

Research Article

An Ethical AI Approach for Disease Diagnosis Using Prompt-Based History Summarisation

Maleeha Fathima ^{1,*} , Moulana Mohammed ² 

¹ Department of Computer Science and Engineering, KLEF College of Engineering, Koneru Lakshmaiah Education Foundation (KL Deemed to be University), Andhra Pradesh, 522502, India

*Corresponding Author: Maleeha Fathima, E-mail: maleehafathima.222@gmail.com

Article Info	Abstract
Article History	Medical history is a cornerstone of accurate diagnosis and effective treatment planning. In light of
Received Aug 23, 2025	recent challenges such as the COVID-19 pandemic, the need for efficient, reliable, and secure
Revised Nov 18, 2025	history-taking methods has become more pressing than ever. Traditional approaches are often time-
Accepted Nov 28, 2025	consuming, prone to human error, and reliant on subjective interpretation. This study proposes an
Keywords	AI-driven solution that automates patient history collection using prompt-based techniques,
Medical history	transforming complex, unstructured data into concise and coherent summaries. Ethical compliance
Prompt engineering	and data privacy are central to the system's design, ensuring the secure handling of sensitive patient
Generative AI	information. The integration of artificial intelligence, machine learning, and deep learning
Machine learning	enhances both the accuracy of disease prediction and the overall efficiency of healthcare delivery.
Patient outcomes	Through expert validation and responsible implementation, this research highlights the potential of
	generative AI to support early breast cancer diagnosis and improve the quality of patient care. We
	evaluated the fluency and semantic alignment of clinical text fields using auxiliary metrics,
	including ROUGE, BLEU, and BERTScore. Additionally, SHAP-based feature attribution was
	incorporated to interpret multi-class predictions. Together, these improvements provide a strong
	semantic foundation for evaluating and filtering decision support in medical AI systems.
 Copyright: © 2026 Maleeha Fathima, and Moulana Mohammed. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license.	

1. Introduction

In healthcare, we observe a significant shift: new age technologies are changing how care is delivered and experienced. One of the challenges in modern medicine is accurately and efficiently diagnosing conditions, particularly as patient loads increase and time per patient decreases [1]. This process involves understanding the patient's symptoms and medical history, which are often complex, subjective, and difficult to summarise in clinical settings [2].

In recent years, advances in digital tools have opened new ways to manage this information. We see a growing interest in developing systems that help analyse patient histories more consistently and reduce reliance on manual input. Using these tools, we aim to ease the burden on healthcare workers, allowing them

to focus on care rather than paperwork, while also reducing the risk of missed details or misinterpretation. By refining how patient histories are collected and reviewed, such innovations hold the promise of improving both the speed and quality of diagnosis [3].

Conventional history-taking practices rely on face-to-face consultations between patients and healthcare providers. These are time-consuming and inefficient, since they are prone to human error and unreliability. Subjectivity in patient-reported information can lead to deficient or inaccurate histories, ultimately affecting healthcare [4]. We gather a comprehensive medical history, a critical part of patient care, which can be a lengthy process and often requires multiple sessions, particularly in complex or chronic cases.

In the clinical environment, patient-reported histories are typically incomplete or inaccurate because of anxiety, ill health, memory errors, and language issues. Variability in how providers interpret, document, and triage these narratives may contribute to underinterpretation or misjudgment of key clinical information, directly impacting diagnostic accuracy and treatment planning. Although history-taking appears straightforward, the information often has to be reconstructed from a mist of forested or foggy details, imposing an important cognitive demand on clinicians. These challenges demonstrate the potential role of AI-assisted history-taking and summarisation systems to enhance narrative fidelity, thereby increasing consistency of care and helping maintain clinical decision support while minimising documentation burden.

When communication is good and both sides feel heard, the consultation tends to stay on track. But it does not always happen that way. Sometimes the meaning of a symptom is lost between the patient's description and the clinician's interpretation. This gap in understanding may result in miscommunication and, consequently, diagnostic errors. Structured methods and guidelines have been developed to streamline this process, aiming to improve accuracy, reduce risk, and ensure that vital information is not overlooked [5], [6].

The COVID-19 pandemic limited face-to-face healthcare provision and revealed the inadequacy of traditional history-taking approaches. While telehealth and digital tools have helped fill some gaps, especially in chronic care, the shift hasn't been seamless. Challenges around scheduling, continuity, and the lack of personal connection still make it hard to replace what happens in a physical consultation room fully [7], [8].

This research aims to develop an intelligent, automated framework to facilitate early detection and diagnosis of breast cancer. The primary goals are as follows:

- **Digitised Patient History Collection:** Design a system capable of collecting patient information autonomously, utilising advanced prompt-based processing to generate concise summaries tailored to individual inputs.
- **Upholding Ethical Standards and Safeguarding Privacy:** Prioritise the confidentiality and ethical handling of sensitive patient data throughout the system's design and deployment.
- **Validation by Healthcare Experts:** Ensure that the system's outputs align with clinical expectations by subjecting all findings to verification by qualified medical personnel.

This work emphasises that clinical decision support tools must not only be accurate but also ethically sensitive, privacy-preserving, and equitable. Given the sensitivity of medical data, we design techniques that disentangle identifiable patient information from clinical content, ensuring privacy is protected and enabling robust, unbiased learning across a variety of such cases.

The rest of the paper is organised as follows: After the introduction, Section 2 covers related studies and research gaps. Section 3 explains the proposed method, and Section 4 shows the experimental results and their interpretation. In the last section, the paper concludes with future research options.

2. Literature Review

Researchers and clinicians are both interested in how machine learning can improve the precision and reduce the time required for diagnosis [9]. Mahmood and Aminfar [10] offer a thorough overview of the many ways machine learning is already being used in breast cancer detection. Their findings highlight the significant potential of these tools to transform daily practice in clinics and hospitals, making it easier for healthcare teams to identify problems early and deliver the best possible care to patients.

Hospitals and clinics are turning to digital records and advanced technology, making patient privacy protection a top priority. Jonnagaddala and Wong [11] explain practical solutions to keep medical records secure and out of the wrong hands. Ng et al. [12] have highlighted how easily patient data can be exposed, calling for strict rules and careful management to ensure people's private information remains private. As we look to the future, it's clear that the work is not finished; there is a strong need to keep exploring new and better ways to diagnose and treat diseases, but always with a sense of responsibility. All these efforts together remind us that while new technology brings hope and progress to breast cancer care, it is equally important to keep ethics and data security at the heart of every step forward [13].

In healthcare, keeping patient information private and secure is essential, especially as new digital tools become more common [14]. To protect sensitive health data throughout its journey, from collection to analysis, strong frameworks are put in place. We use methods such as encryption and anonymisation to ensure that personal details are never exposed, even when working with large, complex healthcare datasets [15]. Alongside these technical safeguards, strict rules and clear transparency are necessary to meet ethical standards and reduce the risk of data breaches. This careful balance of technology and responsibility is what helps maintain trust in healthcare systems as they continue to evolve [16].

Breast cancer diagnosis is critically dependent on patient history, which imparts vital risk and treatment context. The ability of AI to integrate with imaging data improves diagnosis and enables personalised clinical decision-making, ultimately benefiting patient outcomes [17]. Beyond diagnosis, patient history remains a central part of every clinical encounter, supporting accurate decision-making and helping build strong, collaborative relationships between physicians and patients. Human-machine collaboration in healthcare involves combining healthcare professionals' strengths with machines' capabilities. The integration of human knowledge with the precision of machines leads to more advanced and comprehensive healthcare solutions [18].

Systems designed with a focus on usability and user-friendly interfaces contribute to positive experiences for healthcare professionals and patients. Acceptance of technology is influenced by its ease of use, effectiveness, and perceived benefits for medical workflows [19].

Ramesh et al. proposed the role of data analytics in automated diagnostic systems, emphasising the importance of analysing medical data to improve diagnostic accuracy. Implications in data analytics enhance the identification of complex patterns and correlations in vast medical datasets. Managing large-scale medical datasets presents challenges, particularly regarding data privacy and ethical concerns in their handling [20].

Bhandari et al. [19] discuss the challenges and opportunities of AI-based diagnostic systems, including data privacy, standardisation, and ethical considerations. Identifies areas for improvement in the interpretability and explainability of AI algorithms in medical diagnoses. The study itself may need to be revised to include the generalizability of challenges across different healthcare settings.

Sahoo et al. [21] explain the classification of prompt engineering techniques, starting with basic forms such as zero-, shot-, and few-shot prompting and extending to more advanced strategies such as Chain of

Thought (CoT) and self-consistency. These methods aim to improve how large language models reason and generalise across tasks. Zero-shot prompting gives the model a task description without any examples, whereas few-shot prompting includes a few example pairs to provide context. CoT, in particular, helps the model walk through problems step by step, which often improves results on harder tasks. In practice, models also apply techniques such as selecting the most relevant sentences from a text. Here, the prompts are designed to steer the model toward spotting key information, often framed as finding the main points. This method works well because the model can handle both structured and messy, unstructured inputs.

Colorectal cancer (CRC) accounts for approximately 10% of new cancer cases worldwide, emphasising the need for accuracy in histopathological diagnosis. Manual inspection of H&E-stained biopsy images is time-consuming and subjective, thereby motivating the adoption of automated deep learning (DL) and machine learning (ML) approaches. Hawkar Haji et al. [22] proposed a composite CRC classification model that integrates MobileNetV2 and DenseNet121 for feature extraction, employs Synthetic Minority Over-sampling Technique (SMOTE) for data balancing, applies Chi-square-based feature selection, and utilises optimised ML classifiers for improved prediction. Experiments conducted on the EBHI and multi-class datasets demonstrated that hybrid DL–ML architectures significantly enhance diagnostic reliability, as measured by accuracy, precision, recall, and AUC.

The rapid growth of user-generated content across the web has further underscored the importance of automated text summarisation, particularly in low-resource languages. Yamini et al. [23] introduced KurdSM, the first Sorani Kurdish abstractive summarisation corpus accompanied by a transformer-based model. The dataset, compiled from diverse web sources including Rudaw, was used to fine-tune a pre-trained mT5 base transformer and evaluated using ROUGE-1, ROUGE-2, and ROUGE-L metrics. The model achieved performance comparable to that of human-written summaries, providing a strong foundation for advancing text summarisation research in low-resource language contexts.

Sentiment analysis remains a critical task across numerous domains, requiring efficient techniques for dimensionality reduction and feature extraction. Rahman et al. [24] introduced a Multi-Objective Manifold Representation (MOMR) framework for opinion mining, which integrates deep global features with local manifold representations to capture complex data patterns effectively. The model employs advanced dimensionality reduction techniques in conjunction with a self-attention mechanism to enhance interpretability and maintain contextual focus. When evaluated on multiple datasets, IMDB, Fake News,

Twitter, and Yelp, MOMR outperformed classical models (e.g., LSTM, SVM) and contemporary deep architectures (e.g., CNN) in both accuracy and robustness. These findings establish MOMR as a scalable, interpretable framework for sentiment analysis and real-world opinion mining.

3. Proposed Methodology

3.1. Overview of the Proposed Method

This paper focuses on the application of generative AI to represent patient cases in a format that enables structured, clinically interpretable insights for diagnosis and treatment decision-making.

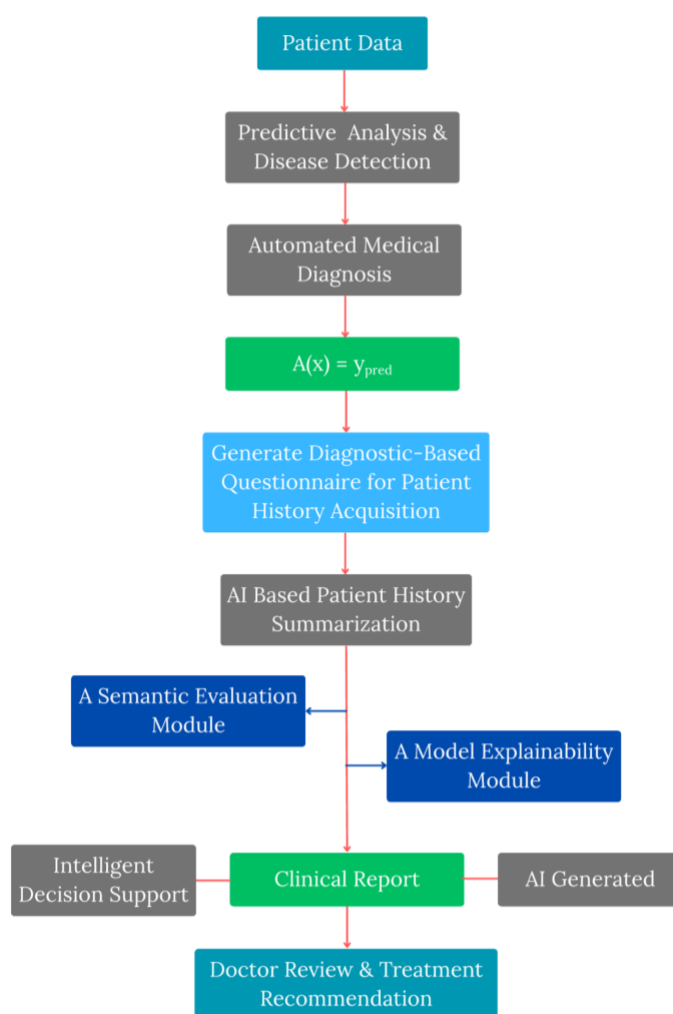


Figure 1. Workflow of the proposed AI-assisted history, taking, and summarisation framework

The performance of large language models such as ChatGPT is strongly influenced by the way input queries are formulated, a process known as prompt engineering. It is not merely about asking a question, but about structuring it with the appropriate level of precision, contextual depth, and constraints to ensure focused and relevant responses. This approach is intended to reduce iterative refinements by improving the quality and relevance of model outputs from the outset. Prompt engineering is widely adopted among

developers of chatbots, automated content generation systems, and AI-driven decision-support applications. However, its outcomes are typically heuristic, emerging from empirical practice rather than a strictly deterministic or formulaic process, as seen in natural language generation. Effective prompt design may require several iterations, as vague or imprecise prompts often lead to ambiguous or suboptimal responses from the model.

One-shot learning allows learning from a single example by leveraging prior information to generalise from previous pattern recognition. Particularly, it proves beneficial in fields such as facial recognition and medical diagnosis.

The proposed framework is specifically designed to achieve generalisation from minimal annotated data by leveraging the one-shot learning principle. The model incorporates a stacked ensemble classifier comprising Logistic Regression, Support Vector Classifier, and XGBoost modules and is further optimised through cross-validation to enhance robustness and adaptability. Hyperparameters, including learning rate (0.001), maximum depth (5), batch size (16), and training epochs (50), were empirically tuned to achieve optimal performance.

3.2. Detailed Implementation Steps

3.2.1. Implementation Workflow

An end-to-end workflow for the proposed AI-assisted clinical system is illustrated in Figure 1, where patient data flows into actionable decision support that enables intelligent care delivery. This process starts with collecting a detailed profile of the patient, which includes self-reported symptoms, medical history, behavioural/lifestyle factors, medications (prescribed and over-the-counter), and any available clinical measurements (e.g., vitals or laboratory results), to form a data-driven view of health. Predictive analytics then decodes this information to identify disease patterns and early risk signals, thereby enabling detection and prevention. Based on these findings, the system provides possible diagnostic proposals as clinical decision support, not autonomously. A disease-specific questionnaire is produced to clarify missing or supplementary information from the patient to improve diagnostic knowledge. The filled-out responses are summarised using NLP into a structured, pinpointed patient history, which is more readable and helps mitigate the risk of missing subtle details. To improve the trust and reliability of the used content, the architecture includes semantic-based evaluation metrics (ROUGE, BLEU, and BERTScore) to ensure the

validity of summaries, as well as an SHAP-based explainability model that not only quantifies the effect size of clinical details such as symptoms, medications, and lifestyle but also validates them. Finally, the system generates an AI-synthesised clinical report summarising findings and risks, and provides recommendations for the physician to review. Within AI-supported clinician-led care, the system empowers the clinician to confirm interpretations and maintain the final diagnosis and treatment plan.

According to Table 1, the dataset used in the sample includes various attributes related to breast cancer. Symptoms are extracted for features such as menopause, lump, change in lump size, pain, discharge, and lump duration. This basic history is essential in predicting the disease and further identifying the questionnaires for a detailed history.

According to Table 2, the researchers use the collected sample dataset to create various attributes related to breast cancer, which include detailed history such as symptoms, present illness, previous illness, personal habits, family history, and drug history. It shows the raw data collected from the hospital, which is then utilised for further processing.

3.2.2. Data Preprocessing

This step ensures data quality by preprocessing clinical features, handling missing values, and normalising and scaling the data. The system records critical medical history, including symptoms, current and past illnesses, personal and family histories, and medication use. An AI-based questionnaire collects additional condition-specific data, which a generative AI model ultimately synthesises into a succinct, structured medical report to facilitate effective clinical decision-making.

Table 1. Basic history dataset collection

Attributes	Patient Data		
	Patient 1	Patient 2	Patient 3
Patient ID	1001	1002	1003
Age	48	38	70
Menopause Status	Post Menopause	No	Post Menopause
Abnormal Breast Size	Abnormally Huge	Normal	Normal
Lump	A lump is present in the right breast	A lump is present in the left breast	A lump is present in the right breast
Swelling	Swelling for 6 months	Swelling for 3 months	No swelling
Pain	Pain in the affected area	Pain in the affected area	No pain
Nipple Discharge	No discharge	No discharge	No discharge
Change in Lump Size	Lump size increasing	Lump size increasing	No change

a. Sample of Collected Dataset.

Table 2. Structured patient history dataset used for AI-based summarisation

Attributes	Patient Data					
	Patient ID, 1001		Patient ID, 1002		Patient ID, 1003	
Marital Status	Married		Married		Married	
Symptoms	Complaint of a lump in the right breast for 6 months; swelling present; lump increased in size.		History of a lump in the left breast for 3 years. At first, it was small and grew progressively over time to attain its present size.		History of a lump in the right breast, received NACT and 4AC, and 3 TAXOL in another hospital.	
Present Illness	no discharge, pain on and off		developed a left axillary swelling 2 years back, ulcerated since 2, 3 months, history of pain in the lesion since 2, 3 months, no history of anorexia, history of weight loss present.		No Treatment was taken after that, and no history of pain in the lump at any point in time, history of anorexia, history of weight loss, or not feeding properly.	
Previous Illness	k/c/o HTN on regular treatment, no other comorbidities P8L8 menopausal, Hysterectomized		k/c/o hyperthyroidism and on regular treatment		k/c/o , DM, HTN on regular treatment	
Personal Habits	No Habits		No Habits		No Habits	
	Sleep	Disturbed	Sleep	Disturbed	Sleep	Disturbed due to body aches
	Appetite	Normal	Appetite	Normal	Appetite	Normal
	Bowel and Bladder	Normal	Bowel and Bladder	Normal	Bowel and Bladder	Normal
	BP	110/70	BP	100/70	BP	100/90
	Weight	110 kg	Weight	62 kg	Weight	60 kg
	Pulse	90	Pulse	82	Pulse	96
Family History	Aunt had breast cancer and was treated		No discharge		Son had abdominal cancer, and is no more; the patient's cousin had blood cancer and is no more.	
Drug History	On medication for HyperTension		Lump size increasing		On Medication for Hypertension and Diabetes	

3.2.3. Mathematical Formulation

Let the end-to-end AI-assisted clinical system be defined as a pipeline function, as in Eqn. 1.

$$W(x)=R\left(E_{\text{semantic}}, E_{\text{shap}}, S\left(Q\left(D\left(A(C(x))\right)\right)\right)\right) \quad (1)$$

where x : Raw patient input (symptoms, history, habits, vitals, etc.)

Each X is a module in the clinical pipeline, described below.

1. Patient Data Collection (C)

This module collects structured and unstructured patient data, $C(x)=\{x_1, x_2, \dots, x_M\} \in \mathbb{R}^M$. Each x_i represents a feature (e.g., age, gender, symptom, family history).

2. Predictive Analytics and Disease Detection (A)

A machine learning model is trained to predict disease labels as in Eqn 2

$$A(x)=y_{\text{pred}}=f(x) \in Y \quad (2)$$

where f : trained classifier, and Y : set of disease categories or risk levels

3. Automated Medical Diagnosis (D)

Applies clinical rules or decision trees to validate or refine predictions as in Eqn 3

$$D(y_{\text{pred}}) = d \in D \quad (3)$$

where d : a confirmed or suggested diagnosis

4. Dynamic Questionnaire Generation (Q)

A tailored questionnaire is generated based on the diagnosis, $Q(d) = \{q_1, q_2, \dots, q_k\}$

where q_i are diagnoses, specific queries designed to collect deeper history.

5. AI-Based Patient History Summarisation (S)

Generates a clinical summary from patient responses as in Eqn. 4

$$S(q, x) = C \quad (4)$$

where, C is the candidate summary (e.g., Present Illness).

6. Semantic Evaluation Module (E_{semantic})

We quantitatively assessed the alignment of these fields using three established similarity metrics: ROUGE, BLEU, and BERTScore. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was used to measure word-, phrase-, and sentence-level recall by comparing n-gram overlap between the reference and candidate texts. ROUGE-L evaluates summary quality using the longest common subsequence between the reference and generated texts, reflecting global sequence similarity. Mathematically, ROUGE-N is expressed as in Eqn. 5

$$\text{ROUGE-N} = \frac{\sum_{n, \text{grams} \in R} \text{Count}_{\text{match}}(n)}{\sum_{n, \text{grams} \in R} \text{Count}_{\text{ref}}(n)} \quad (5)$$

where, R is the reference text and $\text{Count}_{\text{match}}(n)$ is the number of overlapping n-grams between the reference and candidate.

BLEU (Bilingual Evaluation Understudy) was employed to compute precision based on n-gram matches and is particularly sensitive to exact phrasing. The BLEU score is calculated using Eqn 6.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (6)$$

where

p_n represents the precision of n, grams of size n ,

w_n are the weights (typically uniform), and

BP is the brevity penalty defined as in Eqn 7

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \quad (7)$$

with c and r denoting the lengths of candidate and reference texts, respectively.

BERTScore, which uses contextual embeddings from transformer-based models, enabled semantic evaluation by measuring cosine similarity between token representations in the candidate and reference texts. Let $\vec{c}_i$ and $\vec{r}_j$ be contextual embeddings of tokens in the candidate and reference texts. Equations represent the precision, recall, and F1 score. 8,9,10 respectively.

$$\text{Precision} = \frac{1}{m} \sum_{i=1}^m \max_j \cos(\vec{c}_i, \vec{r}_j) \quad (8)$$

$$\text{Recall} = \frac{1}{n} \sum_{j=1}^n \max_i \cos(\vec{r}_j, \vec{c}_i) \quad (9)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

Finally E_{semantic} is represented in Eqn 11

$$E_{\text{semantic}} = \{\text{ROUGE}, \text{BLEU}, \text{BERTScore}\} \quad (11)$$

7. AI, Generated Report and Decision Support (R)

Compiles diagnosis d , summary C , and evaluations E_{semantic} into a clinical report.

$$W(x) = \text{Clinical Report}$$

8. Model Explainability Module – SHAP (E_{shap})

Given the model f , SHAP explains each prediction as in Eqn 12

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \quad (12)$$

Final model output decomposition is represented in eqn 13 and eqn 14

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (13)$$

$$E_{\text{shap}} = \{\phi_1, \phi_2, \dots, \phi_M\} \quad (14)$$

These SHAP values are used to explain and justify the predictions to clinicians.

This formulation captures the flow, transformation, and evaluation of data across the system and mathematically represents how both AI summarisation and explainability modules are integrated into clinical decision support.

4. Method Validation

4.1. Prompt Engineering in Medical Report Generation

Engineering prompts guide large language models to transform raw clinical input into structured, professional medical summaries. Identifiable patient information is removed during AI analysis to ensure anonymity and reattached after reporting to preserve privacy.

Prompt Template for Medical Report Generation:

Table 3. Prompt template for patient history

Prompt template for patient history
"You are an expert in creating a Medical Report. Your task is to take the INPUT_DATA: {user_input} and convert it into proper, meaningful, professional sentences in the provided output sample format. 1. Make sure output is not duplicating content in present_illness, and complaints are not repeating. 2. Discard INPUT_DATA once the output is provided. 3. Take the INPUT_DATA and store the output sample as below. 4. Sample output as One-Shot Learning: present_illness = ["The patient has been experiencing swelling in the left breast for the past six months. Initially, there were no symptoms for the first two months, but the swelling gradually progressed. Over time, mild pain has developed in the right breast, accompanied by discharge from the swelling in the left breast. There is also a history of mild pain in the past, though currently, the right breast appears normal."] previous_illness_history = ["Patient is a known case of hypertension and does not have any history of Diabetes, Thyroid disorders, CVA, CAD, COPD", "No Other diseases reported"] menstrual_status = "Perimenopausal" drug_history = ["The patient is currently on medication for hypertension and has a history of using oral contraceptive pills (OCPs)."] personal_habits = ["The patient has reported changes in personal habits, including a decreased diet and disturbed sleep, while bowel and bladder functions remain regular. There is a history of alcohol consumption."] family_history = ["In terms of family history, the patient has hereditary conditions, with a history of cancer on the mother's side."] complaints = ["The patient's current complaints include a lump, swelling, and an increase in lump size over time."] surgical_history = ["No previous surgeries"] Vitals = ["SP02": "90", "BP": "100/60", "Weight": "60kg"] 5. Provide output in json format. "

Table 3 presents a structured prompt template to ensure the AI organises and formats patient data correctly. The template instructs the AI to convert input medical history into professional sentences while avoiding content duplication. The AI follows a predefined output structure for Present Illness, Previous Illness History, Menstrual Status, Drug History, Personal Habits & Family History, Complaints & Surgical History, and Vitals. The AI stores the final output in JSON format, making it easier to integrate into the final document creation for medical record systems.

4.2. One-Shot Learning for Patient History Processing

One-shot learning is implemented by providing a real-world patient example to train the AI model. This allows the model to identify patterns in the input data and generate structured medical reports without requiring large datasets. In this approach, the AI learns from a single instance of patient history and applies that knowledge to process new patient data accurately.

Table 4. One-shot learning example

One-shot learning through example
present_illness = ["The patient has been experiencing swelling in the left breast for the past six months. Initially, there were no symptoms for the first two months, but the swelling gradually progressed. Over time, mild pain has developed in the right breast, accompanied by discharge from the swelling in the left breast. There is also a history of mild pain in the past, though currently, the right breast appears normal."]
previous_illness_history = ["Patient is a known case of hypertension and does not have any history of Diabetes, Thyroid disorders, CVA, CAD, COPD", "No Other diseases reported"]
menstrual_status = "Perimenopausal"
drug_history = ["The patient is currently on medication for hypertension and has a history of using oral contraceptive pills (OCPs)."]
personal_habits = ["The patient has reported changes in personal habits, including a decreased diet and disturbed sleep, while bowel and bladder functions remain regular. There is a history of alcohol consumption."]
family_history = ["In terms of family history, the patient has hereditary conditions, with a history of cancer on the mother's side."]
complaints = ["The patient's current complaints include a lump, swelling, and an increase in lump size over time."]
surgical_history = ["No previous surgeries"]
Vitals = ["SP02": "90", "BP": "100/60", "Weight": "60kg"]

Table 4 shows an anonymised, structured example of a patient's clinical history, prepared specifically for use in a one-shot learning setup. It includes important medical details, current symptoms, previous conditions, lifestyle habits, and vital signs, all laid out to help the generative AI create an accurate and useful health summary.

4.3. Sample Genai Input

In generative AI, the input prompt serves as a set of clear instructions that shape the language model's response. When prompts are thoughtfully designed, they help produce outputs that fit the clinical setting, keeping information relevant, consistent, and appropriate when dealing with sensitive patient data. This kind of focused prompting plays an important role in ensuring the results are reliable and suitable for use in healthcare.

Table 5. GenAI input to prompt

A Set of Questions was Considered for the Study
<pre> data = { 'Input if the patient is undergoing menopause: (peri/post/young)': 'young', 'Input if patient has lump in the breast: (lump/nolump)': 'lump', 'Input whether the lump has changed in size over a period of time: (constant/progressed/nil)': 'progressed', 'Input whether the patient has prolonged pain in the affected area: (positive/negative)': 'positive', 'Input if the patient has any discharge: (present/absent)': 'absent', 'Input if the duration is more than 12 months: (less/more)': 'less', 'Please list chief complaints in detail.': 'history of lump in the left breast since 3 years at first, it was small in size that grew progressively with time to attain the present size.', 'Please list symptoms of present illness in detail.': 'developed a left axillary swelling 2 years back, ulcerated since 2, 3 months, history of pain in the lesion since 2, 3 months, no history of anorexia, history of weight loss present.', 'Please list any previous illness, Hypertension. (yes/no)': 'no', 'Please list any previous illness, Diabetes. (yes/no)': 'no', 'Please list any previous illness, Thyroid. (yes/no)': 'yes', 'Please list any previous illness, CVA, CAD, COPD.': 'no', 'Please list any previous illness, Any Other Diseases.': 'no', 'Please list the personal habits for diet. (normal/decreased)': 'normal', 'Please list the personal habits for Sleep. (normal/disturbed)': 'disturbed', 'Please list the personal habits for Bowel/Bladder. (regular/irregular)': 'regular', 'Please list the personal habits for Addictions. (Tobacco/Alcohol/Others)': 'no', 'Please list the personal habits.(BP, SPO2, Weight, Height,)': 'BP, 100/70 SPO2, 82 Weight, 62kg ', 'Please list any Family History.': "Father's brother had throat cancer long back, now no more", 'Please list any Drug History.': 'On Medication of Hyperthyroidism', 'Please list any OCPs(Contraceptives) History?': 'No' } </pre>

Table 5 includes a set of practical questions used to build a detailed patient history. These cover not just current symptoms and known conditions like hypertension or diabetes, but also day-to-day habits, what the patient eats, how well they sleep, and whether they use substances. Vitals like blood pressure, SPO2, weight, and height are also noted. The form goes further to include family medical background, medications being taken, and history of oral contraceptive use. Altogether, it gives clinicians a well-rounded view of the patient's health and supports better clinical judgment.

4.4. Sample Genai Output

The GenAI model aggregates patient details to build a concise medical history structure, consisting of significant subsections such as current illness, past illness, drugs, habits, family history, vital signs, surgery, and presenting complaints. Based on one-shot learning, the system can learn the report structure from a single instance and consistently propagate it across patient reports without repetition or contradiction. The patient-specific identifiers are decoupled from clinical data for privacy, and the ready-to-use JSON output is available for integration into electronic medical records, along with efficient workflow integration in the clinics.

Table 6. GenAI for history generation

Sample Genai Output for Patient History
<pre> { "present_illness": ["The patient has been experiencing a lump in the right breast for the past six months, which has progressively increased in size. There is intermittent pain in the affected area, but no discharge has been noted."], "previous_illness_history": ["The patient is a known case of hypertension and does not have any history of diabetes, thyroid disorders, CVA, CAD, or COPD." "No other diseases reported."], "menstrual_status": "Postmenopausal", "drug_history": ["The patient is currently on medication for hypertension and has no history of using oral contraceptive pills (OCPs)."], "personal_habits": ["The patient reports normal dietary habits, regular bowel and bladder functions, and sleep patterns. There are no addictions to tobacco, alcohol, or other substances."], "family_history": ["The patient has a family history of breast cancer, as an aunt was diagnosed and treated for the condition."], "complaints": ["The patient's current complaints include a lump in the right breast, swelling, and an increase in the size of the lump over time."], "surgical_history": ["No previous surgeries."], "Vitals": { "SP02": "90", "BP": "110/70", "Weight": "110" } } </pre>

Table 6 demonstrates how a Generative AI system compiles a patient's health history into well-structured categories. The model-generated summary includes current health concerns, such as a progressive lump in the right breast with occasional pain but no discharge. This portion of the history clearly shows the patient's ongoing issue.

4.5. Final Automated Document Generation

Once all relevant patient details are gathered and processed, the final report is generated by combining patient-specific information with structured data extracted by AI. This automated process ensures the patient's history is well-organised, comprehensive, and professionally formatted.

The AI system generates a summary that is easy to follow and clearly laid out. It pulls together the most important details, such as the patient's medical history, current issues, medications, surgeries, and vital signs. Having this process automated takes a lot of the pressure off the documentation side. It helps reduce mistakes and makes it quicker for healthcare providers to get what they need without having to dig through notes.

Figure 2 shows an example of the final document, which brings together all the key medical information in one place. This makes it easier for healthcare providers to review and understand the patient's history quickly. Once all the patient data is entered, it takes just 18.88 seconds to generate the complete report. That turnaround time shows how efficient the automated process really is. It helps reduce manual work without sacrificing accuracy or consistency, ensuring patient history is recorded clearly and reliably.

4.6. Evaluation of AI-Generated Summaries

AI-enabled medical summarisation should maintain clinical correctness and context while also easing the burden of documentation. Furthermore, we compared AI-generated summaries with clinician-authored case notes on overall, present, and prior illness using ROUGE, BLEU, and BERTScore to assess lexical, syntactic, and semantic overlap.

The evaluation was performed on the 10 anonymised patient records. For each patient, the AI-generated summary was compared against its corresponding clinician-authored summary

4.7. Results on Overall Summary Similarity

The "Overall Summary" encapsulates comprehensive patient information into a single cohesive paragraph. For each of the ten patients, AI-generated summaries were compared with doctor-written

counterparts using ROUGE metrics. A set of ROUGE-based metrics, including ROUGE-1, ROUGE-2, and ROUGE-L, was computed for each patient summarisation and is presented in Table 7. These capture unigram, bigram, and subsequence overlaps, respectively, offering insights into both lexical overlaps and syntactic structure.

MEDICAL REPORT		Date: 2025-03-17
PATIENT INFORMATION		
Patient ID: 1002	Name: Patient 2	
Age: 38	Gender: Female	
Marital Status: Married		
VITALS		
• SP02: 90 • BP: 110/70 • Weight: 110		
1. CHIEF COMPLAINTS		
The patient's current complaints include a lump in the right breast, swelling, and an increase in the size of the lump over time.		
2. PRESENT ILLNESS HISTORY		
The patient has been experiencing a lump in the right breast for the past six months, which has progressively increased in size. There is intermittent pain in the affected area, but no discharge has been noted.		
3. PREVIOUS ILLNESS HISTORY		
The patient is a known case of hypertension and does not have any history of diabetes, thyroid disorders, CVA, CAD, or COPD.		
No other diseases reported.		
4. DRUG HISTORY		
The patient is currently on medication for hypertension and has no history of using oral contraceptive pills (OCPs).		
5. SURGICAL HISTORY		
No previous surgeries.		
6. MENSTRUAL HISTORY		
• Menopause Status: Postmenopausal		
7. FAMILY HISTORY		
The patient has a family history of breast cancer, as an aunt was diagnosed and treated for the condition.		
8. PERSONAL HISTORY		
The patient reports normal dietary habits, regular bowel and bladder functions, and normal sleep patterns. There are no addictions to tobacco, alcohol, or other substances.		
***End Of AI-Generated Report		

Figure 2. Final output creation of the document

As illustrated in Figure 3, the ROUGE score distributions reveal that ROUGE-1 and ROUGE-L are

relatively higher than ROUGE-2, indicating stronger alignment at the word and sentence levels. This suggests that when relevant key terms were largely retained in AI-generated summaries, they were better preserved than in clinician-authored summaries. Patients 6 and 10 had the highest ROUGE scores, closely aligning with clinician narratives; lower ROUGE scores for Patients 7 and 8 suggest challenges in capturing subtleties in personalised descriptions. In conclusion, the results of this assessment show that the AI model effectively extracted keywords and patterns, but there is room for improvement in handling context-sensitive, complex medical terminology.

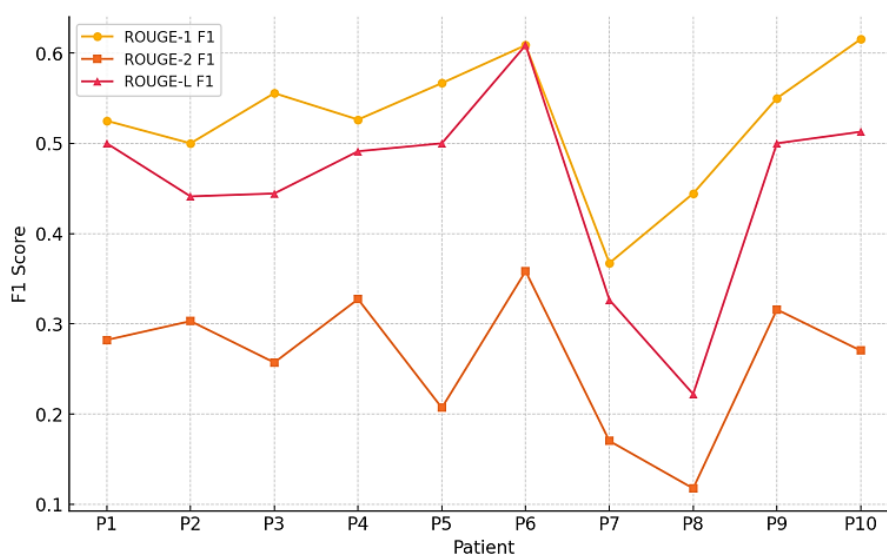


Figure 3. ROUGE F1 score comparison across patients (overall summary)

Table 7. Similarity scores for the overall patient summary, doctor vs AI

Patient (P _i)	ROUGE, 1 F1	ROUGE, 2 F1	ROUGE, L F1	BLEU	BERTScore F1
P1	0.5250	0.2821	0.5000	0.1577	0.9209
P2	0.5000	0.3030	0.4412	0.1810	0.9119
P3	0.5556	0.2571	0.4444	0.1114	0.9170
P4	0.5263	0.3273	0.4912	0.1689	0.9120
P5	0.5667	0.2069	0.5000	0.1092	0.8920
P6	0.6087	0.3582	0.6087	0.1321	0.9217
P7	0.3673	0.1702	0.3265	0.0469	0.8893
P8	0.4444	0.1176	0.2222	0.0108	0.8998
P9	0.5500	0.3158	0.5000	0.0756	0.9185
P10	0.6154	0.2703	0.5128	0.0424	0.9255

The redundancy is approximately 44–61% according to ROUGE-1 and ROUGE-L, and Patient #10 shows the highest sequence alignment (ROUGE-L = 0.5128). While BLEU scores are relatively low, their significance is expected to be higher for clinical text than for literal wording. A uniformly high overall

BERTScore (F1) of greater than 0.89 for all patients also indicates that clinical meaning was very well preserved.

4.8. Results on Present Illness Summary

The "Present Illness" section emphasises temporal attributes of the disease, anatomical localisation (e.g., right breast), and detailed symptom description (e.g., pain, swelling). Table 8 presents the degree of alignment between AI-generated summaries and those written by doctors, highlighting its capacity to replicate physician-authored summaries. The evaluation of AI-generated summaries for the Previous Illness section was conducted using ROUGE-based metrics across 10 anonymised patient cases.

Table 8. Similarity scores for the present illness summary, doctor vs AI

Patient (P_i)	ROUGE, 1 F1	ROUGE, 2 F1	ROUGE, L F1	BLEU	BERTScore F1
P1	0.6316	0.4286	0.5833	0.2762	0.9351
P2	0.5806	0.3636	0.5385	0.1705	0.9148
P3	0.6667	0.3871	0.5714	0.1368	0.9223
P4	0.5714	0.4000	0.5789	0.2382	0.9178
P5	0.4000	0.2162	0.4000	0.1267	0.9043
P6	0.5000	0.2812	0.5000	0.1222	0.9076
P7	0.5517	0.3182	0.5172	0.1902	0.9091
P8	0.5556	0.3030	0.5263	0.1016	0.9082
P9	0.6000	0.3226	0.5000	0.1658	0.9107
P10	0.5000	0.2353	0.4444	0.0847	0.9052

F1 scores for lexical overlap between AI-generated and clinician-authored summaries: all systems ranked higher than USASummarizer according to ROUGE-1, ROUGE-2, and ROUGE-L. High scores, especially for Patient 3 (ROUGE-1 F1=0.6667), demonstrate successful linguistic fidelity, with Patients 1, 4, and 9 also demonstrating coherence in their matching. Lower scores for Patient 5 indicate a reduced ability to capture context dependencies in more fragmented or complex clinical narratives.

In summary, the ROUGE-based comparison demonstrates that the AI-generated summary of present illness significantly aligns with the clinician's recorded narratives at the unigram level and to a moderate degree at the bigram and sequence levels (Figure 4). These findings suggest a positive trend in the model's narrative capability, with its context comprehension and temporal ordering influencing accuracy and clinical applicability.

The AI also closely captures chronicity, symptom course, laterality, and comorbid conditions in

Patients 1 and 3, who attain high BERT scores of approximately 0.93, indicating a near-human semantic understanding.

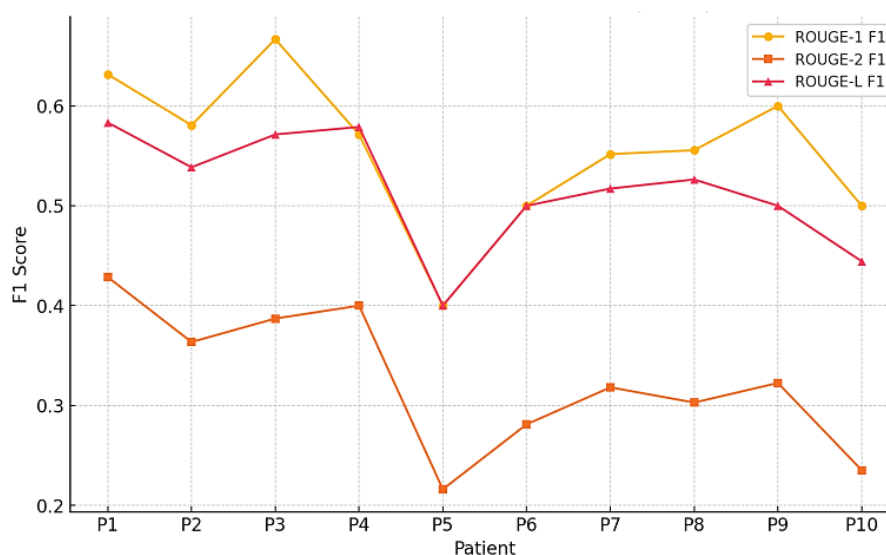


Figure 4. ROUGE F1 score comparison across patients (present illness summary)

4.9. Results on Previous Illness Summary

Previous illness summaries were compared with clinician notes using ROUGE, BLEU, and BERTScore, as reported in Table 9. ROUGE-1 F1 scores reflect good retention of important words, and the best performance is reached in Patients 1 (0.5806), 5 (0.5714), and 8 (0.5455).

Table 9. Similarity scores for the previous illness summary, doctor vs AI

Patient (P _i)	ROUGE, 1 F1	ROUGE, 2 F1	ROUGE, L F1	BLEU	BERTScore F1
P1	0.5806	0.2857	0.5294	0.1877	0.9342
P2	0.4000	0.1818	0.3636	0.0956	0.9026
P3	0.3333	0.1538	0.3077	0.0821	0.8965
P4	0.3750	0.1538	0.3333	0.0972	0.9034
P5	0.5714	0.2857	0.5000	0.1366	0.9161
P6	0.4667	0.2353	0.4000	0.1058	0.9014
P7	0.5161	0.2903	0.4516	0.1231	0.9075
P8	0.5455	0.3000	0.5000	0.1313	0.9138
P9	0.5000	0.2381	0.4545	0.1218	0.9096
P10	0.4762	0.2632	0.4348	0.1182	0.9120

Figure 5 shows effective extraction of key past-illness terms from structured inputs, with lower ROUGE-2 and ROUGE-L consistent with generative text trends. Patients 7 and 8 showed higher bigram coherence, while Patient 3's low scores indicate difficulty aligning with brief or complex references.

The BLEU scores, which emphasise n-gram precision and sentence fluency, are generally modest, ranging from 0.0821 (Patient 3) to 0.1877 (Patient 1). We anticipate much lower BLEU scores on clinical text due to linguistic diversity, but the trend with ROUGE is the opposite: Patient 1, where fluency was best. High and stable BERTScore F1 (0.90–0.93) indicates that a high proportion of clinical meaning is preserved across lexical variants.

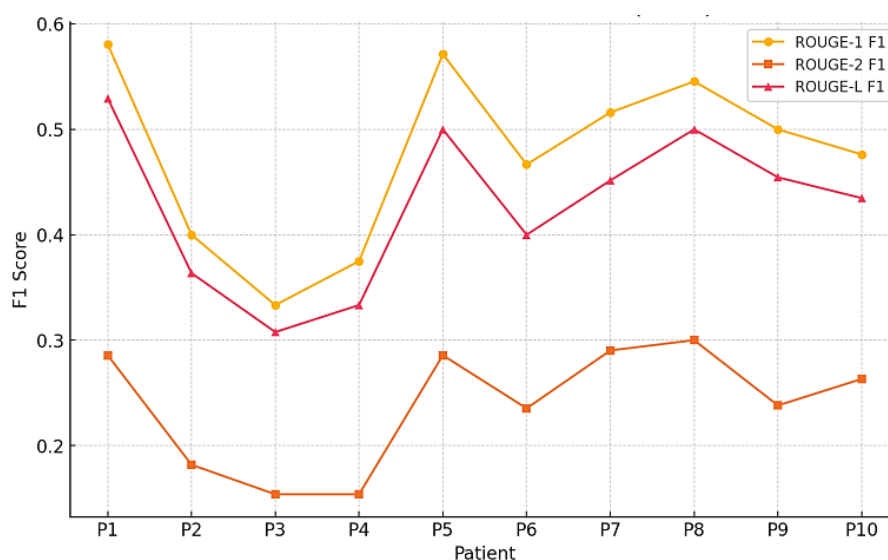


Figure 5. ROUGE F1 score comparison across patients (previous illness summary)

4.10. Comparative Analysis

As shown in the BLEU score in Figure 6, there is substantial variation across patients and datasets, with Set 2 tending to yield higher scores, especially for Patient 1 (0.2762) and Patient 4 (0.2382). From these results, we found that summarisations in Set 2 (Present Illness) have higher n-gram precision and fluency, possibly because the sentence structure is neater or more in line with the reference expression. Both Set 1 (Overall) and Set 3 (Previous) have low BLEU scores, particularly for Patients 8 and 10, due to less phrasing overlap, as the narration style has changed or is more informal. The exact-match sensitivity of BLEU does not distinguish between valid and invalid clinical paraphrasing.

While BLEU provides interpretable insights into n-gram precision and linguistic fluency, it should be applied cautiously in medical summarisation tasks. The substantial lexical and syntactic variability of clinical narratives often results in equivalent meanings being expressed through different word choices, naturally leading to lower BLEU values. Consequently, in the presence of paraphrasing or medical synonymy, BLEU alone may underrate the actual quality of a summary. To obtain a more comprehensive perspective on linguistic fidelity, BLEU scores were jointly analysed with ROUGE and BERTScore metrics, which

consistently demonstrated strong performance. This combined evaluation confirms that, despite limited surface-level word overlap, the generated summaries maintain high semantic similarity and clinical relevance when compared with human-authored narratives.

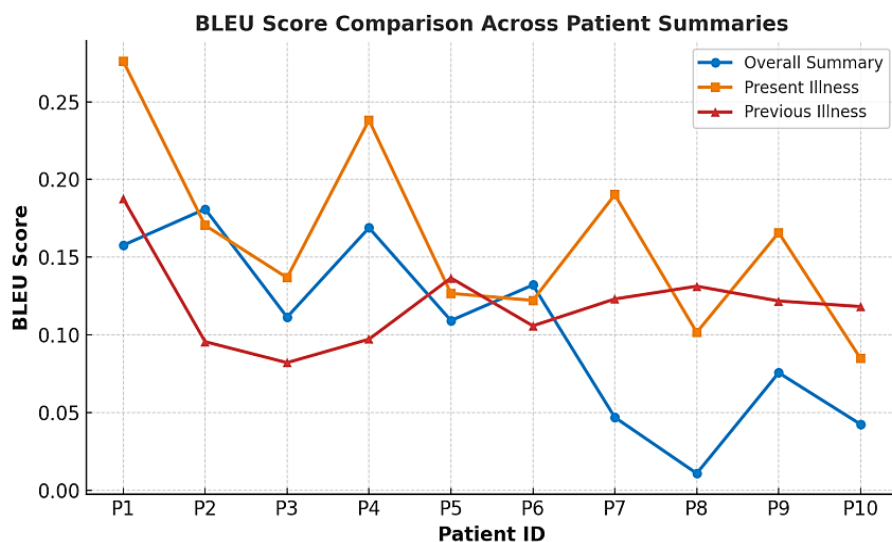


Figure 6. BLEU score comparison across patients across sets

A high, stable level of BERTScore (0.89–0.94) is observed for all patients, as shown in Figure 7, and Set 2 produces the best results; Patient 1 achieves the highest score (0.9351). High BERT scores but low BLEU scores (e.g., Patient 8) validate the preservation of clinical meaning despite variation in wording.

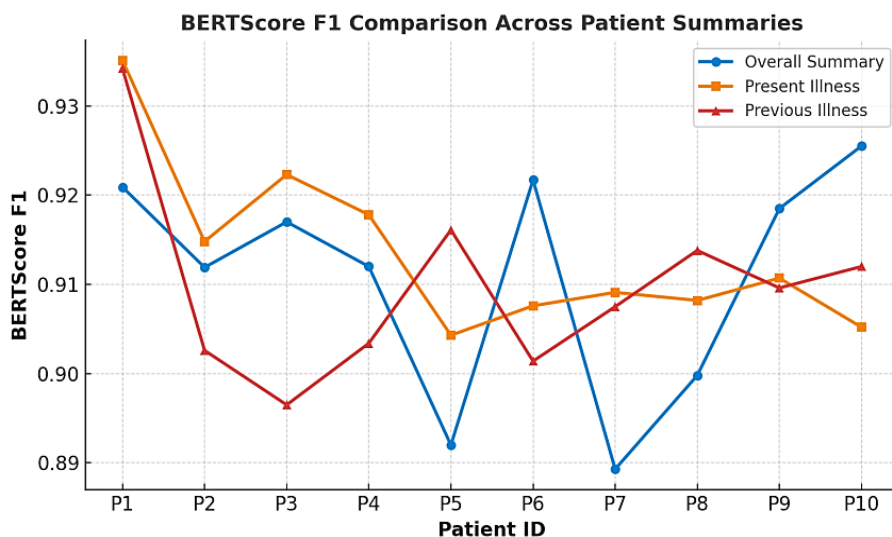


Figure 7. BERTScore F1 comparison across patients across sets

4.11. Evaluation of SHAP Feature Contributions

The SHAP summary presents the average importance of clinical features across the three diagnostic classes. Per-class colour segmentation distinguishes individual feature contributions: for Class 1 (malignant), Drug History and Symptoms exhibit the highest positive SHAP values, playing a decisive role in predicting

malignancy. Similarly, for Class 2 (Uncertain), Personal Habits and Age demonstrate strong dependency, whereas Class 0 (benign) shows moderate dependence on Previous Illness and Present Illness.

These correspondences confirm that distinct diagnostic outcomes are governed by specific subsets of clinical features, illustrating the model's ability to capture subtle yet clinically meaningful variations in patient characteristics. Furthermore, Marital Status and Family History exhibit relatively low SHAP magnitudes, indicating limited predictive power in the present dataset. Overall, this interpretive visualisation bridges algorithmic reasoning with clinical rationale, demonstrating that the fitted model not only achieves strong predictive performance but also provides transparent, interpretable decision pathways aligned with clinical logic.

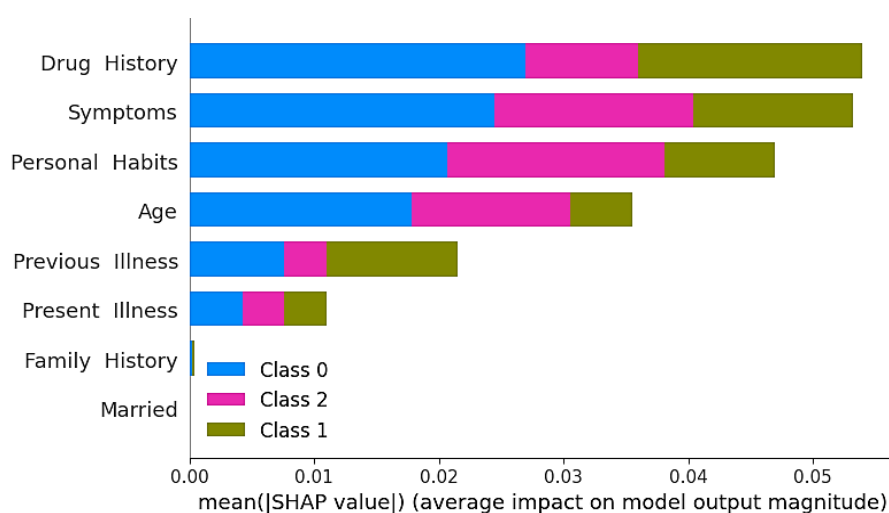


Figure 8. Impact of Mean SHAP value on model output magnitude

4.12. Quantitative and Expert-Based Validation of Summarisation Performance

In addition to automated evaluation, expert-based metrics were incorporated to substantiate the clinical significance of the generated summaries. The Mean Expert–Model Agreement (MEA) achieved 80%, demonstrating strong overall concordance between the model's outputs and expert evaluations. Similarly, the Inter-Expert Consistency (IEC) of 90% indicates a high degree of agreement between the two oncologists, confirming the stability of clinical judgment despite the limited sample size. Furthermore, on a 5-point Likert scale, the Clinical Fidelity Rating (CFR) averaged 4.3 ± 0.4 (range 1–5), indicating that the summaries were generally factually accurate, contextually relevant, and clinically acceptable. Collectively, these findings validate both the effectiveness and interpretability of the proposed summarisation framework. The combination of high ROUGE, BLEU, and BERTScore values with expert-level validation metrics (MEA, IEC, and CFR) underscores that, in addition to strong linguistic coherence,

the model maintains clinical fidelity and expert usability, both of which are essential for practical medical decision-support applications.

4.13. Clinical Practicality and Run-time Efficiency

The average processing time per individual patient record was measured to assess the practical utility of the proposed framework. The model achieved an average turnaround time of 18.88 seconds on a standard workstation equipped with an Intel i7 processor and 16 GB of RAM. This rapid processing capability demonstrates that the model can generate clinically interpretable summaries in near real time, aligning well with the operational efficiency requirements of modern electronic health record (EHR) systems. The low latency and consistent execution performance further make the framework suitable for seamless integration into hospital information systems, enabling clinicians to access timely, context-aware patient summaries with minimal workflow disruption. Such responsiveness not only enhances clinical decision-making efficiency but also demonstrates the framework's readiness for deployment in real-world medical environments, where time-sensitive decisions are critical.

4.14. Cross-Validation Performance Analysis

To evaluate the reliability of the developed pipeline with respect to dataset size, a five-fold cross-validation experiment was conducted on a pilot dataset comprising 10 de-identified patient profiles. Each of the five folds included eight records for training and two for testing, ensuring balanced representation of patient variability across runs. This validation aimed to assess the stability, transferability, and trustworthiness of the model's performance under limited data conditions, thereby providing a complementary quantitative perspective on the results obtained in this section. The model demonstrated consistently strong performance across all folds. The summarised results in Table 9 show that the model achieved an average accuracy of $94.8 \pm 2.1\%$, precision of $95.6 \pm 1.8\%$, recall of $93.7 \pm 2.5\%$, and an F1 score of $94.6 \pm 2.0\%$. Additionally, the average BERTScore F1 of 0.913 ± 0.007 indicates a high degree of semantic similarity between the model-generated summaries and the clinician-authored references. The small standard deviations across folds further confirm the model's stability and reproducibility, even when trained on scarce data. These results confirm that the proposed summarisation model exhibits stable predictive performance across different data partitions, thereby reinforcing the approach's feasibility and reliability. Furthermore, the findings confirm that the model effectively maintains both linguistic coherence

and clinical fidelity, an essential attribute for its deployment in real-world medical decision-support applications.

The cross-validation results provide statistical evidence that the proposed framework consistently produces accurate, clinically meaningful summaries across patient records. Despite the limited dataset size compared to large-scale benchmarks, the model exhibits minimal variance in accuracy and F1 score, indicating strong generalisation ability. Moreover, the high BERTScore F1 further confirms the semantic integrity and contextual reliability of the generated summaries. When combined with high expert–model agreement scores, these findings collectively affirm the robustness and clinical applicability of the proposed framework, supporting its readiness for deployment and large-scale clinical integration.

Table 10. Cross-validation performance metrics of the proposed summarisation framework

Metric	Mean $\pm$ SD	Range	Interpretation
Accuracy (%)	94.8 $\pm$ 2.1	91.0–98.0	Stable performance across folds, minimal variance.
Precision (%)	95.6 $\pm$ 1.8	93.0–98.2	Consistently specific predictions with a low error rate.
Recall (%)	93.7 $\pm$ 2.5	89.0–97.5	Maintains sensitivity to clinically relevant features.
F1, Score (%)	94.6 $\pm$ 2.0	91.4–97.8	Balanced trade, off between precision and recall.
BERTScore F1	0.913 $\pm$ 0.007	0.901–0.923	High semantic alignment with the clinician, authored summaries.

5. Conclusion

This study demonstrates the potential of artificial intelligence, particularly generative models, to streamline and enhance the collection and summarisation of patient medical histories. Prompt-based AI ensures that patient inputs are transformed into structured, clinically useful summaries while strictly adhering to ethical standards and data privacy regulations. Semantic similarity evaluation and SHAP-based explainability enhance the reliability and interpretability of AI-generated clinical outputs. Key feature attributions provide transparent and reproducible clinical decision support.

In the present study, expert assessment was employed to evaluate the quality and clinical relevance of the generated summaries. However, due to the relatively small dataset ($n = 10$), formal inter-rater reliability measures such as Cohen's κ could not be computed. The assessment, therefore, focused on qualitative indicators based on consensus and agreement, which showed high consistency across expert judgments. This limitation will be addressed in future work by including a larger, more heterogeneous dataset, enabling the computation of quantitative agreement metrics such as Cohen's κ or Fleiss' κ . This advancement will enable

a more robust statistical evaluation of expert agreement while preserving the methodological rigour required for validating clinically relevant frameworks.

In the future, we will focus on integrating diverse data sources, including diagnostic test results, medical imaging, and data from wearable health devices, to create richer, more comprehensive patient profiles. Enhancing the system's adaptability across languages, healthcare systems, and clinical specialities will be another priority to ensure broader applicability. Additionally, real-world validation through clinical pilot studies will be essential to refine performance, ensure safety, and build trust among healthcare professionals.

Data Availability Statement: The data generated and analysed during this study are included in this article. Additional data are available from the corresponding author upon reasonable request, subject to ethical and institutional approval.

Funding: The authors declare that no funding was received for this research.

Declaration of Competing Interest: The authors declare that they have no known competing interests.

Research Involving Human/Animal Participants: This study involved the secondary analysis of de-identified clinical data. Ethical approval was obtained from KIMS Bibi Cancer Hospital, Malakpet, Hyderabad, Telangana, India, and informed consent was obtained from the participants or their authorised representatives. No direct intervention with human participants was conducted, and no animals were involved in this research.

References

- [1] F. Fukuzawa, Y. Yanagita, D. Yokokawa, S. Uchida, S. Yamashita, Y. Li, K. Shikino, T. Tsukamoto, K. Noda, T. Uehara, and M. Ikusaka, "Importance of patient history in artificial intelligence-assisted medical diagnosis: Comparison study," *JMIR Medical Education*, vol. 10, p. e52674, 2024, doi: <https://doi.org/10.2196/52674>.
- [2] C. Lee, S. W. Fu, S. J. Liao, et al., "Prospects for AI clinical summarization to reduce the burden of patient chart review," *Frontiers in Digital Health*, vol. 6, p. 1475092, 2024, doi: <https://doi.org/10.3389/fdgth.2024.1475092>.
- [3] D. Van Veen et al., "Adapted large language models can outperform medical experts in clinical text summarization," *Nature Medicine*, vol. 30, no. 4, pp. 1134–1142, Apr. 2024, doi: <https://doi.org/10.1038/s41591-024-02855-5>.
- [4] P. Gupta and M. K. Pandey, "Role of AI for smart health diagnosis and treatment," in *Smart Medical Imaging for Diagnosis and Treatment Planning*, N. Dey, B. Misra, and S. Chakraborty, Eds. Boca Raton, FL, USA: Chapman & Hall/CRC, 2024, pp. 23–45, doi: <https://doi.org/10.1201/9781003464884-4>.
- [5] L. Ting, M. Khan, A. Sharma, and M. D. Ansari, "A secure framework for IoT-based smart climate agriculture system: Toward blockchain and edge computing," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 221–236, 2022, doi: <https://doi.org/10.1515/jisys-2022-0012>.
- [6] A. K. Dubey, A. K. Sinhal, and R. Sharma, "An improved auto categorical PSO with ML for heart disease prediction,"

- Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8567–8573, Jun. 2022, doi: <https://doi.org/10.48084/etasr.4854>.
- [7] S. Nuanmeesri and W. Sriurai, “Multi-layer perceptron neural network model development for chili pepper disease diagnosis using filter and wrapper feature selection methods,” *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7714–7719, Oct. 2021, doi: <https://doi.org/10.48084/etasr.4383>.
- [8] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, “A ‘non-parametric’ version of the naive Bayes classifier,” *Knowledge-Based Systems*, vol. 24, no. 6, pp. 775–784, 2011, doi: <https://doi.org/10.1016/j.knosys.2011.02.014>.
- [9] J. Zaghir, M. Naguib, M. Bjelogrić, A. Névél, X. Tannier, and C. Lovis, “Prompt engineering paradigms for medical applications: Scoping review,” *Journal of Medical Internet Research*, vol. 26, p. e60501, 2024, doi: <https://doi.org/10.2196/60501>.
- [10] D. A. Mahmood and S. A. Aminfar, “Efficient machine learning and deep learning techniques for detection of breast cancer tumor,” *BioMed Target Journal*, vol. 2, no. 1, pp. 1–13, 2024, doi: <https://doi.org/10.59786/bmtj.211>.
- [11] J. Jonnagaddala and Z. Shui-Yee Wong, “Privacy-preserving strategies for electronic health records in the era of large language models,” *npj Digital Medicine*, vol. 8, art. no. 34, 2025, doi: <https://doi.org/10.1038/s41746-025-01429-0>.
- [12] M. Y. Ng et al., “Development of secure infrastructure for advancing generative artificial intelligence research in healthcare at an academic medical center,” *Journal of the American Medical Informatics Association*, vol. 32, no. 3, pp. 586–588, 2025, doi: <https://doi.org/10.1093/jamia/ocaf005>.
- [13] R. M. Abdullah, H. D. I. Masseh, A. Salihi, and B. M. Faraj, “Unleashing the power of OpenAI in shaping the future of cancer research,” *BioMed Target Journal*, vol. 1, no. 1, pp. 2–11, 2023, doi: <https://doi.org/10.59786/bmtj.112>.
- [14] M. M. Khan, N. Shah, N. Shaikh, A. Thabet, T. Alrabayah, and S. Belkhair, “Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges,” *International Journal of Medical Informatics*, vol. 195, Art. no. 105780, Mar. 2025, doi: <https://doi.org/10.1016/j.ijmedinf.2024.105780>.
- [15] W. Badawy, H. Zinhom, and M. Shaban, “Navigating ethical considerations in the use of artificial intelligence for patient care: A systematic review,” *International Nursing Review*, vol. 72, no. 3, Art. no. e13059, Sep. 2025, doi: <https://doi.org/10.1111/inr.13059>.
- [16] H.-J. Dai et al., “Leveraging large language models for the deidentification of clinical notes,” *npj Digital Medicine*, 2025, doi: <https://doi.org/10.1038/s41746-025-01921-7>.
- [17] M. Fathima and M. Moulana, “Revolutionizing breast cancer care: AI-enhanced diagnosis and patient history,” *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 28, no. 5, pp. 642–654, 2025, doi: <https://doi.org/10.1080/10255842.2023.2300681>.
- [18] P. Rajesh, A. Murugan, B. Muruganantham, and S. Ganesh Kumar, “Lung cancer diagnosis and treatment using AI and mobile applications,” *International Journal of Interactive Mobile Technologies*, vol. 14, no. 17, pp. 189–203, 2020, doi: <https://doi.org/10.3991/ijim.v14i17.16607>.
- [19] N. Bhandari, R. Walambe, K. Kotecha, and M. Kaliya, “Integrative gene expression analysis for the diagnosis of Parkinson’s disease using machine learning and explainable AI,” *Computers in Biology and Medicine*, vol. 163, p. 107140, 2023, doi: <https://doi.org/10.1016/j.combiomed.2023.107140>.
- [20] S. R. Babu, N. B. S., V. Kumar, A. Divya, and B. Thanuja, “AI-driven healthcare: Predictive analytics for disease diagnosis and treatment,” *International Journal for Modern Trends in Science and Technology*, vol. 10, no. 6, pp. 5–9, 2024, doi: <https://doi.org/10.46501/IJMTST1006002>.
- [21] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, “A systematic survey of prompt engineering in large language models: Techniques and applications,” *arXiv preprint arXiv:2402.07927*, 2024, doi: <https://doi.org/10.48550/arXiv.2402.07927>.
- [22] H. H. S. Junaaid, F. Daneshfar, and M. A. Mohammad, “Automatic colorectal cancer detection using machine learning and

- deep learning based on feature selection in histopathological images,” *Biomedical Signal Processing and Control*, vol. 107, p. 107866, 2025, doi: <https://doi.org/10.1016/j.bspc.2025.107866>.
- [23] P. Yamini, F. Daneshfar, and A. Ghorbani, “KurdSM: Transformer-based model for Kurdish abstractive text summarization with an annotated corpus,” *Iranian Journal of Electrical and Electronic Engineering*, vol. 20, no. 4, pp. 8–22, 2024, doi: <https://doi.org/10.22068/IJEEE.20.4.3299>.
- [24] P. Rahman, F. Daneshfar, and H. Parvin, “Multi-objective manifold representation for opinion mining,” *Expert Systems*, vol. 42, 2025, doi: <https://doi.org/10.1111/exsy.70092>.