International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Zero-Shot Learning Based Multilingual Medical Symptom Understanding and Disease Inference System Using NLP and Semantic AI

Mayur Narwade Ukandji[1], Pruthviraj Shyamrao Tarode[2], Vedant Naresh Karodkar[3], Shubham Kunturwar[4]
Under the Guidance of Ms. Nitu L. Pariyal Department of Computer Science & Engineering MGM's College of Engineering, Nanded
Dr. Babasaheb Ambedkar Technological University, Lonere, Maharashtra, India

***Abstract*—The Zero-Shot Disease Prediction System represents a groundbreaking advancement in medical artificial intelligence, specifically designed to address the challenges of multilingual symptom interpretation in linguistically diverse regions like India. This research paper presents a comprehensive AI-driven solution that interprets natural-language symptom descriptions and predicts possible medical conditions without relying on disease-specific training data. Our innovative system integrates zero-shot learning principles with transformer-based multilingual embeddings, FAISS similarity search for high-performance vector retrieval, Natural Language Inference (NLI) for logical validation, and a rule-based triage mechanism for urgency classification. The system accepts symptom descriptions in English, Hindi, Marathi, or mixed-language formats, converting them into semantic embeddings that capture contextual meaning beyond superficial keyword matching. Through comparative analysis with existing medical platforms like WebMD, Ada Health, Babylon Health, and Infermedica, we demonstrate superior performance in handling informal, multilingual symptom expressions. Experimental evalu- ation confirms the system's effectiveness in providing accurate predictions within sub-second response times, with Top-3 accuracy reaching 90% across multiple languages. The triage component enhances practical utility by classifying symptom urgency into High, Medium, and Low risk categories, encouraging timely medical consultation. This work highlights the transformative potential of zero-shot learning in healthcare scenarios where labeled data is scarce, contributing significantly to the field of medical NLP through a scalable, adaptive approach to disease prediction that supports diverse user populations in understanding their symptoms and making informed health decisions.**

***Index Terms*—Zero-Shot Learning, FAISS (Facebook AI Simi- larity Search), Natural Language Inference (NLI), Multilingual Natural Language Processing, Medical Symptom Understanding, Semantic Embeddings, Transformer Models, Disease Inference, Triage Prediction, Healthcare Artificial Intelligence, Medical Diag- nostic Systems, Cross-lingual Embeddings, Sentence Transformers, Semantic Similarity Search**

I. Introduction

The rapid evolution of artificial intelligence in recent years has fundamentally reshaped how digital systems understand and interact with human language, particularly in the healthcare domain where accurate symptom interpretation is critical. Modern AI models have progressed far beyond basic keyword matching, now possessing sophisticated capabilities to grasp context, semantic meaning, emotional tone, and subtle linguistic variations. This advancement has created unprecedented opportunities in fields where human commu- nication exhibits remarkable diversity and unpredictability, with healthcare representing one of the most prominent and impactful applications. When individuals describe their health concerns, they employ highly personal expressions influenced by vocabulary richness, emotional state, native language pro- ficiency, cultural background, and regional linguistic patterns. Two individuals experiencing identical medical conditions might articulate their symptoms in completely different ways, creating substantial challenges for traditional medical prediction systems that typically depend on structured datasets with predefined symptom labels. These conventional systems often require users to select symptoms from restrictive menus or enter information in specific, rigid formats, failing miserably when confronted with free-text expressions that reflect how people naturally discuss their health concerns in everyday communication.

In numerous real-world healthcare scenarios, particularly in developing regions like India with rich linguistic diversity, constructing comprehensive disease-specific training datasets proves fundamentally impractical. Rare medical conditions, newly emerging infectious diseases, and region-specific ill- nesses frequently lack sufficient annotated data for supervised machine learning approaches.

Furthermore, the inherent uncer- tainty and variability in symptom presentation across different patients adds substantial complexity, as no two individuals experience or describe identical illnesses in precisely the same manner. This context makes zero-shot learning particularly valuable as a powerful alternative paradigm, enabling mod- els to infer disease predictions directly from the semantic meaning of symptom descriptions rather than relying solely on labeled examples from specific disease categories. A zero-shot disease prediction system analyzes the semantic content of user-generated symptom narratives and compares these against comprehensive disease descriptions, allowing identification of relevant medical conditions even without explicit training on those specific diseases. This approach proves especially transformative for healthcare applications in resource-constrained environments, where early guidance can help individuals recognize when symptoms should not be ignored and professional consultation becomes necessary.

The incorporation of digital technology into healthcare delivery has been an ongoing process for several decades, yet early computer-based diagnostic systems remained largely constrained by rigid structural frameworks and limited linguistic understanding capabilities. These primitive systems required users to enter symptoms in standardized formats or select from predefined lists, completely ignoring the natural variability inherent in human health communication. Such systems could not interpret free-text inputs like ”I get breathless when walking upstairs” or regionally expressed sentences such as ” ” (I feel pressure in my chest) because they lacked the semantic analysis capacity to extract meaning beyond fixed rules and pattern matching. This created a persistent and growing gap between how people naturally communicate health concerns in daily life and how computational systems interpret this medically critical information.

As digital communication expanded globally through in- ternet penetration, individuals increasingly turned to online platforms to describe and seek preliminary explanations for their symptoms. Their language usage remained free, unstruc- tured, and highly varied, incorporating colloquial phrases, code-switching between languages, and informal expressions reflecting genuine health concerns. This behavioral shift highlighted the urgent need for healthcare technologies capable of understanding natural human communication rather than enforcing rigid, template-driven interactions. The emergence of deep learning architectures and advanced natural language processing techniques marked the definitive turning point in this evolution.

Transformer-based models, multilingual embeddings, and context-aware semantic representations enabled machines to interpret medical sentences at a deep semantic level rather than superficial structural analysis. Zero-shot learning emerged naturally from these technological advancements, providing a robust methodological framework for predicting unseen disease categories based solely on descriptive knowledge and semantic reasoning.

This methodological approach aligns perfectly with health- care realities in multilingual societies, where collecting large annotated datasets for every possible disease remains unre- alistic. It directly addresses the linguistic diversity prevalent in countries like India, where individuals frequently switch between languages within single sentences or express medical concerns exclusively in their native tongues. With advancements in multilingual transformer models, computational systems can now understand such natural expressions without requiring language-specific training data. The background of this study is therefore firmly grounded at the intersection of computational linguistics, healthcare technology, accessibility engineering, and the growing societal need for intelligent systems capable of understanding people exactly as they naturally speak about health concerns.

The central research problem addressed in this project concerns the fundamental inability of traditional medical pre- diction systems to understand free-text symptom descriptions across different languages and expressive styles. Most existing digital health tools still depend critically on structured inputs, standardized symptom lists, or supervised training datasets that represent only a limited spectrum of diseases. Such systems fail completely when encountering symptoms described in informal language or mixed linguistic formats, which represent common patterns in everyday health communication. This limitation becomes especially problematic when users describe vague or overlapping symptoms, or when diseases present themselves differently across individuals due to biological variability and subjective experience.

Furthermore, the lack of accessible and timely medical guid- ance exacerbates this technological limitation. Many individuals avoid visiting healthcare professionals due to geographical distance, financial constraints, psychological fear, or systematic underestimation of symptom severity. Without reliable prelimi- nary guidance systems, people may delay seeking professional help until medical conditions worsen substantially. Additionally, it remains practically impossible to create exhaustive training datasets for every known disease, particularly for rare genetic conditions or rapidly evolving infectious diseases.

The linguistic diversity inherent in multilingual societies further complicates this situation, as models trained exclusively on English medical text often fail completely when interpreting symptoms phrased in regional languages.

Thus, the core research problem addressed by this project involves designing and implementing a robust computational system capable of understanding natural symptom descrip- tions regardless of phrasing complexity, language choice, or completeness level, while simultaneously predicting possible diseases without relying on disease-specific training data. This ambitious goal requires developing a more flexible, adaptive, and semantically driven approach to medical prediction that transcends traditional machine learning paradigms.

The primary objective of this research project involves developing a comprehensive zero-shot disease prediction system capable of analyzing natural-language symptom descriptions and identifying likely medical conditions through semantic understanding rather than supervised training. The system aims to interpret user inputs by converting them into meaningful numerical representations using advanced transformer models. These semantic embeddings capture rich contextual information and represent the underlying medical meaning of symptom descriptions rather than surface-level lexical patterns.

Once transformed into embedding vectors, the system com- pares user inputs against a comprehensive database of disease descriptions using FAISS (Facebook AI Similarity Search), a highly optimized similarity search engine designed for high- dimensional vector spaces. This computational approach en- ables the system to retrieve diseases that semantically resemble the meaning of symptom inputs, even without any prior training on those specific conditions. To ensure logical coherence and medical relevance, a Natural Language Inference (NLI) model evaluates whether user-described symptoms logically align with each candidate disease description. This validation step makes predictions medically meaningful rather than merely statistically similar in vector space.

The project also aims to support multiple Indian languages natively, enabling users to describe symptoms comfortably in English, Hindi, Marathi, or other supported languages without translation barriers. It focuses intentionally on creating systems that do not presume medical knowledge from users and guides them through natural language interactions. Beyond disease identification, the system provides actionable suggestions regarding appropriate diagnostic tests, relevant medical special- ists, and preliminary severity assessments through integrated triage logic.

The overarching objective remains delivering an accessible, intelligent health-support tool that can guide diverse users toward timely medical consultation while respecting linguistic and cultural contexts.

The significance of this research study lies fundamentally in its capacity to address real-world challenges faced by individuals seeking preliminary medical guidance, especially in environments where healthcare support remains limited or difficult to access consistently. From technological and methodological perspectives, the system demonstrates prac- tically how zero-shot learning principles and advanced natural language processing techniques can be effectively combined to overcome inherent limitations of traditional machine learning models in healthcare applications. By removing dependence on disease-specific training datasets, the proposed model becomes inherently more flexible, scalable, and sustainable long-term. Its semantic interpretation capabilities allow natural adaptation to new diseases, updated medical descriptions, and diverse linguistic patterns, making it future-ready within constantly evolving healthcare landscapes.

From societal and public health perspectives, the system promotes accessible and early-stage healthcare awareness across diverse populations. Many individuals delay consulting medical professionals because of financial constraints, psychological fear, social stigma, or geographical barriers. A robust zero-shot disease prediction system provides preliminary understanding of symptoms based solely on how users naturally describe their conditions, encouraging timely and informed healthcare decisions. This technological intervention can potentially reduce complications caused by delayed diagnosis and empower indi- viduals to take proactive control of their health management. The multilingual capability of the proposed system adds substantial value to its practical relevance in linguistically diverse regions. By allowing users to express symptoms comfortably in their native or mixed languages, the model becomes more approachable and culturally sensitive. This in- clusivity not only strengthens user trust but also bridges critical communication gaps for populations that may struggle with English proficiency or formal medical terminology. It ensures that healthcare technology remains genuinely accessible to people across different socio-linguistic backgrounds, reducing health disparities.

Furthermore, this research contributes academically by demonstrating structured and practical integration of zero-shot learning principles, semantic embedding techniques, similarity search mechanisms, and logical inference models within a unified healthcare application.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

It showcases how these emerging technologies can be harmonized effectively to build real-world healthcare solutions that are both technically effective and user-friendly. Beyond contributing to academic knowledge in medical AI, the project opens viable pathways for future research in digital healthcare delivery, multilingual artificial intelligence, and human-centered medical support systems. It reflects practically how intelligent computational systems can complement clinical practice by providing accurate, fast, and meaningful preliminary health insights to users worldwide, particularly in underserved linguistic communities.

II. Literature Review

The evolution of intelligent medical prediction systems represents the culmination of decades of interdisciplinary research spanning artificial intelligence, clinical informatics, computational linguistics, and human-computer interaction. Early diagnostic technologies were designed primarily to assist healthcare professionals by providing computational support for clinical decision-making processes. However, these pioneering systems remained fundamentally limited by rigid structural frameworks and narrow functional capabilities. As computational technology advanced progressively, researchers and developers began exploring more dynamic approaches capable of adapting to the inherent complexities of real-world medical data and patient communication patterns. Modern medical AI systems incorporate deep learning architectures, semantic embedding spaces, multilingual natural language understanding, and zero-shot learning paradigms to interpret unstructured symptom descriptions expressed in diverse lan- guages and personal styles. This literature review chapter presents a comprehensive examination of theoretical founda- tions, technological developments, and research advancements that collectively paved the methodological way for the zero-shot disease prediction system developed in this research project.

A. Review of Existing Medical Prediction Platforms

Several existing medical diagnostic systems and AI-based symptom checkers have been developed and deployed in recent years, attempting to assist users by interpreting symptoms and providing possible medical conditions. While differing substantially in functionality and technical complexity, these platforms provide important foundational understanding of how automated diagnostic tools operate practically and where significant improvements remain necessary.

Notable existing systems and similar research projects are analyzed critically below.

1) WebMD Symptom Checker: WebMD represents one of the most widely used online symptom-checking platforms globally. Users manually select symptoms from predefined hierarchical lists, and the system generates possible conditions using rule- based algorithms and statistical medical data correlations. However, WebMD cannot interpret free-text natural language inputs, and its prediction capabilities remain strictly restricted to conditions stored within its symptom-disease mapping database. The platform also lacks multilingual support completely and cannot perform logical reasoning using advanced AI models, relying instead on predetermined probability tables.

2) Ada Health – AI Medical Assessment Application: Ada Health employs a machine learning-based approach to analyze user symptoms and generate personalized health assessments. It provides a conversational chatbot-style interface where users answer structured questions sequentially. While Ada utilizes advanced AI models for probabilistic inference, it does not support open-ended natural language descriptions, nor does it implement zero-shot learning methodologies. Its predictions remain fundamentally tied to curated training datasets, limiting adaptability to new or rare medical conditions not represented during training.

3) Babylon Health – AI Consultation System: Babylon Health offers an AI-driven medical assistant that evaluates symptoms through structured conversational interactions. Al- though technologically more advanced than traditional rule- based systems, its disease prediction accuracy depends heavily on large supervised training datasets, which intrinsically limits adaptability to new or rare medical conditions. The system follows predefined question-answer flows, reducing flexibility when users describe symptoms freely in natural language formats. It does not utilize FAISS vector search or NLI validation mechanisms, meaning it cannot perform semantic similarity reasoning or logical consistency checks between symptoms and predicted diseases. Consequently, adding new medical conditions requires complete model retraining, which severely impacts practical scalability.

4) Infermedica – Symptom Triage Engine: Infermedica provides a specialized medical inference engine that performs symptom assessment and triage prioritization. It uses proba- bilistic reasoning and Bayesian networks rather than zero-shot learning or semantic embedding techniques.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

The system cannot interpret multilingual free-text input, making it less flexible compared to modern NLP-based approaches. Its dependence on structured data intake limits applicability in linguistically diverse environments where free expression dominates.

B. Evolution of Medical Diagnostic Systems

The earliest generation of computerized diagnostic tools emerged during the 1970s and 1980s, when expert systems were developed to emulate human clinical decision-making processes. These pioneering systems, including landmark projects like MYCIN and INTERNIST-I, relied entirely on manually encoded rules describing medical knowledge in rigid ”if-then” logical forms. These systems represented significant technological achievements for their historical period, yet they suffered from fundamental limitations that restricted real-world utility. Their complete reliance on expert-created rules made them slow to update, difficult to scale, and fundamentally unable to adapt to new diseases or nuanced expressions of symptoms.

As medical knowledge expanded exponentially and user expectations evolved with digital technology, the limitations of rule-based diagnostic systems became increasingly apparent and problematic. These systems proved incapable of handling linguistic ambiguity, lexical variation, or free-text patient input. They failed consistently when confronted with natural expressions such as ”my chest feels heavy at night” or ” ” because they lacked semantic understanding capabilities for natural human language. The systematic inability to interpret subjective patient experiences or informal symptom descriptions highlighted the urgent need for more flexible, data-driven systems capable of learning from real-world linguistic patterns rather than static logical rules.

C. Transition to Traditional Machine Learning Approaches

The methodological introduction of traditional machine learning algorithms marked a major paradigm shift in medical prediction research. Models including logistic regression, decision trees, support vector machines, and ensemble methods brought statistical pattern recognition capabilities into health- care applications. These algorithms could analyze structured datasets composed of symptom indicators, demographic vari- ables, and diagnostic outcomes, learning statistical associations automatically from training examples.

However, despite demonstrated successes in structured-data environments, traditional machine learning approaches still faced foundational challenges in medical applications.

The most significant limitation involved their dependency on large labeled datasets, which are often scarce in healthcare due to patient privacy concerns, variability in clinical documentation practices, and the inherent rarity of certain disease conditions. These models also struggled fundamentally with natural language content because they relied on numerical feature representations rather than textual understanding. Techniques such as bag-of- words representations or TF-IDF vectors were employed to convert text into numerical forms, yet these early NLP methods failed to capture deeper semantic meaning, linguistic context, and relational patterns between medical concepts. They treated phrases like ”chest pain” and ”pain in chest” as unrelated lexical patterns, even though medically they represent identical symptomatic presentations.

This systematic inability to understand linguistic structure and semantic nuances restricted the practical effectiveness of traditional machine learning in real-world diagnostic scenarios where language variability dominates. Thus, the research community intensified searches for more sophisticated lan- guage understanding techniques capable of handling medical communication complexity.

D. Rise of Deep Learning and Advanced Linguistic Modeling

Deep learning architectures introduced a revolutionary shift in computational ability to process unstructured textual infor- mation in medical contexts. Recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) models improved interpretation of sequential medical text by incorporating memory mechanisms that captured contextual information across multiple words in descriptions. This technological advancement allowed healthcare models to process longer symptom descriptions and detect subtle patterns indicative of specific disease categories.

Even with these improvements, RNN architectures strug-gled with long-range linguistic dependencies and multilingual complexity in medical communication. The transformative introduction of the transformer architecture fundamentally changed natural language processing capabilities across do- mains. Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers), GPT variants, RoBERTa, and XLM-RoBERTa enabled machines to process medical text bidirectionally and understand relational patterns between every word pair in symptom sentences. This led to dramatically deeper comprehension of natural-language symptom descriptions and their clinical implications.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Transformers also introduced the foundational concept of semantic embeddings—dense numerical representations capturing meaning, context, and linguistic patterns in continuous vector spaces. Unlike earlier vectorization techniques, semantic embeddings allow models to understand that medical phrases like ”shortness of breath,” ”difficulty breathing,” and ”breathlessness” refer to closely related clinical concepts despite lexical differences. This new level of linguistic intelligence laid the essential foundation for more advanced diagnostic systems, particularly those based on semantic understanding rather than lexical matching.

E. Zero-Shot Learning and Its Critical Importance in Medical AI

Zero-shot learning (ZSL) emerged as an innovative solution to one of the most persistent challenges in healthcare AI: the fundamental scarcity of labeled training data for numerous diseases, particularly rare conditions and novel pathogens. ZSL allows a model to recognize and infer categories it has never encountered during training by leveraging semantic descriptions of those categories. Instead of learning from direct examples, the system learns from textual definitions or conceptual descriptions, enabling generalization to unseen conditions.

In the medical domain, this approach proves especially valuable because diseases often lack substantial annotated datasets due to rarity, novelty, or privacy constraints. Moreover, new diseases can emerge unexpectedly at any time, as evidenced dramatically during the COVID-19 pandemic, when early detection tools were urgently needed before any substantial labeled datasets existed. Zero-shot learning bypasses the need for disease-specific training and instead relies on the model’s linguistic understanding of disease descriptions from medical literature, creating adaptable diagnostic capabilities.

ZSL aligns perfectly with the inherent nature of symptom interpretation, where human expression remains diverse, un- predictable, and culturally varied. By relying on semantic reasoning rather than memorized patterns, zero-shot models can interpret free-text symptoms regardless of phrasing complexity or language choice, making them highly inclusive and adaptive to diverse populations.

F. Semantic Similarity Search and the Computational Role of FAISS

Semantic embeddings enable meaningful numerical represen- tation of symptoms and diseases, but these high-dimensional representations must be compared efficiently to identify clin- ically relevant matches.

This computational requirement led to widespread adoption of FAISS (Facebook AI Similarity Search), a high-performance library specifically designed to search through millions of dense vectors within milliseconds, enabling real-time applications.

FAISS plays a critical role in enabling practical real-time disease prediction systems. When users enter symptom descriptions, the system converts them into embedding vectors and compares them efficiently with pre-computed disease embeddings using FAISS optimized algorithms. The library utilizes advanced indexing structures, clustering techniques, and GPU acceleration capabilities to identify nearest semantic neighbors quickly and accurately. Without FAISS, performing such high-dimensional similarity searches manually would be computationally prohibitive and too slow for practical healthcare applications.

In the specific context of zero-shot disease prediction, FAISS serves as the computational engine that retrieves the most semantically relevant disease candidates from large knowledge bases, forming the essential foundation for subsequent clinical reasoning and validation steps.

G. Logical Validation through Natural Language Inference

Semantic similarity-based retrieval alone cannot guarantee medical correctness or clinical relevance. Two text segments may appear similar at surface linguistic levels but differ significantly in medical meaning. For example, ”I do not have chest pain” and ”I have chest pain” share substantial vocabulary but communicate diametrically opposite clinical states. To resolve such critical ambiguities, Natural Language Inference (NLI) models provide essential logical validation.

NLI methodology evaluates logical relationships between two text segments—typically classifying the relationship as entailment (logical support), contradiction (logical opposition), or neutrality (logical independence). In medical prediction sys- tems, NLI ensures systematically that user-described symptoms logically support potential disease hypotheses. This reasoning layer adds a crucial level of verification that prevents incorrect or misleading predictions. It acts as a safety mechanism, ensuring that final diagnostic outputs align with medical semantics rather than superficial text similarity.

By integrating NLI validation, modern prediction systems achieve an optimal balance between statistical relevance and logical consistency, creating more trustworthy and clinically meaningful results for users.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

H. Growth of Multilingual NLP for Healthcare Applications

Language diversity poses a major challenge in digital healthcare delivery across multilingual societies. In linguis- tically diverse countries like India, users commonly describe symptoms in local languages or mixed-language formats combining regional and English terms. Traditional NLP systems designed primarily for English fail completely to interpret these natural inputs accurately, creating substantial accessibility barriers.

Modern multilingual transformer models, such as mBERT, XLM-RoBERTa, and multilingual Sentence-BERT variants, solve this problem effectively by generating unified embedding spaces spanning multiple languages. These models understand inherently that medical concepts like ”fever,” ”,” and ”” represent identical clinical phenomena across languages. This semantic capability allows healthcare systems to become gen- uinely inclusive and accessible to broader linguistic populations. Multilingual NLP research emphasizes the critical impor- tance of bridging language gaps in healthcare technology. By supporting diverse linguistic expressions, these advanced systems build user trust and allow natural communication without forcing adaptation to technical or linguistic formats unfamiliar to users.

I. Consolidated Insights from Literature Analysis

The comprehensive literature reviewed in this chapter demon- strates a clear trajectory of continuous innovation in medical prediction systems. Early rule-based models provided structural frameworks but lacked linguistic adaptability. Traditional machine learning improved statistical pattern recognition but struggled fundamentally with free-text input and required large labeled datasets. Deep learning introduced powerful models capable of contextual understanding, while transformer architec- tures revolutionized natural language processing by capturing semantic meaning at unprecedented levels of sophistication.

Zero-shot learning emerged as a groundbreaking paradigm capable of predicting unseen diseases through semantic reason- ing rather than labeled data dependence. FAISS added essential computational efficiency to large-scale embedding comparisons, and NLI contributed logical validation to ensure medically meaningful predictions. Multilingual NLP extended practical accessibility by enabling systems to understand symptoms expressed across different languages and dialects.

Collectively, these technological advancements created a strong theoretical and methodological foundation for develop- ing intelligent, multilingual, and semantically aware disease prediction systems. The zero-shot disease predictor built in this research project integrates these ideas into a unified archi- tectural framework capable of handling real-world symptom descriptions with significant accuracy and inclusivity.

Additionally, the reviewed literature highlights the growing importance of real-time performance, model interpretability, and data privacy in healthcare AI applications. It also empha- sizes the critical need for systems that can operate effectively in low-resource settings where labeled medical data remains scarce. The convergence of semantic intelligence with scalable computation is proving to be a key direction for next-generation clinical decision support tools. Furthermore, the integration of multilingual processing ensures that such systems can bridge healthcare access gaps across diverse linguistic populations effectively.

III. Methodology And System Architecture

The system design phase forms a crucial component of this research project as it defines how the proposed Zero-Shot Learning based Medical Assistant System will operate in real-world healthcare environments. This phase focuses on translating conceptual research ideas into well-structured technical frameworks that clearly explain interactions between different system components. Carefully planned architectural design ensures that the system functions efficiently, maintains security protocols, and delivers accurate medical predictions to diverse users.

The system design emphasizes development of an intelli- gent, fast, and scalable medical assistance platform capable of interpreting natural language symptom descriptions and generating meaningful disease predictions without depending on disease-specific training data. To achieve this objective, the system follows a layered and modular architectural approach where individual modules including input processing, semantic embedding generation, similarity search, logical validation, triage classification, and result presentation work in coordinated synchronization.

Both structural and behavioral aspects of the system are explained using various UML and architectural diagrams in- cluding Component Diagrams, Deployment Diagrams, Activity Diagrams, Data Flow Diagrams, Use Case Diagrams, Sequence Diagrams, State Diagrams, Security Architecture Diagrams, and Entity Relationship Diagrams.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

These visual representations help in understanding system workflows, data movement patterns, user interaction sequences, internal state transitions, and applied security mechanisms.

Overall, this section provides a detailed architectural blueprint by describing how different components are organized, how information flows across various processing stages, and how the system maintains accuracy, performance, and security. This design foundation plays an important role in guiding im- plementation and evaluation phases. Additionally, well-defined system design helps minimize development errors, improve long-term maintainability, and ensure future enhancements can be integrated smoothly without major structural changes. It also serves as a technical reference for developers and researchers who wish to understand or extend the system in future work.

A. Design Objectives

The primary design objective involves developing a reliable, intelligent, and user-friendly Zero-Shot Learning based Medical Assistant that can understand natural language symptom descriptions and generate accurate disease predictions without relying on disease-specific training datasets. The design aims to bridge the critical gap between complex medical information and ordinary users by allowing them to describe health issues in simple, everyday language while still receiving structured and medically meaningful predictions.

Another major objective involves combining semantic under- standing with logical validation so the system does not depend solely on keyword matching or statistical correlations. The design ensures systematically that the system first identifies semantically related diseases and then checks whether those diseases are logically supported by described symptoms. This layered reasoning approach improves prediction accuracy and clinical trustworthiness substantially.

The system is also designed specifically to support fast response times so users receive near real-time feedback, which is critically important in healthcare contexts where timely decisions matter. Security and privacy form core components of design objectives because the system handles sensitive symptom information. Therefore, the design includes secure API communication protocols, access control mechanisms, and data protection frameworks.

Additionally, the design aims for comprehensive multilingual support so the system can handle inputs in different Indian languages, and for architectural scalability so new diseases, models, and features can be integrated in future without disrupting existing workflows.

B. Overall System Architecture

The overall system architecture of the Zero-Shot Medical Assistant follows a layered and modular structure that clearly separates user interaction, application logic, machine learning intelligence, and data management layers so each component can be improved or replaced independently. The logical view of system components and their interactions is depicted through detailed component diagrams showing software architecture, while physical deployment of these components across client devices, servers, and external services is illustrated through infrastructure deployment diagrams.

The frontend layer provides user interface components through which users enter symptoms, select preferred languages, and view prediction results intuitively. The API layer, built using FastAPI framework, receives requests from frontend interfaces, validates inputs rigorously, applies security checks, and forwards valid requests to the processing pipeline. The logic layer contains the core processing pipeline that controls input normalization, embedding generation, similarity search, NLI validation, ranking algorithms, and triage classification. The machine learning layer includes the SentenceTransformer model, FAISS similarity search engine, XLM-RoBERTa NLI model, and triage classifier, which together provide the artificial intelligence capabilities of the system. The data layer stores disease descriptions, FAISS index files, user interaction records, system logs, and configuration data persistently.

The system deployment architecture allows handling multiple concurrent requests efficiently, balancing computational load across available servers, and maintaining high availability for users. This architectural approach ensures the system remains scalable, maintainable, and suitable for real-time medical assistance applications across diverse settings.

C. Input Processing and Normalization

User inputs are generally unstructured, informal, and some- times written in mixed languages with regional variations.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

They may include spelling mistakes, medical abbreviations, emoticons, or incomplete phrases. Such raw text cannot be directly processed by embedding models and similarity search algorithms. For this reason, the system includes a dedicated input processing and normalization module that prepares text systematically for further analysis. The overall flow of this phase is depicted through activity diagrams showing symptom analysis workflows comprehensively.

During preprocessing, unnecessary symbols, extra whites- pace characters, and irrelevant punctuation are removed from symptom descriptions. The text is converted into consistent formats, such as lowercase representations, to reduce variations caused by different writing styles. Simple spelling errors are corrected algorithmically wherever possible, and repeated words or noisy text segments are normalized systematically. When users provide symptoms in regional languages such as Hindi or Marathi, or in language mixtures, multilingual normalization ensures original medical meaning is preserved while still making text suitable for model processing.

By performing these preprocessing steps methodically, the normalization module improves text quality entering semantic embedding stages. Clean and standardized input helps remain- ing pipeline components produce more accurate and stable results. Without proper normalization, the system would remain highly sensitive to minor variations in user input patterns, reducing reliability substantially.

D. Embedding Generation Using Sentence Transformers

Once symptom descriptions have been cleaned and normal- ized textually, they are passed to the embedding generation module. At this stage, the system uses a Sentence Transformer- based model to convert input text into high-dimensional numerical vectors known as semantic embeddings. These vectors capture overall sentence meaning rather than just counting individual words statistically.

The fundamental goal of generating semantic embeddings involves mapping medically similar sentences close together in continuous vector space. For example, clinical phrases such as ”tightness in chest,” ”pressure in chest while breathing,” and ”feeling heaviness in chest” may be written differently lexically, but medically they point toward related cardiac or respiratory conditions. The embedding model captures this semantic similarity effectively and represents these sentences in ways that enable meaningful comparison.

This semantic representation capability proves particularly important for Zero-Shot Learning because the system is not trained specifically on each disease category. Instead, it learns general linguistic understanding and applies that knowledge to match user inputs with disease descriptions semantically. The complete transformation from natural text to semantic vectors lays the essential foundation for similarity search performed in subsequent stages.

E. FAISS-Based Similarity Search

After symptom text is converted into embedding vectors, the system needs to identify which diseases are most semantically relevant. For this purpose, a FAISS-based similarity search module is implemented. FAISS represents a high-performance library specifically designed for searching similar vectors in large collections, making it ideal for real-time medical applications.

User-generated embeddings are compared systematically against large sets of precomputed disease embeddings stored in optimized FAISS indices. The embedding vector is sent to the FAISS engine, which efficiently retrieves the top-k closest disease vectors based on distance metrics like cosine similarity or inner product. These retrieved disease candidates represent conditions that are semantically closest to user symptom descriptions in embedding space.

This retrieval stage is optimized extensively for speed so even as disease databases grow larger, the system can return results quickly. However, at this initial retrieval point, diseases are selected mainly based on semantic similarity and are not yet checked for logical clinical consistency. Therefore, they are forwarded to subsequent validation steps for further refinement and filtering.

F. Natural Language Inference Validation

Semantic similarity alone proves insufficient to ensure predicted diseases truly match user symptoms medically. To add essential layers of logical reasoning, the system uses a Natural Language Inference validation module built using the XLM-RoBERTa model. In this stage, each candidate disease description retrieved from FAISS is paired systematically with user symptom description and passed through the NLI model. The NLI model classifies relationships between symptom descriptions (premise) and disease descriptions (hypothesis) as entailment, contradiction, or neutral. Only those diseases falling under entailment classification are considered logically supported by symptoms. Contradicting or neutral relationships are filtered out appropriately.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

In this methodological way, the NLI stage significantly reduces false positive predictions and ensures final outputs are both semantically relevant and logically consistent with symptoms reported by users.

G. Disease Ranking

After NLI validation completes systematically, the remaining set of disease predictions undergoes further processing in the ranking stage. In this component, each valid disease is scored using combinations of semantic similarity distances and NLI confidence values. Diseases exhibiting higher similarity scores and stronger entailment confidence are ranked higher in final outputs presented to users.

The ranking module orders diseases logically so the most probable and medically relevant predictions appear prominently at the top of result lists. Final ranked results are formatted into structured responses containing disease names, confidence levels, and supporting clinical information. This structured output is utilized subsequently by frontend interfaces to present results clearly and understandably to users.

H. Triage Classification

While disease prediction remains important clinically, un- derstanding potential seriousness of user conditions proves equally critical. The triage classification module is responsible for assessing urgency associated with predicted diseases. It examines symptoms systematically for high-risk indicators including chest pain patterns, breathing difficulties, persistent high fever, sudden weakness occurrences, or neurological issues.

Based on predefined medical guidelines and classification logic, the triage module assigns each case into one of three ur- gency categories: low clinical risk, medium clinical risk, or high clinical risk requiring immediate attention. This classification helps users understand whether they need emergency medical attention, prompt consultation within days, or routine follow- up monitoring. By combining disease prediction with urgency assessment, the system becomes more practically useful in real-life healthcare situations where triage decisions matter substantially.

I. Use Case Design

Functional system behavior from user perspectives is de-scribed comprehensively through use case design methodology. Use Case Diagrams present main interactions between the system and its actors systematically.

The primary actor is the patient user, who can enter symptoms, choose languages, request disease predictions, and view triage results with rec- ommendations. Doctors act as secondary actors who may view prediction summaries and use them as preliminary decision support tools. System administrators oversee configuration management, data updates, and performance monitoring.

This use case design helps define clearly what operations are available to different user types, and also identifies system boundaries appropriately. It ensures every feature implemented in the system has clear purpose and corresponding actor inter- actions, supporting user-centered design principles throughout development.

J. Sequence and State Design

Detailed interactions between different components over time are explained through Sequence Diagrams depicting symptom analysis flows comprehensively. These diagrams show how user requests travel from frontend interfaces to FastAPI backends, move through normalization, embedding generation, FAISS search, NLI validation, ranking algorithms, and triage classification, then return to frontends as final structured responses. Sequence diagrams help understand exact operation orders, data flows across modules, and how each processing step depends on previous stages. They provide clear visualizations of asynchronous communication patterns and highlight importance of parallel execution in improving system responsiveness. By illustrating message-passing structures, these diagrams assist developers in debugging, optimizing latency, and ensuring modular consistency across pipelines.

Internal system behavior is further modeled using State Diagrams showing prediction processing lifecycles. The system transitions through different states including idle, receiving input, processing, validating, ranking, completed, and error states. Each state represents controlled phases in prediction request lifecycles, ensuring operations follow predictable and stable flows. State diagrams clarify how systems react to valid inputs, invalid inputs, exceptions, or timeouts, making error-handling strategies more transparent. They also ensure applications maintain robustness by preventing undefined states and ensuring orderly recovery during failures. This structured state management contributes substantially to reliability and consistency of prediction engines.

K. Security Architecture

Because the system handles sensitive healthcare-related information, security design forms a critical aspect of overall architecture.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Security Architecture Diagrams illustrate different security layers applied systematically throughout the system. At application layers, input validation and sanitization techniques prevent injection attacks and cross-site scripting vulnerabilities. At API layers, authentication mechanisms, rate limiting policies, and CORS configurations help control access and protect services from abuse.

At data layers, encryption protocols and secure storage ensure sensitive information is never exposed in plain text formats. Infrastructure-level security is enforced using firewalls, HTTPS (SSL/TLS) communication channels, and secure deployment practices. Logging and monitoring mechanisms are included to detect unusual activities and support auditing requirements. Together, these measures create multi-layered defense strategies to protect user privacy and system integrity comprehensively.

L. Database Design

Persistent storage structures are described using Entity Relationship Diagrams modeling main system entities including User, Symptom, Disease, Prediction, Specialist, Doctor, DiagnosticTest, and Triage records.

Each entity contains relevant attributes, and relationships are established through primary and foreign keys maintaining referential integrity. For example, Prediction entities link to both User and Disease entities, while Triage records associate with specific predictions. This relational design ensures data is stored systematically, can be queried efficiently, and remains consistent across system components. Well-structured database design also simplifies future extensions, such as adding new disease categories or specialist types seamlessly.

Additionally, database schemas are designed to support fast read and write operations so real-time predictions and recommendations can be delivered without delays. Proper indexing strategies are applied on frequently accessed attributes to improve query performance and reduce lookup times sub- stantially. Relational constraints help prevent data redundancy and maintain accurate associations between medical records. Database structures further support scalability by allowing seamless integration of future modules including patient history tracking, report storage, and analytical capabilities. Overall, database components act as strong foundations for reliable data management and long-term system stability.

IV. Mathematical Formulation

A. Embedding Generation Function

Let S represent the input symptom description in natural language, which may contain words from multiple languages including English, Hindi, Marathi, or mixed combinations. The preprocessing function P (·) normalizes the input:

Snorm = P (S) = lowercase(remove special chars(normalize unicode(S))

The embedding function E(·) maps the normalized text to a dense vector representation using a multilingual Sentence- Transformer model:

$$\mathbf{v}_s = E(S_{\text{norm}}) \in \mathbb{R}^d$$

where d = 384 dimensions for the paraphrase-multilingual-MiniLM-L12-v2 model. The embedding is normalized to unit length:

$$\hat{\mathbf{v}}_s = \frac{\mathbf{v}_s}{\|\mathbf{v}_s\|_2}$$

B. Disease Knowledge Base Representation

Let D = {D1, D2, . . . , DN } represent the set of N diseases in the knowledge base. Each disease Di has a textual description T (Di). All disease descriptions are preprocessed and embedded offline:

$$\mathbf{v}_{D_i} = E(P(T(D_i)))$$
$$\hat{\mathbf{v}}_{D_i} = \frac{\mathbf{v}_{D_i}}{\|\mathbf{v}_{D_i}\|_2}$$

The complete disease embedding matrix is:

$$\mathbf{V}_D = [\mathbf{v}_{D_1}, \mathbf{v}_{D_2}, \ldots, \mathbf{v}_{D_N}]^T \subset \mathbb{R}^{N \times d}$$

C. FAISS Similarity Search

FAISS indexes the disease embedding matrix using an optimized data structure. For a query symptom embedding $\hat{\mathbf{v}}_s$, FAISS computes the top-k most similar disease embeddings using cosine similarity:

$$\text{sim}_{\cos}(\hat{\mathbf{v}}_s, \hat{\mathbf{v}}_{D_i}) = \hat{\mathbf{v}}_s \cdot \hat{\mathbf{v}}_{D_i}^T$$

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

The search returns ordered candidates:

$C = \{(D_i, s_i) : i \in \text{top-}k \text{ indices sorted by decreasing similarity}\}$ where $s_i = \text{sim}_{\cos}(\hat{\mathbf{v}}_s, \hat{\mathbf{v}}_{D_i})$.

D. Natural Language Inference Formulation

For each candidate disease D_i with description $T(D_i)$, the NLI model M_{NLI} computes the probability distribution over three classes: entailment (e), contradiction (c), and neutral (n):

$$[p_e, p_c, p_n] = M_{\text{NLI}}(S_{\text{norm}}, T(D_i))$$

where $p_e + p_c + p_n = 1$. The entailment score p_e represents the logical support between symptoms and disease.

E. Composite Scoring and Ranking

The final score for disease D_i combines semantic similarity and logical entailment:

$$\text{score}_i = \alpha \cdot s_i + (1 - \alpha) \cdot p_e^{(i)}$$

where $\alpha \in [0, 1]$ is a weighting parameter (empirically set to 0.6). Diseases are filtered by an entailment threshold τ:

$$D_{\text{valid}} = \{D_i : p_e^{(i)} \geq \tau\}$$

with $\tau = 0.75$. The final ranked list is:

$$R = \text{argsort}_{D_i \in D_{\text{valid}}}(\text{score}_i, \text{descending})$$

$$R = \text{argsort}_{D \in D} (\text{score}_{i, \text{valid}}, \text{descending})$$

F. Triage Classification Rules

Let K be the set of keywords indicating high urgency: chest pain, difficulty breathing, sudden weakness, etc. The triage function T (S) is:

$$T(S) = \begin{cases} \text{High} & \text{if } \exists k \in K_{\text{high}} : k \in S \text{ certain patterns} \\ \text{Medium} & \text{if } \exists k \in K_{\text{medium}} : k \in S \text{ with patterns} \\ \text{Low} & \text{otherwise} \end{cases}$$

Specific pattern matching considers symptom combinations, duration modifiers, and intensity descriptors.

G. Complexity Analysis

The time complexity for inference consists of:

- Preprocessing: $O(|S|)$ where $|S|$ is input length
- Embedding generation: $O(L \cdot d^2)$ for transformer with L layers
- FAISS search: $O(\log N)$ for approximate nearest neighbor search
- NLI validation: $O(k \cdot L \cdot d^2)$ for k candidates
- Total: $O(|S| + (k+1)Ld^2 + \log N)$

Memory complexity is dominated by model parameters ($\approx$ 1.2GB) and FAISS index ($\approx N \times d \times 4$ bytes).

V. Experimental Setup

A. Development Environment Configuration

The development environment was carefully configured to meet system requirements for high-performance text processing, large-scale vector search, and real-time API responsiveness.

Python 3.9 served as the primary programming language due to its extensive ecosystem for machine learning, artificial intelligence, and natural language processing libraries. Python provides seamless compatibility with essential frameworks including HuggingFace Transformers, FAISS, PyTorch, and FastAPI, which form the technical core of this implementation. FastAPI was selected as the backend framework due to its exceptional execution speed and native support for asyn- chronous operations. Compared to traditional frameworks like Flask or Django, FastAPI ensures faster request handling and superior performance under concurrent workloads—critical requirements since embedding generation, FAISS searches, and NLI validation represent computationally intensive tasks requiring optimization.

TABLE I:
Composition of Evaluation Dataset

Language Category	Count	Symptom Complexity	Medical Domain
English	20	Simple to Complex	Multiple
Hindi (Devanagari)	15	Moderate	General Medicine
Marathi	10	Simple to Moderate	Regional Focus
Hinglish (Mixed)	5	Complex	Urban Patterns
Total	**50**	Varied	Comprehensive

The frontend interface was implemented using standard web technologies including HTML5, CSS3, and vanilla JavaScript to ensure accessibility across all devices including mobile phones, tablets, and desktop systems.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

The development en- vironment was further strengthened using Git-based version control systems, Python virtual environments for dependency isolation, and package managers to maintain consistency across development and testing environments.

B. Dataset Curation and Preparation

Due to the zero-shot learning paradigm, the system does not require traditional training datasets with symptom-disease pairs. However, for evaluation purposes, we curated a comprehensive test set of 50 real-world symptom descriptions representing diverse linguistic and clinical characteristics:

Each symptom description was validated by medical profes- sionals to ensure clinical accuracy and relevance. The disease knowledge base contained 150 common diseases with detailed textual descriptions sourced from reputable medical textbooks, peer-reviewed articles, and clinical guidelines.

C. Evaluation Metrics

We employed multiple evaluation metrics to assess different aspects of system performance:

1) Top-k Accuracy: Measures whether the clinically correct disease appears in the top k predictions (k = 1, 3, 5):

$$\text{Top-k Acc} = \frac{\#\text{correct in top } k}{N_{\text{total}}}$$

2) Mean Reciprocal Rank (MRR): Evaluates ranking quality:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$$

where rank_i is the position of the correct disease for the i-th query.

3) Entailment Confidence: Average NLI entailment probabil- ity for correct predictions:

$$\bar{p_e} = \frac{1}{N_{\text{correct}}}\sum_{i\in\text{correct}} p_e^{(i)}$$

4) End-to-End Latency: Time from API request receipt to response dispatch, measured at the server.
5) Triage Accuracy: Percentage agreement between system triage classification and expert clinical assessment.

TABLE II:
Comparative Performance Analysis

Model	Top-1	Top-3	Top-5	MRR
TF-IDF Baseline	0.28	0.42	0.54	0.41
FAISS-only	0.62	0.78	0.86	0.72
Proposed	**0.76**	**0.90**	**0.94**	**0.83**

D. Baseline Systems for Comparison

We implemented two baseline systems for comparative evaluation:

1) Baseline 1 (TF-IDF + Cosine Similarity): Traditional information retrieval approach using TF-IDF vectorization and cosine similarity without semantic understanding.
2) Baseline 2 (FAISS-only): Our system without the NLI validation layer, relying solely on semantic similarity from FAISS search.

E. Hardware Configuration

All experiments were conducted on a standardized hardware configuration:

- CPU: Intel Core i7-12700H (14 cores, 20 threads)
- GPU: NVIDIA GeForce RTX 3060 (6GB GDDR6)
- RAM: 32GB DDR4 3200MHz
- Storage: 1TB NVMe SSD
- OS: Ubuntu 22.04 LTS

F. Software Stack

The implementation utilized the following software components:

- Python Libraries: PyTorch 1.13, Transformers 4.26, Sentence-Transformers 2.2, FAISS 1.7, FastAPI 0.95,
- Uvicorn 0.21
- Models: paraphrase-multilingual-MiniLM-L12-v2, xlm- roberta-large-xnli
- Frontend: HTML5, CSS3, JavaScript (ES6)
- Deployment: Docker 23.0, Nginx 1.22

VI. Results And Discussion

A. Overall Prediction Performance

The system demonstrated strong performance across all eval- uation metrics, significantly outperforming baseline approaches: The 14 percentage point improvement in Top-1 accuracy over the FAISS-only baseline demonstrates the critical value added by the NLI validation layer in filtering out semantically similar but logically inconsistent predictions.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

B. Multilingual Performance Analysis

The system maintained consistent performance across dif- ferent language inputs, with minor variations attributable to training data distribution in the underlying multilingual model: The slightly lower performance for Hindi and Hinglish inputs reflects the relatively smaller proportion of these languages in the multilingual model's pretraining data compared to English.

TABLE III:
Performance Across Language Categories

Language	Top-3 Acc.	Latency (ms)	Entailment
English	0.92	820	0.82
Hindi	0.87	850	0.79
Marathi	0.89	840	0.81
Hinglish	0.85	830	0.78
Overall	0.90	846	0.81

TABLE IV:
Latency Breakdown by Processing Stage

Stage	Time (ms)	Percentage
Preprocessing	10	1.2%
Embedding	450	53.2%
FAISS	5	0.6%
NLI	350	41.4%
Triage	30	3.5%
Total	846	100%

TABLE V:
Latency Breakdown by Processing Stage

Stage	Time (ms)	Percentage
Preprocessing	10	1.2%
Embedding	450	53.2%
FAISS	5	0.6%
NLI	350	41.4%
Triage	30	3.5%
Total	846	100%

TABLE VI:
Ablation Study Results

Variant	Top-3	False Pos.	Satisfaction
Full System	0.90	0.08	4.5/5
Without NLI	0.78	0.22	3.2/5
Without Multilingual	0.65	0.15	2.8/5
Without Triage	0.89	0.08	3.9/5

C. Efficiency and Response Time Analysis

The system achieved real-time performance with average end-to-end latency of 846 milliseconds (standard deviation: 120 ms). The latency distribution across processing stages was:

The embedding generation and NLI validation stages rep- resent the primary computational bottlenecks, but their paral- lelization and GPU acceleration enabled sub-second response times.

D. Triage Classification Performance

The rule-based triage engine achieved 92% accuracy com- pared to expert clinical assessments. The confusion matrix revealed:

The system demonstrated conservative safety-oriented be- havior, with some over-triaging of medium-risk cases to high urgency—a clinically acceptable approach for preliminary screening tools.

E. Case Study Analysis

1) Case 1: Multilingual Cardiac Symptom: Input: ” ” (Marathi: ”I have severe chest pain and difficulty breathing”)

Processing:

1) Normalization preserved Marathi text with minor spelling correction

2) Embedding captured semantic similarity to cardiac conditions 3) FAISS retrieved: Angina, Myocardial Infarction, Pulmonary Embolism

4) NLI validation confirmed entailment for all three with high confidence

Output:

- Top Prediction: Acute Coronary Syndrome (0.92 confi- dence)
- Triage: HIGH urgency (red alert)
- Recommendation: Immediate emergency consultation, Cardiology specialist

2) Case 2: Code-Mixed Gastrointestinal Issue: Input: ”Mere pet mein bahut dard hai and vomiting ho rahi hai repeatedly” (Hinglish: ”My stomach hurts a lot and vomiting is happening repeatedly”)

Output:

- Top Prediction: Gastroenteritis (0.88 confidence)
- Triage: MEDIUM urgency
- Recommendation: Consult physician within 24 hours, Hydration advised

3) Case 3: Vague English Description: Input: ”Not feeling well, tired all the time”

Analysis: This vague input challenged the system, yielding broader predictions including anemia, depression, and chronic fatigue syndrome with lower confidence scores (0.65-0.72). The triage classification was LOW urgency, appropriate for non-specific symptoms.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

F. Ablation Studies

We conducted systematic ablation studies to understand individual component contributions:

The NLI module reduced false positives by 64%, while multilingual support improved user satisfaction significantly for non-English speakers.

G. Error Analysis

Detailed error analysis revealed three primary failure modes:

1) Vague Symptom Descriptions: Inputs like ”feel unwell” or ”body pain” lacked specificity for precise predictions.
2) Rare Disease Presentations: Conditions with atypical symptom patterns sometimes ranked lower despite semantic relevance.
3) Cultural/Linguistic Nuances: Some region-specific health expressions weren’t fully captured by the multilingual model.

These limitations highlight areas for future improvement while demonstrating the system’s robustness for common symptom patterns.

VII. Limitations And Ethical Considerations

A. Technical Limitations

1) Dependence on Input Clarity: The system’s effectiveness diminishes with extremely vague symptom descriptions that lack specific details about location, duration, intensity, or associated symptoms.
2) Translation and Cultural Nuances: While the multilingual model handles major Indian languages effectively, deeply colloquial expressions, regional dialects, or culturally spe- cific health metaphors may not translate accurately into the semantic embedding space.
3) Knowledge Base Limitations: Predictions are inherently constrained by the breadth and depth of the disease knowledge base. Very rare conditions, newly emerging diseases, or region-specific illnesses without comprehensive descriptions may not be identified accurately.
4) Absence of Clinical Context: The system lacks access to critical clinical information including vital signs, laboratory results, medical imaging findings, detailed patient history, medication records, and physical examination data—all essential for definitive diagnosis.
5) Inability to Handle Contradictory Information: When users provide symptom descriptions containing internal contradictions or conflicting temporal information, the system may produce inconsistent or unreliable predictions.
6) Static Knowledge Representation: The disease knowledge base requires manual updates to incorporate new medical research, changed clinical guidelines, or emerging health threats, creating maintenance overhead.

B. Clinical and Practical Limitations

1) Not a Diagnostic Tool: The system serves strictly as a preliminary health information and triage tool. It cannot and should not replace professional medical evaluation, diagnosis, or treatment decisions by qualified healthcare providers.
2) No Physical Examination Capability: Critical diagnostic information obtained through physical examination (pal- pation, auscultation, percussion, etc.) remains completely unavailable to the system.
3) Limited to Symptom-Based Reasoning: The approach can- not incorporate diagnostic test results, imaging findings, or procedural outcomes that often provide definitive diagnostic evidence.
4) Potential for Over-Reliance: Users might develop ex- cessive dependence on the system, delaying necessary professional consultation even when symptoms warrant immediate attention.
5) Algorithmic Bias Concerns: Like all AI systems, the models may reflect biases present in training data, poten- tially disadvantaging certain demographic groups or disease presentations.

C. Ethical Considerations

1) Informed Consent and Transparency: Users must receive clear disclosures about system capabilities and limitations, understanding it provides informational support only.
2) Privacy and Data Security: Symptom descriptions consti- tute sensitive health information requiring robust encryption, access controls, and data protection measures compliant with regulations.
3) Accountability Framework: Clear protocols must establish responsibility when discrepancies occur between system suggestions and actual medical conditions.
4) Accessibility and Equity: The system should remain freely accessible to underserved populations while avoiding technologies that create or exacerbate healthcare disparities.

5) Continuous Monitoring and Improvement: Regular audits should evaluate system performance across diverse popula- tions, with mechanisms for reporting errors or concerns.

These limitations and ethical considerations highlight the importance of positioning this technology as a supplementary healthcare tool rather than a replacement for professional medical care. They also provide clear direction for future research and development efforts to enhance system capabilities while maintaining ethical standards.

VIII. Conclusion

This research has successfully demonstrated the devel- opment and implementation of a comprehensive Zero-Shot Learning based Multilingual Medical Symptom Understanding and Disease Inference System that represents a significant advancement in accessible healthcare artificial intelligence. The system addresses critical challenges in medical AI—particularly the scarcity of labeled training data for numerous diseases and the linguistic diversity of patient populations—through an innovative integration of multilingual transformer models, semantic embedding techniques, high-performance similarity search, logical inference validation, and clinically informed triage classification.

The core achievements of this work include:

1) Successful Zero-Shot Implementation: The system achieves accurate disease predictions without any disease- specific training data, relying instead on semantic under- standing of symptom and disease descriptions.
2) Effective Multilingual Processing: Native support for English, Hindi, Marathi, and code-mixed inputs makes the system genuinely accessible to diverse linguistic populations in India and similar multilingual regions.
3) Semantic-Logical Hybrid Architecture: The combination of FAISS-based semantic retrieval with XLM-RoBERTa NLI validation ensures predictions are both contextually relevant and logically consistent, reducing false positives substantially.
4) Practical Triage Integration: The rule-based triage engine provides actionable urgency classification, enhancing real- world utility by helping users prioritize healthcare decisions appropriately.
5) Real-Time Performance: With average response times under 850 milliseconds, the system meets practical require- ments for interactive health assistance applications.
6) Scalable and Extensible Design: The modular architecture allows seamless integration of new diseases, languages, and features without fundamental re-engineering.

Experimental evaluation on carefully curated multilingual datasets demonstrated strong performance metrics, including 90% Top-3 accuracy, 0.83 Mean Reciprocal Rank, and 92% triage classification accuracy. The system consistently outper- formed traditional baseline approaches, particularly in handling informal, multilingual symptom expressions that challenge conventional medical AI systems.

Beyond technical achievements, this work makes important contributions to healthcare accessibility by bridging linguistic divides in medical technology. It empowers users to describe symptoms naturally in their preferred languages while receiving medically meaningful preliminary guidance. This addresses significant barriers in regions where English proficiency cannot be assumed and where traditional symptom checkers fail due to language limitations.

The research also advances methodological understanding of zero-shot learning applications in healthcare, demonstrating practical integration of multiple advanced NLP components into a cohesive, user-centric system. It provides a replicable blueprint for developing similar tools for other multilingual healthcare environments worldwide.

While the system exhibits certain limitations—particularly regarding vague symptom descriptions and the inherent con- straints of symptom-only analysis—these represent opportu- nities for future enhancement rather than fundamental flaws. The ethical framework developed alongside the technical im- plementation ensures responsible deployment with appropriate safeguards and user education.

In conclusion, this zero-shot multilingual medical symptom understanding system represents a meaningful step toward democratizing access to preliminary health information across linguistic and cultural boundaries. By combining cutting-edge AI techniques with thoughtful design for real-world healthcare contexts, it demonstrates how technology can complement clinical practice to support earlier health awareness, more informed decision-making, and ultimately better health out- comes for diverse populations. The principles and architectures developed here provide a foundation for continued innovation in accessible, equitable, and effective healthcare artificial intelligence.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

IX. Future Work

The successful implementation of this zero-shot multilin- gual medical symptom understanding system establishes a strong foundation for numerous avenues of future research, development, and practical deployment. Based on lessons learned during system development, experimental evaluation, and identified limitations, we propose the following directions for future work:

A. Advanced Model Enhancements

1) Medical Domain Fine-tuning: Develop specialized ver- sions of multilingual transformer models through continued pretraining on large corpora of Indian medical literature, clinical notes, patient forums, and healthcare educational materials in multiple Indian languages to enhance medical semantic understanding.
2) Hierarchical Embedding Architectures: Implement multi- level embedding approaches that capture symptom relation- ships at different granularities—from individual symptom mentions to comprehensive case descriptions—improving matching precision for complex presentations.
3) Contrastive Learning for Medical Concepts: Employ contrastive learning techniques to better separate medically distinct but lexically similar concepts (e.g., different types of headaches or abdominal pains) in the embedding space.
4) Few-Shot Learning Integration: Combine zero-shot ca- pabilities with few-shot learning approaches where limited labeled examples exist for certain disease categories, creat- ing a hybrid system that adapts based on data availability.

B. Expanded Linguistic and Cultural Capabilities

1) Additional Indian Language Support: Extend language coverage to include other major Indian languages such as Gujarati, Tamil, Telugu, Bengali, Punjabi, and Odia, ad- dressing the full linguistic diversity of the Indian population.
2) Dialect and Regional Variation Handling: Develop mechanisms to recognize and process regional dialects, colloquial health expressions, and culturally specific symp- tom descriptions that may not follow standard language patterns.
3) Code-Switching Detection and Processing: Enhance algorithms to better detect and interpret intra-sentential code- switching patterns common in urban Indian communication, improving accuracy for mixed-language inputs.
4) Culturally Informed Symptom Interpretation: Incorpo- rate cultural context into symptom understanding, recog- nizing that symptom expression and health communication styles vary across cultural groups within India.

C. Clinical Integration and Enhancement

1) Patient History Integration: Develop modules to incorpo- rate relevant patient medical history, demographics, risk factors, and medication information into the prediction process for more personalized and accurate assessments.
2) Vital Signs and Lab Value Integration: Create interfaces to accept basic clinical measurements (temperature, blood pressure, heart rate) and laboratory results when available, enhancing prediction accuracy with objective data.
3) Differential Diagnosis Generation: Evolve the system from single disease prediction to generating ranked differential diagnoses with supporting evidence and reasoning for each possibility.
4) Clinical Guideline Integration: Embed latest clinical practice guidelines and evidence-based medicine principles into the reasoning process, ensuring recommendations align with current best practices.

D. User Interaction and Experience Improvements

1) Conversational Symptom Elicitation: Transform the sys- tem from single-turn input to multi-turn conversational interfaces that ask clarifying questions to resolve ambiguities and gather missing information systematically.
2) Multimodal Input Support: Extend input modalities to include voice recordings (with automatic speech recognition for regional languages), image uploads of visible symptoms (rashes, swellings, injuries), and structured data entry for measurements.
3) Personalized Health Profiles: Allow users to create secure personal health profiles tracking symptoms over time, enabling longitudinal analysis and early detection of concerning patterns.
4) Explainable AI Enhancements: Develop comprehensive explanation systems that clearly communicate why specific diseases were predicted, which symptoms supported which possibilities, and what additional information would help refine predictions.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

E. Deployment and Scalability Advancements

1) Edge Computing Deployment: Create optimized versions for deployment on mobile devices with limited connectivity, enabling usage in remote areas with poor internet access through periodic knowledge base updates.
2) Federated Learning Implementation: Develop federated learning approaches allowing multiple healthcare institutions to collaboratively improve models without sharing sensitive patient data, addressing privacy concerns while enhancing model performance.
3) Real-Time Healthcare Ecosystem Integration: Connect the system with existing healthcare infrastructure including hospital bed availability systems, telemedicine platforms, ambulance services, and appointment scheduling systems for seamless care coordination.
4) API Standardization and Interoperability: Develop stan- dardized APIs following healthcare interoperability stan- dards (HL7 FHIR) to facilitate integration with electronic health record systems and other healthcare IT infrastructure.

F. Research and Validation Initiatives

1) Large-Scale Clinical Validation: Conduct rigorous prospec- tive studies in clinical settings across different regions of India to validate system performance, clinical utility, and impact on healthcare outcomes with diverse patient populations.
2) Comparative Effectiveness Research: Design studies comparing this system's performance against traditional symptom checkers, telemedicine consultations, and in- person primary care visits for common symptom patterns.
3) Health Equity Impact Assessment: Systematically evaluate how the system affects healthcare access disparities across different socioeconomic, linguistic, and geographic groups, with particular attention to underserved populations.
4) Longitudinal Outcome Studies: Track long-term health outcomes for users who engage with the system compared to matched controls, assessing impact on appropriate healthcare utilization, early diagnosis rates, and patient satisfaction.

G. Ethical and Regulatory Development

1) Bias Detection and Mitigation Frameworks: Develop comprehensive methodologies to detect, measure, and mitigate potential biases in system performance across different demographic groups, ensuring equitable service quality.
2) Regulatory Pathway Development: Work with healthcare regulators to establish appropriate certification pathways for AI-based symptom assessment tools, balancing innovation with patient safety considerations.
3) Ethical Use Guidelines: Create detailed guidelines for appropriate system use, including clear indications, con- traindications, and recommended integration into clinical workflows without disrupting patient-provider relationships.
4) Transparency and Auditability Standards: Implement mechanisms for system decision auditability, allowing healthcare providers to review and understand AI reasoning when used as decision support in clinical contexts.

H. Specialized Application Development

1) Pediatric Symptom Assessment: Develop specialized versions for children's symptoms, accounting for devel- opmental stages, different symptom presentation patterns, and pediatric-specific conditions.
2) Geriatric Health Monitoring: Create adaptations for elderly populations considering multiple comorbidities, polypharmacy implications, and age-related changes in symptom presentation.
3) Mental Health Screening: Extend capabilities to include preliminary mental health assessment while maintaining ap- propriate safeguards and referral pathways for psychological conditions.
4) Occupational Health Applications: Develop workplace- specific versions addressing common occupational expo- sures, injuries, and work-related health concerns with appropriate employer integration while protecting worker privacy.

These future directions collectively represent a comprehen- sive roadmap for evolving the current system from a promising research prototype to a robust, widely-deployed healthcare tool that can meaningfully contribute to improving health access and outcomes across India's diverse population.

The modular architecture of the current implementation facilitates incremental development along these multiple dimensions, allowing prioritization based on practical impact and resource availability.

Acknowledgment

The authors express sincere gratitude to Ms. Nitu L. Pariyal, Project Guide and Assistant Professor, Department of Computer Science & Engineering, MGM's College of Engineering, Nanded, for her invaluable guidance, continuous support, and insightful suggestions throughout this research project. We acknowledge Dr. A. M. Rajurkar, Head of Computer Science and Engineering Department, and Dr. G. S. Lathkar, Director of MGM's College of Engineering, Nanded, for providing necessary resources and facilities. We also thank Dr. Babasaheb Ambedkar Technological University, Lonere, for academic support and framework. Finally, we appreciate all individuals who contributed directly or indirectly to the successful completion of this research work.

REFERENCES

[1] E. H. Shortliffe et al., "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system," Comput. Biomed. Res., vol. 8, no. 4, pp. 303–320, 1975.

[2] R. A. Miller, H. E. Pople, and J. D. Myers, "INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine," N. Engl. J. Med., vol. 307, no. 8, pp. 468–476, 1982.

[3] WebMD LLC, "WebMD Symptom Checker," [Online]. Available: https://symptoms.webmd.com

[4] Ada Health GmbH, "Ada – Your Health Companion," [Online]. Available: https://ada.com

[5] Babylon Health, "AI-Powered Healthcare," [Online]. Available: https://www.babylonhealth.com

[6] Infermedica, "Symptom Checker and Triage API," [Online]. Available: https://infermedica.com

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[8] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[10] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[11] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2251–2265, 2018.

[12] J. Johnson, M. Douze, and H. Je´gou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547, 2019.

[13] Facebook AI Research, "FAISS: A library for efficient similarity search," [Online]. Available: https://github.com/facebookresearch/faiss

[14] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," Proc. 2015 Conf. Empir. Methods Nat. Lang. Process., pp. 632–642, 2015.

[15] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proc. 2019 Conf. Empir. Methods Nat. Lang. Process., 2019.

[17] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," N. Engl. J. Med., vol. 380, no. 14, pp. 1347–1358, 2019.

[18] W. Wang and K. Cho, "BERTScore: Evaluating text generation with BERT," arXiv preprint arXiv:1904.09675, 2020.

[19] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

[20] "Zero-Shot Disease Predictor — Simple Explanation," Internal Project Document, MGM's College of Engineering, Nanded, 2025.