stable performance, achieving a mean F1-score of $91.37 \pm 0.71\%$ and a mean MCC of 0.8528 ± 0.0138 . The coefficient of variation (CV) for F1 was below 1%, confirming that the strong classification performance is not dependent on a particular example selection. On BCCD, GPT-4.1 with 5S-KE-CoT exhibited moderate stability, with a mean F1-score of $87.46 \pm 1.91\%$ and a mean MCC of 0.7863 ± 0.0207 . Notably, recall showed higher variability ($94.10 \pm 5.03\%$), indicating that the model's sensitivity to the positive class fluctuated across example sets while precision remained relatively stable ($82.66 \pm 0.54\%$). On MMD, Gemini 2.5 Flash Lite with 3S-KE-CoT-ExR showed the greatest sensitivity to example selection. While Set A achieved an accuracy of 82.10% and F1 of 82.91%, Sets B and C dropped to 73.76% and 74.85% accuracy, respectively, yielding a mean F1 of $79.62 \pm 2.87\%$ and a mean accuracy of $76.90 \pm 4.53\%$. This approximately 8-percentage-point accuracy drop between the best and worst runs highlights that prompt-driven performance on MMD is notably dependent on the specific examples provided. This finding underscores a practical limitation of few-shot LLM inference: the model's behavior can vary substantially with different demonstration contexts, and single-run results should be interpreted with caution. Overall, the robustness analysis confirms that WBCD results are highly reliable, BCCD results are moderately stable, and MMD results exhibit non-negligible sensitivity to few-shot example selection. These observations directly inform the practical recommendation that any deployment of prompt-based LLM classification should incorporate multiple example sets and report aggregated performance rather than relying on a single favorable configuration.

V. CONCLUSION

This study presented a comparative evaluation of 12 LLMs for breast-cancer-related classification across three datasets and 13 prompting strategies. The experiments show that there is no single universally

TABLE XI. ROBUSTNESS ANALYSIS ACROSS DIFFERENT FEW-SHOT EXAMPLE SETS

Dataset	Set	Acc%	Prec%	Rec%	F1%	MCC
WBCD	A	93.67	86.97	97.64	92.00	0.8654
WBCD	B	92.05	85.10	96.80	90.60	0.8380
WBCD	C	93.17	86.50	97.20	91.50	0.8550
WBCD	M±S	92.96 ± 0.83	86.19 ± 0.97	97.21 ± 0.42	91.37 ± 0.71	0.8528 ± 0.0138
BCCD	A	88.79	82.61	88.37	85.39	0.7640
BCCD	B	90.52	83.23	97.75	89.14	0.8050
BCCD	C	88.79	82.15	96.18	87.86	0.7900
BCCD	M±S	89.37 ± 1.00	82.66 ± 0.54	94.10 ± 5.03	87.46 ± 1.91	0.7863 ± 0.0207
MMD	A	82.10	77.20	89.53	82.91	0.6512
MMD	B	73.76	66.20	93.77	77.61	0.5247
MMD	C	74.85	67.26	93.77	78.33	0.5425
MMD	M±S	76.90 ± 4.53	70.22 ± 6.06	92.36 ± 2.45	79.62 ± 2.87	0.5728 ± 0.0686

best model. Instead, the optimal model-strategy pair is dataset-dependent: Cogito-v1-preview-qwen-32B performed best on WBCD, GPT-4.1 and GPT-4o were strongest on BCCD, and Gemini 2.5 Flash Lite achieved the leading result on MMD. The results also show that prompt engineering plays a decisive but non-uniform role. Proprietary GPT models generally benefited from knowledge-enhanced few-shot prompting, whereas some open-source models reached their best performance with simpler 3-shot or 5-shot strategies. In addition, several smaller models showed substantial instability across prompts, confirming that prompt sensitivity must be considered explicitly in structured medical evaluations.

When compared with state-of-the-art traditional ML/DL baselines, the best LLM configurations still remain below specialized supervised methods. However, the performance gap varies by dataset and becomes relatively narrow on MMD for the best single run. The robustness analysis further revealed that performance stability itself is dataset-dependent: WBCD results were highly consistent (F1 CV < 1%), whereas MMD showed non-negligible variation across different example sets (F1 standard deviation of 2.87 percentage points), indicating that prompt-based approaches require careful validation before practical deployment. Taken together, these findings indicate that LLMs should not currently be regarded as replacements for task-specific supervised diagnostic systems, but rather as flexible exploratory decision-support models that can be evaluated without task-specific retraining. Despite the improved results, false negatives and false positives remain clinically significant for several datasets and models. Therefore, the proposed approach should not be positioned as an autonomous screening or early warning system. Instead, it should be viewed as an exploratory AI-assisted baseline requiring expert oversight, calibration, and prospective validation before any clinical use.

In summary, this study provides a comprehensive benchmark for LLM-based structured clinical classification, demonstrating that while LLMs can approximate supervised baselines in some settings, their prompt-sensitive behavior and residual error rates necessitate expert oversight for any clinical application. Future work should extend the robustness evaluation to a larger number of seeds and additional datasets, investigate uncertainty calibration, explore the integration of tabular-specific architectures with language model reasoning, and extend the evaluation to larger and more diverse clinical cohorts for safer and more reliable medical decision support.

CREDiT AUTHORSHIP CONTRIBUTION STATEMENT

Habibe Karayıgıt: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Project administration, Writing – original draft.

Filiz Kalelioğlu: Conceptualization, Supervision, Project administration, Validation, Writing – review and editing.

DATA STATEMENT

The three primary breast cancer datasets used in the current study are publicly available from the UCI Machine Learning Repository. The WBCD dataset, which is the core dataset based on cytological features, can be obtained from <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>. The BCCD, which focuses on metabolic biomarkers, is available at <https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra>. Finally, the MMD, a dataset derived from features in mammography reports, is accessible via the link <https://archive.ics.uci.edu/dataset/161/mammographic+mass>.

DECLARATION OF CONFLICTS OF INTEREST

We have no conflict of interest to declare.

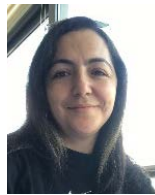
ACKNOWLEDGMENT

The authors did not receive support from any organization for the submitted work.

REFERENCES

- [1] M. Methamem, I. Ghadhab, S. Hidar, and R. Briki, "Breast Cancer in Men: A Serie of 45 Cases and Literature Review," *Pan African Medical Journal*, vol. 36, no. 183, p. 36, 2020, doi: <https://doi.org/10.11604/pamj.2020.36.183.22574>
- [2] World Health Organization, "Breast cancer." Accessed: Jul. 06, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>
- [3] S. Łukasiewicz, M. Czelelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies," *Cancers*, vol. 13, no. 17, p. 4287, 2021, doi: <https://doi.org/10.3390/cancers13174287>
- [4] G. I. Balali, "Breast Cancer: A Review of Mammography and Clinical Breast Examination for Early Detection of Cancer," *Open Access Library Journal*, vol. 7, no. 10, 2020. [Online]. Available: https://www.scirp.org/html/103621_103621.htm
- [5] National Cancer Institute, "Female Breast Cancer – Cancer Stat Facts." Accessed: Jul. 06, 2025. [Online]. Available: <https://seer.cancer.gov/statfacts/html/breast.html>
- [6] N. M. Kuderer, D. C. Dale, J. Crawford, L. E. Cosler, and G. H. Lyman, "Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients," *Cancer*, vol. 106, no. 10, pp. 2258–2266, May 2006, doi: <https://doi.org/10.1002/cncr.21847>
- [7] A. I. Ionescu et al., "Prediction of Subclinical and Clinical Multiple Organ Failure Dysfunction in Breast Cancer Patients—A Review Using AI Tools," *Cancers*, vol. 16, no. 2, p. 381, Jan. 2024, doi: <https://doi.org/10.3390/cancers16020381>
- [8] C. Vigneron, J. Charpentier, F. Coussy, J. Alexandre, F. Pène, and M. Jamme, "When breast cancer comes to the ICU: outcomes and prognostic factors," *Acta Oncologica*, vol. 62, no. 4, pp. 358–363, 2023, doi: <https://doi.org/10.1080/0284186X.2023.2197123>
- [9] M. Y. Yan, L. T. Gustad, and Ø. Nytrø, "Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review," *Journal of the American Medical Informatics Association*, vol. 29, no. 3, pp. 559–575, Jan. 2022, doi: <https://doi.org/10.3390/biomedinformatics4020062>
- [10] K. Nassiri and M. A. Akhloufi, "Recent Advances in Large Language Models for Healthcare," *BioMedInformatics*, Vol. 4, Pages 1097–1143, vol. 4, no. 2, pp. 1097–1143, Apr. 2024, doi: <https://doi.org/10.3390/biomedinformatics4020062>
- [11] J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, "Large Language Models in Medicine: The Potentials and Pitfalls," *arXiv preprint arXiv:2309.00087*, Aug. 2023. Accessed: Feb. 09, 2024. [Online]. Available: <https://arxiv.org/abs/2309.00087>
- [12] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," in *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2020, pp. 179–183, doi: <https://doi.org/10.15439/2020F20>
- [13] H. Huang et al., "ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model," *International Journal of Oral Science*, vol. 15, no. 1, pp. 1–13, Jul. 2023, doi: <https://doi.org/10.1038/s41368-023-00239-y>
- [14] A. Egli, "ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology?," *Clinical Infectious Diseases*, vol. 77, no. 9, pp. 1322–1328, Nov. 2023, doi: <https://doi.org/10.1093/cid/ciad407>
- [15] H. Y. Zhou et al., "A Transformer-Based Representation-Learning Model with Unified Processing of Multimodal Input for Clinical Diagnostics," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 743–755, Jun. 2023, doi: <https://doi.org/10.1038/s41551-023-01045-x>
- [16] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D. K. Kim, "Deep Learning Application to Clinical Decision Support System in Sleep Stage Classification," *Journal of Personalized Medicine*, vol. 12, no. 2, p. 136, Jan. 2022, doi: <https://doi.org/10.3390/jpm12020136>
- [17] T. H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: <https://doi.org/10.1371/journal.pdig.0000198>
- [18] S. Liu et al., "Assessing the Value of ChatGPT for Clinical Decision Support Optimization," *medRxiv preprint*, Feb. 2023, doi: <https://doi.org/10.1101/2023.02.21.23286254>
- [19] A. Rao et al., "Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot," *Journal of the American College of Radiology*, vol. 20, no. 10, pp. 990–997, Oct. 2023, doi: <https://doi.org/10.1016/j.jacr.2023.05.003>
- [20] A. Rao et al., "Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow," *medRxiv preprint*, 2023, doi: <https://doi.org/10.1101/2023.02.21.23285886>
- [21] J. Lee et al., "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: <https://doi.org/10.1093/bioinformatics/btz682>
- [22] R. Luo et al., "BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining," *Briefings in Bioinformatics*, vol. 23, no. 6, pp. 1–11, Nov. 2022, doi: <https://doi.org/10.1093/bib/bbac409>
- [23] Y. Jiang and C. Poellabauer, "Medication Error Detection Using Contextual Language Models," *Studies in Computational Intelligence*, vol. 1060, pp. 91–99, 2023, doi: https://doi.org/10.1007/978-3-031-14771-5_7
- [24] D. P. Panagoulas, F. A. Palamidis, M. Virvou, and G. A. Tsihrintzis, "Evaluating the Potential of LLMs and ChatGPT on Medical Diagnosis and Treatment," in *Proceedings of the 14th International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2023, pp. 1–9, doi: <https://doi.org/10.1109/IISA59645.2023.10345968>
- [25] A. P. Javidan, T. Feridooni, L. Gordon, and S. A. Crawford, "Evaluating the Progression of Artificial Intelligence and Large Language Models in Medicine Through Comparative Analysis of ChatGPT-3.5 and ChatGPT-4 in Generating Vascular Surgery Recommendations," *JVS-Vascular Insights*, p. 100049, Dec. 2023, doi: <https://doi.org/10.1016/j.jvsvi.2023.100049>
- [26] A. Čartolovni, A. Tomićić, and E. Lazić Mosler, "Ethical, Legal, and Social Considerations of AI-Based Medical Decision-Support Tools: A Scoping Review," *International Journal of Medical Informatics*, vol. 161, p. 104738, May 2022, doi: <https://doi.org/10.1016/j.ijmedinf.2022.104738>
- [27] B. Vasey et al., "Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI," *Nature Medicine*, vol. 28, no. 5, pp. 924–933, May 2022, doi: <https://doi.org/10.1038/s41591-022-01772-9>

- [28] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, Feb. 2020, doi: <https://doi.org/10.1038/s41746-020-0221-y>
- [29] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proceedings of SPIE - Biomedical Image Processing and Biomedical Visualization*, vol. 1905, Jul. 1993, pp. 861–870, doi: <https://doi.org/10.1117/12.148698>
- [30] W. H. Wolberg and O. L. Mangasarian, "Breast Cancer Wisconsin (Original) Data Set," UCI Machine Learning Repository, 1992, [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Original\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))
- [31] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep Convolutional Neural Networks for breast cancer screening," *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 19–30, Apr. 2018, doi: <https://doi.org/10.1016/j.cmpb.2018.01.011>
- [32] S. Karthik, R. Srinivasa Perumal, and P. V. S. S. R. Chandra Mouli, "Breast Cancer Classification Using Deep Neural Networks," in *Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques*, Volume 1, Singapore: Springer, Jan. 2018, pp. 227–241, doi: https://doi.org/10.1007/978-981-10-6680-1_12
- [33] UCI Machine Learning Repository, "Breast Cancer Wisconsin (Original) Data Set." Accessed: Oct. 27, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>
- [34] UCI Machine Learning Repository, "Breast Cancer Coimbra Data Set." Accessed: Oct. 27, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra>
- [35] UCI Machine Learning Repository, "Mammographic Mass Data Set." Accessed: Oct. 27, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/161/mammographic+mass>
- [36] DAIR.AI, "Prompt Engineering Guide." Accessed: Feb. 14, 2024. [Online]. Available: <https://www.promptingguide.ai/>
- [37] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, Jan. 2023, doi: <https://doi.org/10.1145/3560815>
- [38] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 35, Dec. 2022, pp. 24824–24837
- [39] OpenAI, "Models - OpenAI API." Accessed: Feb. 14, 2024. [Online]. Available: <https://platform.openai.com/docs/models>
- [40] A. Yang et al., "Qwen2 Technical Report," *arXiv preprint arXiv:2407.10671*, Sep. 2024. Accessed: Mar. 27, 2026. [Online]. Available: <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
- [41] Google, "gemma-3-4b-it." Hugging Face. Accessed: Mar. 27, 2026. [Online]. Available: <https://huggingface.co/google/gemma-3-4b-it>
- [42] Mistral AI, "Mistral-7B-Instruct-v0.3." Hugging Face, 2024. Accessed: Feb. 14, 2024, [Online]. Available: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>
- [43] Microsoft, "Phi-3-medium-14B-instruct." Hugging Face, 2024. Accessed: Mar. 27, 2026, [Online]. Available: <https://huggingface.co/microsoft/Phi-3-medium-128k-instruct>
- [44] DeepSeek AI, "DeepSeek-R1-Distill-Llama-8B." Hugging Face. Accessed: Mar. 27, 2026. [Online]. Available: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>
- [45] OpenAI, "gpt-oss-20b." Hugging Face, 2025. Accessed: Mar. 27, 2026, [Online]. Available: <https://huggingface.co/openai/gpt-oss-20b>
- [46] TeichAI, "Qwen3-4B-Thinking-2507-Gemini-3-Flash-VIBE." Hugging Face. Accessed: Mar. 27, 2026. [Online]. Available: <https://huggingface.co/TeichAI/Qwen3-4B-Thinking-2507-Gemini-3-Flash-VIBE>
- [47] Meta, "Llama-3.1-8B-Instruct." Hugging Face, 2024. Accessed: Feb. 14, 2024, [Online]. Available: <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>
- [48] Meta, "Llama-2-7b-chat-hf." Hugging Face. Accessed: Feb. 14, 2024. [Online]. Available: <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>
- [49] Google AI, "Gemini 2.5 Flash-Lite | Gemini API." Accessed: Mar. 27, 2026. [Online]. Available: <https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite>
- [50] DeepCogito, "cogito-v1-preview-qwen-32B." Hugging Face. Accessed: Mar. 27, 2026. [Online]. Available: <https://huggingface.co/deepcogito/cogito-v1-preview-qwen-32B>
- [51] S. Saha, A. Chaturvedi, J. Mahapatra, and U. Garain, "sudoLLM: On Multi-role Alignment of Language Models," *arXiv preprint arXiv:2505.14607*, 2025. Accessed: May 28, 2025. [Online]. Available: <https://arxiv.org/abs/2505.14607>
- [52] A. Dunton, R. Vyas, H. Shah, S. Sankathala, and M. Masum, "InvadeAI: Interactive Advertising with Multimodal AI," in *Proceedings of the IEEE Conference on Artificial Intelligence (CAI)*, 2025, pp. 126–131.
- [53] J. Sun, Y. Tang, A. Li, C. J. Maddison, and K. S. Meel, "Enumerate-Conjecture-Prove: Formally Solving Answer-Construction Problems in Math Competitions," *arXiv preprint arXiv:2505.18492*, 2025. Accessed: May 28, 2025. [Online]. Available: <https://arxiv.org/abs/2505.18492>
- [54] H. Karayigit, Ç. İ. Aci, and A. Akdagli, "Detecting Abusive Instagram Comments in Turkish Using Convolutional Neural Network and Machine Learning Methods," *Expert Systems with Applications*, vol. 174, p. 114802, Jul. 2021, doi: <https://doi.org/10.1016/j.eswa.2021.114802>
- [55] H. Karayigit, A. Akdagli, and Ç. İ. Aci, "Homophobic and Hate Speech Detection Using Multilingual-BERT Model on Turkish Social Media," *Information Technology and Control*, vol. 51, no. 2, pp. 356–375, Jun. 2022, doi: <https://doi.org/10.5755/j01.itc.51.2.29988>
- [56] M. H. Alshayegi, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, p. 103141, Jan. 2022, doi: <https://doi.org/10.1016/j.bspc.2021.103141>
- [57] A. Alsabry, A. A. Shujaaddeen, and M. A. A. Mosleh, "Improving Breast Cancer Prediction Using Adaptive Synthetic Sampling: A Study on the Coimbra Dataset," in *Proceedings of the 5th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2025, doi: <https://doi.org/10.1109/eSmarTA66764.2025.11132130>
- [58] S. Bacha, H. Ning, M. Belarbi, D. S. Sarwatt, and S. Dhelim, "A novel double pruning method for imbalanced data using information entropy and Roulette wheel selection for breast cancer diagnosis," *Knowledge-Based Systems*, vol. 330, p. 114403, Nov. 2025, doi: <https://doi.org/10.1016/j.knsys.2025.114403>



Habibe Karayigit

Habibe Karayigit received the B.Sc. degree in Electronics and Computer Education from Firat University, Elazığ, Türkiye, in 2002, a second B.Sc. degree in Computer Engineering from Iskenderun Technical University, Hatay, Türkiye, in 2023, the M.Sc. degree in Electrical and Electronics Engineering from Mustafa Kemal University, Hatay, Türkiye, in 2014, and the Ph.D. degree in the same field from Mersin University, Mersin, Türkiye, in 2022. She has been with the Ministry of National Education of Türkiye as a computer teacher since 2002, and since 2020 she has served as a subject-matter expert in the TÜBİTAK Science High School Textbook Preparation Committee. Her research interests include large language models, transformer-based architectures, attention-based deep learning, sentiment and opinion mining, hate-speech and abusive-language detection, data balancing techniques, computer vision, time-series forecasting, and machine learning, with applications to Turkish-language understanding, social-media analytics, and medical informatics.



Filiz Kalelioğlu

Filiz Kalelioğlu is a Professor in the Faculty of Education at Başkent University, Türkiye, specializing in Educational Technology. She holds a Ph.D., M.S., and B.S. in Educational Technology, as well as a second M.S. degree in Computer Engineering. In 2017, she was awarded a TÜBİTAK fellowship and served as a visiting scholar at King's College London. Her research interests include technology-enhanced learning, learning design, technology integration, computer science education, data mining, learning analytics, and machine learning. Dr. Kalelioğlu has authored numerous publications in national and international journals, as well as book chapters in edited volumes.