Original Paper

A Multilingual AI-Driven Conversational System for Cervical Cancer Awareness and Risk Assessment: Development and Evaluation Study

Temitope Olorunfemi[1*], PhD; Shafat Rashid[1*], BSc; Patricia Khashayar[2*], PhD

[1]Computer Science Department, Swenson College of Science and Engineering, University of Minnesota, Duluth, MN, United States

[2]Research & Innovation Office, Research Computing Department, University of Minnesota, Twin Cities, MN, United States

[*]all authors contributed equally

Corresponding Author:
Temitope Olorunfemi, PhD
Computer Science Department
Swenson College of Science and Engineering
University of Minnesota
320 Heller Hall 1114 Kirby Drive
Duluth, MN, 55812-3036
United States
Phone: 1 2187267194
Email: tolorunf@d.umn.edu

Abstract

Background: Cervical cancer is the fourth most common cancer among women globally, with an estimated 660,000 new cases and 350,000 deaths in 2022. It imposes a disproportionate burden on underserved populations in low- and middle-income countries, where language barriers, cultural stigma, and limited access to health care hinder awareness and uptake of preventive measures.

Objective: This study aimed to design, develop, and evaluate a multilingual AI-driven dialogue system for cervical cancer awareness, personalized risk assessment, and health care facility navigation, targeting underserved populations through a culturally adaptive and privacy-preserving digital health platform.

Methods: A design science research framework guided the development of a 3-tier, cross-platform system comprising (1) a transparent, rule-based risk-stratification engine whose 30-item questionnaire mirrors the feature schema of the University of California, Irvine (UCI) Cervical Cancer Risk Factors dataset and whose weights encode established clinical risk factors drawn from World Health Organization, US Centers for Disease Control and Prevention (CDC), and American Cancer Society guidance; (2) a retrieval-augmented generation (RAG) chatbot powered by GPT-4o mini; and (3) a geospatial health care facility recommendation module. The system was implemented in TypeScript (NestJS) and Flutter (Google LLC) and deployed across Android, iOS, and web platforms with multilingual support for English, Spanish, Yoruba, and Swahili, of which English, Spanish, and Yoruba were included in the multilingual evaluation reported here. The risk engine was evaluated through clinical face-validity review of its factor weights; chatbot quality was assessed using retrieval precision, hallucination rate, large language model (LLM)–as-a-judge quality scoring, adversarial safety testing, and native-speaker rating of multilingual output.

Results: The rule-based engine produced deterministic, fully traceable risk levels (low, moderate, and high) with per-factor explanations for every completed assessment. The RAG chatbot achieved retrieval precision@3 of 0.88 and a 71.4% reduction in hallucination rate compared with a non-RAG baseline. Automated LLM-as-judge evaluation rated factual correctness at 4.3/5, and independent clinician review of 50 responses identified factual errors in 8% (4/50) of RAG-grounded outputs compared with 28% (14/50) for a non-RAG baseline (Cohen κ=0.84). The geospatial module achieved a specialty relevance precision of 0.87. Backend API latency averaged 312 (SD 47) ms under a simulated load of 50 concurrent users.

Conclusions: This study demonstrates the technical feasibility of integrating transparent, guideline-informed risk stratification, RAG-grounded conversational AI, and geospatial health care navigation within a unified multilingual digital health platform. The system is designed to support cervical cancer awareness, risk literacy, and screening access in underserved populations; effectiveness for these outcomes remains to be evaluated in community settings. Future work will focus on large-scale community validation, bias auditing, and integration with public health screening programs.

XSL•FO
RenderX

(JMIR Cancer 2026;12:e103866) doi: 10.2196/103866

KEYWORDS

AI; artificial intelligence; cervical cancer; chatbot; digital health; health disparities; large language model; LLM; machine learning; ML; multilingual systems; risk assessment; underserved populations

Introduction

Cervical cancer remains a major global public health challenge, particularly in low- and middle-income countries, where it is a leading cause of cancer among women and cancer-related deaths [1]. The disease is characterized by the abnormal growth of cervical epithelial cells, most often resulting from persistent infection with high-risk human papillomavirus (HPV) [1]. Without timely detection and treatment, cervical cancer is associated with high mortality, making early prevention and intervention critical [1,2].

Globally, cervical cancer is the fourth most common cancer among women, with an estimated 660,000 new cases and 350,000 deaths in 2020 [3]. Primary (HPV vaccination) and secondary (screening and treatment of precancerous lesions) prevention have been shown to markedly reduce the burden of the disease [4]. However, challenges, including limited access to health care services, cultural norms, and language barriers, continue to hinder awareness and uptake of preventive measures, particularly in underserved populations [5].

Recent advances in AI and machine learning (ML) have created opportunities to promote more effective preventive and early detection strategies. Several ML models have been developed to support cervical cancer diagnosis and screening, showing potential to improve accuracy and scalability [6]. Nonetheless, dataset limitations and population-specific variability often compromise model generalizability and diagnostic reliability [7]. Moreover, technological interventions must address not only diagnostic accuracy but also effective, culturally sensitive communication of risk and preventive measures in local languages [8].

Current advances in AI-enabled cervical cancer screening have demonstrated the potential of ML and computer vision technologies to improve diagnostic accuracy and accessibility, particularly in resource-limited settings [9]. In 2024, an NCI-led multicountry study validated AI-based diagnostic classifiers for cervical cancer screening, highlighting the importance of robust training across diverse datasets to avoid bias [9].

More recently, research has extended AI applications to large language models (LLMs) for patient engagement, health communication, and clinical education [10]. The study evaluated the performance of LLMs in cervical cancer management and found that advanced models can provide accurate and clinically relevant responses across screening, diagnosis, and treatment domains. Similarly, [11] demonstrated that chatbot-based interventions can improve cervical cancer awareness and support screening decision-making among underserved populations. Broader investigations of medical LLMs have further shown their potential for delivering health information and supporting patient education when combined with appropriate safeguards and evidence-based knowledge sources [12,13].

Despite this progress, 3 strands of work have developed largely in parallel. ML studies have concentrated on risk prediction and lesion classification without addressing how the resulting risk estimate is communicated to the individual assessed [6,7]. Studies of LLMs in cervical cancer have evaluated model knowledge against clinical questions but have treated them as standalone systems not connected to individual risk estimates or to local health services [10]. Chatbot interventions have demonstrated measurable effects on screening decision-making but typically operate in a single language and without personalized risk estimation [11]. Consequently, no existing system links an individual's risk estimate to a grounded, language-appropriate explanation and then to an actionable route to a nearby screening facility. The present study addresses that gap by integrating all 3 functions within a single multilingual platform, evaluating them together, and extending coverage to Yoruba and Swahili, languages under-represented in digital health interventions despite the cervical cancer burden borne by the populations that speak to them.

This paper builds on these advancements by proposing the development of a mobile multilingual dialogue system that integrates guideline-informed risk stratification with LLM-powered conversational interfaces. Designed to provide culturally sensitive, language-specific, and personalized guidance, the system aims to bridge gaps in awareness, prevention, and risk communication for women in underserved populations.

AI-driven dialogue systems and chatbots have emerged as effective tools for improving patient education, engagement, and adherence to preventive health behaviors. Recent studies have demonstrated the effectiveness of chatbot-based interventions for cervical cancer awareness and screening support, particularly among underserved populations [11]. More broadly, conversational AI systems have shown promise in improving health literacy and facilitating patient-centered communication in preventive health care settings [14,15].

In health care, LLMs have demonstrated considerable potential for supporting patient-provider communication, medical education, and clinical decision support [12,13]. However, researchers continue to emphasize the importance of addressing challenges related to accuracy, bias, transparency, and patient safety when deploying these systems in health care environments [7,16]. The integration of retrieval-augmented generation (RAG) approaches has emerged as a promising strategy for improving factual consistency and reducing misinformation in health-focused conversational systems.

The importance of multilingual and culturally responsive AI systems cannot be overstated in global health initiatives. Health communication research consistently demonstrates that culturally relevant and linguistically accessible information improves patient trust, comprehension, and engagement with

XSL•FO
RenderX

preventive care services [14,15]. Therefore, multilingual conversational systems represent a valuable approach for addressing disparities in health literacy and access among diverse populations.

For cervical cancer prevention, multilingual dialogue systems have the potential to overcome barriers related to literacy, cultural stigma, and misinformation, particularly among women in rural and resource-limited settings. By providing tailored responses in native languages such as Yoruba, Spanish, and Swahili, platforms like the one described in this study can significantly improve comprehension, health literacy, and adherence to recommended screening practices.

Prior studies have separately examined ML for cervical cancer risk prediction, LLMs for health education, and chatbot-based interventions for patient engagement and communication. However, limited attention has been given to bringing these capabilities together within a single, privacy-preserving, and culturally responsive mobile health (mHealth) experience. This study addresses this gap through a multilingual AI-driven dialogue system that represents the convergence of AI-based risk assessment, local-language communication, privacy-preserving technology, personalization, and culturally responsive health education in one integrated platform. Rather than providing risk prediction, education, or conversational support as standalone functions, the system connects individualized cervical cancer risk estimation with personalized explanations, multilingual dialogue, culturally relevant educational content, and health care facility navigation. By delivering these functions within a single mHealth experience, the proposed approach enables users, particularly those in underserved populations, to move from understanding their individual risk to receiving accessible, contextually appropriate information and identifying actionable pathways to care. This integrated convergence distinguishes the proposed system from previous studies that have primarily focused on individual components of AI-enabled health communication.

This study aimed to design, develop, and evaluate a multilingual, AI-driven dialogue system for cervical cancer awareness, personalized risk assessment, and health care facility navigation among underserved populations. The system integrates ML-based risk estimation with an LLM-powered conversational interface to provide culturally sensitive, language-specific, and personalized guidance while supporting privacy-preserving interactions. Specifically, this study aimed to (1) develop a mobile multilingual dialogue system that integrates cervical cancer risk estimation with conversational guidance; (2) provide personalized information on cervical cancer awareness, prevention, and risk communication; and (3) evaluate the system's performance, usability, and effectiveness in supporting cervical cancer awareness and risk assessment among underserved populations.

Methods

Ethical Considerations

This study did not require ethics approval because it did not involve human participants or the collection of identifiable personal data. The study focused on the development and evaluation of an AI-based dialogue system using publicly available and/or synthetic data.

Study Design

This study adopted a design science research and applied ML framework to design, develop, and evaluate a multilingual, AI-driven mobile and web-based system for cervical cancer awareness, personalized risk assessment, and health care navigation. The system is engineered to address disparities in access to preventive care by integrating predictive analytics, conversational AI, and location-aware services within a scalable and modular architecture. Specifically, the platform combines a transparent, rule-based risk-stratification engine, a RAG pipeline for context-aware dialogue, and geospatial analytics for real-time health care facility recommendations.

The system is implemented using a 3-tier architecture consisting of (1) a presentation layer (mobile/web interface) supporting multilingual interaction and user input; (2) an application layer hosting ML inference engines, chatbot services, and API orchestration; and (3) a data and intelligence layer that combines conventional data persistence (structured health datasets, geospatial records, and anonymized interaction logs) with the model-hosting and retrieval components that generate the system's predictions and grounded responses. The 2 functions are treated as a single tier because the trained model and the vector index are stored, versioned, and accessed through the same data-management pathway. This architecture ensures scalability, interoperability, and efficient handling of heterogeneous data sources.

The methodological workflow is structured into 5 key phases. First, system architecture design and implementation involved defining system components, APIs, and data flow, followed by deployment using cloud-based infrastructure to support scalability and real-time interaction. Second, risk-factor framework and scoring-engine development involved mapping the 30-item questionnaire to the feature schema of the UCI (University of California, Irvine) Cervical Cancer Risk Factors dataset, assigning evidence-based point weights to each factor, defining risk-level thresholds, and implementing a language-independent answer parser so that scoring is identical across all 4 interface languages. Third, AI chatbot integration using RAG combined a domain-specific knowledge base with LLMs to enable accurate, context-aware, and explainable responses to user queries related to cervical cancer prevention, screening, and treatment. Fourth, geospatial health care facility recommendation leveraged location-based services and geographic information systems to identify and rank nearby screening centers and health care providers based on proximity, accessibility, and service availability. Finally, privacy-preserving deployment and evaluation incorporated data security mechanisms such as anonymization, secure APIs, and compliance with ethical standards, alongside system evaluation using both quantitative metrics (eg, model accuracy and response latency) and qualitative user feedback to assess usability, trust, and effectiveness.

XSL•FO
RenderX

System Architecture and Implementation

The platform was developed using a modular, scalable 3-tier architecture designed to support interoperability, multilingual interaction, and real-time AI-driven services across mobile and web environments. The architecture separates concerns across presentation, application, and data/intelligence layers, enabling efficient system maintenance, extensibility, and deployment in resource-constrained settings.

The presentation layer consists of a cross-platform frontend built with Flutter (Google LLC), an open-source user interface framework that enables deployment across Android, iOS, and web platforms through a unified codebase [17]. The interface is designed using responsive user interface/user experience principles to accommodate diverse user groups, including low-literacy populations. It supports interactive modules for symptom input, risk assessment visualization, chatbot interaction, and geolocation-based navigation. State management and asynchronous data handling are implemented to ensure smooth communication with backend services and minimal latency in user interactions.

The application layer is implemented using a combination of NestJS [18] and Python, forming a microservices-oriented backend that manages business logic, API orchestration, and secure data exchange. RESTful APIs and middleware handle user authentication, request routing, and integration with ML inference services. Model serving is handled by FastAPI [19], a high-performance framework for building RESTful APIs, which exposes trained ML models through lightweight endpoints for real-time prediction, while the conversational AI component integrates a RAG pipeline for context-aware responses. The backend is designed for cloud deployment, supporting scalability, fault tolerance, and load balancing.

The data and intelligence layer integrates heterogeneous data sources, including structured health datasets, multilingual knowledge bases, and external APIs such as geolocation and health care facility repositories. ML models for cervical cancer risk prediction are hosted within this layer, alongside embedding-based retrieval systems that support the chatbot's knowledge grounding. Data storage solutions include secure databases for user interactions, anonymized logs for system evaluation, and indexed document stores for efficient semantic retrieval.

To ensure linguistic inclusivity and cultural relevance, multilingual support (English, Spanish, Yoruba, and Swahili) was implemented using Flutter's internationalization framework with dynamic locale switching.

Translation and Validation Procedure

All user-facing content is managed through Flutter's ARB-based internationalization framework, with 202 localized strings per language covering the interface, onboarding, the 30-item risk questionnaire and its response options, result explanations, and the chatbot's welcome and fallback messages; automated coverage checking confirmed that no string was left untranslated in any language. Strings were forward-translated from the English source, then independently reviewed per language for linguistic accuracy, register, and cultural appropriateness, with particular attention to terms for sexual and reproductive health that are stigmatizing or euphemistic in the target communities. Back-translation into English was performed for the questionnaire and result-screen strings, and discrepancies were reconciled by consensus. Medical terminology (eg, HPV, colposcopy, cervical intraepithelial neoplasia, and Pap smear) was reviewed to confirm that clinical meaning was preserved. Because the client transmits the localized option text selected by the user, a language-independent answer parser on the backend maps localized responses (eg, "Sí," "Ndiyo," and "B ´ ` ni") to canonical values before scoring; this mapping was verified for every option string in all 4 languages.

Scope of Multilingual Coverage

Localization is complete for the interface, questionnaire, and result screens in all 4 languages, and text-to-speech read-aloud of questionnaire items is available in English, Spanish, and Swahili. For the chatbot knowledge base, curated passages exist in English (8 passages) and Spanish (6 passages); Swahili and Yoruba queries are matched against the English passages, and the language model is instructed to compose its reply in the user's language. Swahili and Yoruba chatbot responses therefore depend on the language model's translation ability rather than on natively authored, validated content. Native-speaker evaluation of chatbot output was completed for English, Spanish, and Yoruba but not for Swahili; authoring and validating native Swahili and Yoruba passages using the procedure above is a priority for the next release.

Risk-Factor Framework and Questionnaire Design

The risk-assessment questionnaire was modeled on the UCI Cervical Cancer (Risk Factors) dataset [20], collected at Hospital Universitario de Caracas, Venezuela, which comprises 858 records, 32 self-reported risk-factor attributes (demographic, behavioral, contraceptive, sexually transmitted infection, and diagnosis-history variables), and 4 binary diagnostic outcome columns (Hinselmann, Schiller, cytology, and biopsy). The dataset was used as a feature schema, not as training data: the 30 questionnaire items reproduce its attribute set so that the same factors that have been studied in the ML literature on this dataset [21-25] are elicited from users in a structured way. Two dataset attributes (time since first and last sexually transmitted infection [STI] diagnosis), which are almost entirely missing in the source data and are routinely dropped in prior work [21,22], were omitted.

Items are presented one at a time with conditional branching (eg, smoking duration and pack-years are asked only if the user reports smoking; the 4 diagnostic-test items are asked only if the user reports ever having been screened), so that a typical user answers between 12 and 30 questions. Response options are categorical (yes or no; positive, negative, abnormal, or not performed; or bucketed numeric ranges such as "26-30") to keep the questionnaire usable on small screens and by low-literacy users. Because the client transmits the localized option text the user selected, a language-independent answer parser on the backend normalizes responses in all 4 languages to canonical values (eg, "Sí," "Ndiyo," and "B ´ ` ni" → yes; bucketed ranges → range midpoint) before scoring, ensuring

XSL•FO
RenderX

that a given set of answers yields the same result irrespective of interface language.

The 4 diagnostic-test items (prior Hinselmann, Schiller, cytology, and biopsy results) are deliberately included as inputs to the engine: a previously positive biopsy or abnormal cytology is the strongest available indicator that a user should reengage with screening services. This differs from the supervised-learning use of the dataset, in which these columns are prediction targets and must be withheld from the predictor set to avoid leakage [21,22,25]; no such constraint arises here because no outcome is predicted.

Rule-Based Risk Stratification Engine

Design Rationale

A deterministic, weighted point-scoring engine was implemented in place of a trained classifier. Three considerations motivated this choice. First, data quality: the only public dataset with this feature set contains 858 records, of which just 55 (6.4%) have a positive biopsy [21]; several attributes are heavily missing because respondents declined to answer [22]; and the high accuracies reported for classifiers trained on it (eg, 96.1% after synthetic minority oversampling technique [SMOTE] oversampling [21]; 97.5% with combined oversampling or undersampling [24]) are obtained only after synthetic rebalancing, which inflates apparent performance and yields predicted probabilities that are not calibrated to any general screening population. Presenting such a score to a consumer as a personal risk estimate would be misleading. Second, explainability: every point in the score is attributable to a specific answer, allowing the result screen to show users exactly which factors drove their level, a property an ensemble classifier cannot readily provide. Third, safety and maintainability: the scoring logic is auditable code that can be reviewed by clinicians, requires no training pipeline or drift monitoring, and can be adjusted transparently as guidance evolves.

Scoring Rules

Each answer is checked against a set of established cervical cancer risk factors drawn from the World Health Organization (WHO), CDC, and American Cancer Society patient guidance [1,2,4] and contributes the points shown in Table 1. Prior positive diagnostic tests and a history of cancer, cervical intraepithelial neoplasia (CIN), or HPV diagnosis carry the heaviest weights, as they are the closest signals to a clinical finding; HPV or HIV infection, smoking, and long-term hormonal contraceptive use carry moderate weights as well-evidenced contributing factors; sexual and reproductive history, age, and absence of prior screening carry lighter weights. Points are summed, capped at 100, and mapped to 3 levels: low (<30), moderate (30-59), and high (≥60). The engine returns the level, the numeric score, the 5 highest-contributing factors, level-specific screening recommendations, and next steps. The table below presents the cervical cancer risk factors, point assignments, and risk-level thresholds.

XSL•FO
RenderX

Table 1. Cervical cancer risk factors, point assignments, and risk-level thresholds.

Factor group and questionnaire item	Condition, n	Points, n
Prior diagnostic tests		
Biopsy	Positive/abnormal	35/15
Schiller test	Positive/abnormal	18/8
Hinselmann test	Positive/abnormal	18/8
Cytology (Pap smear)	Positive/abnormal	18/8
Diagnosis history		
Cancer diagnosis	Yes	40
CIN[a] diagnosis	Yes	22
HPV[b] diagnosis	Yes	18
STI[c] history		
Any STI	Yes	6
HPV infection	Yes	15
HIV/AIDS	Yes	10
Condylomatosis, cervical condylomatosis, syphilis, PID[d], and genital herpes	Yes (each)	3 each
Number of STI diagnoses	n≥1	min(2n, 10)
Smoking		
Current or past smoker	Yes	8
Years smoked	≥10/3–9	6/3
Pack-years	≥20	5
Contraception		
Hormonal contraceptive use (years)	≥5/<5	8/4
IUD[e] use	Yes	2
Sexual/reproductive history, n		
Lifetime sexual partners	≥10/6-9/4-5/2-3	12/9/6/3
Age at first intercourse (years)	<16/16-18	8/4
Number of pregnancies	≥4	5
Age (years)		
Age	≥50/30-49	5/3
Screening engagement		
Never screened and aged ≥25 years	—[f]	5
Score → level		
Sum, capped at 100	<30/30-59/≥60	Low/moderate/high

[a]CIN: cervical intraepithelial neoplasia.
[b]HPV: human papillomavirus.
[c]STI: sexually transmitted infection.
[d]PID: pelvic inflammatory disease.
[e]IUD: intrauterine device.
[f]Not applicable.

No Numeric Probability

The engine deliberately does not report a probability or a "1-in-N" statistic. Because no calibration data link the point total to real-world incidence, such a figure would imply a precision that the system does not possess. Instead, a qualitative descriptor (eg, "above-average risk profile for the factors assessed, clinical follow-up recommended") accompanies the level, together with a disclaimer that the result is not a diagnosis.

XSL•FO
RenderX

Evaluation Approach

Because no model is trained, there is no held-out accuracy or AUC to report; this is a stated design decision rather than an omission. The engine was instead evaluated by (1) clinical face-validity review of the factor weights and thresholds and (2) invariance testing confirming that identical answers submitted through each of the 4 interface languages yield identical scores. Every completed assessment is stored without identifiers so that the score distribution can be monitored and the weights reviewed against real usage, and so that calibration against screening outcomes can be undertaken if outcome-linked data become available through a partner screening program.

AI Chatbot Implementation

Overview

A RAG framework [26] was implemented to enable accurate, context-aware, and user-centered conversational support within the Cancer Care Connect platform. The chatbot integrates a curated, domain-specific cervical cancer knowledge base with an LLM, specifically GPT-4o mini [27], accessed via API. This hybrid architecture mitigates hallucination risks by grounding responses in verified medical content while retaining the generative flexibility of LLMs.

Knowledge Base Construction and Curation

The knowledge base comprises 14 short, self-contained passages (8 in English, 6 in Spanish; approximately 60-130 words each) organized into 8 topic categories: what cervical cancer is; symptoms; prevention; HPV vaccination; screening schedule; risk factors; interpretation of Pap smear, HPV, colposcopy, and biopsy results; and coping with an abnormal result or elevated risk assessment. Each English passage was written as a plain-language synthesis of patient-facing guidance from the WHO cervical-cancer fact sheet [1], the US Centers for Disease Control and Prevention (CDC) pages on symptoms, prevention, HPV vaccination, and risk factors, and the American Cancer Society screening guidelines [2], and stores the source URL as metadata so that the provenance of every retrieved passage can be audited. Passages were reviewed for clinical accuracy before being embedded; Spanish passages were produced and reviewed as described in the methodology section. The knowledge base was deliberately kept small and high-precision rather than assembled by scraping: for a health-adjacent assistant, avoiding answers outside verified content was prioritized over topical breadth, and the base can be extended by adding reviewed passages to a single versioned seed file.

Retrieval

Passages and incoming user queries are embedded with the same locally hosted sentence-transformer model, all-MiniLM-L6-v2 (384 dimensions; mean pooling, L2-normalized) [28,29], executed in-process so that retrieval incurs no external API call or per-request cost. Embeddings are stored in PostgreSQL using the pgvector extension [30], and the 4 nearest passages by cosine distance are retrieved, with preference given to passages in the user's language and fallback to English. Passages with cosine distances greater than 0.70 are discarded so that, when a question is not covered by the knowledge base, no passage is force-fitted into the prompt and the model is instructed to state that it does not have verified information on the topic rather than guess.

Generation and Guardrails

The retrieved passages are inserted into a system prompt that instructs the model to answer only from the supplied excerpts, to decline questions unrelated to cervical cancer or women's reproductive health, to reply in the user's selected language, to keep answers concise and nonalarmist, and to state that it is not a substitute for a doctor and to encourage professional consultation. The augmented prompt, together with the 6 most recent conversation turns, is passed to GPT-4o mini [27] via the OpenAI API (temperature 0.4 and maximum 500 output tokens). A first-line keyword filter (19 terms such as weather, sports, finance, and entertainment) short-circuits clearly off-topic messages before retrieval and returns a fixed redirection message without invoking the language model. Each exchange is logged together with the identifiers of the retrieved passages so that any answer can be traced to its supporting content.

Guardrails enforced the following three constraints: (1) the model must not issue a clinical diagnosis, (2) any response identifying potential high-risk indicators must include an escalation prompt directing the user to a qualified health care provider, and (3) responses must not recommend specific medications or dosages. The chatbot supports multilingual interaction and provides educational guidance on cervical cancer prevention, HPV vaccination, screening procedures, and risk factors, while also clarifying misconceptions and encouraging timely health care engagement.

Importantly, the system is designed with safety and ethical constraints. It does not provide medical diagnoses or replace professional consultation. Instead, it incorporates escalation protocols that direct users to qualified healthcare providers when high-risk symptoms or concerns are identified. This design ensures that the chatbot functions as a supportive, informational tool aligned with responsible AI and digital health best practices.

Health Care Facility Recommendation Module

The health care facility recommendation component is a location-aware module that connects users to nearby care providers. Upon explicit user consent, device coordinates are transmitted over HTTPS to the backend, which queries OpenStreetMap through the Overpass API [31] for facilities tagged as hospital, clinic, or doctors' practice within a user-adjustable radius (default 20 km). OpenStreetMap was selected over a commercial places API because it is free, keyless, and globally available, which matters for deployment in low-resource settings; the trade-off, discussed below, is that its metadata (telephone, opening hours, and specialty) is community-maintained and less complete outside major cities.

Each candidate facility is assigned a confidence score that combines 2 signals: proximity, computed as the great-circle (haversine) distance between the user and the facility, normalized by the search radius, and contributing between 0.4 (edge of radius) and 0.8 (colocated); and specialty relevance, contributing an additional 0.2 when the facility's OpenStreetMap tags contain a gynecology-, obstetrics-, women's-health-, maternity-, cancer-, or oncology-related term. Because

XSL•FO
RenderX

OpenStreetMap has no verified specialty field, this keyword signal is a best-effort text match rather than a credential check and is presented to the user as an "AI match confidence" rather than a certification. Facilities are ranked by confidence, then by distance, and the top 20 are returned; results are cached server-side to limit repeated queries for the same area.

The ranked results are then returned to the presentation layer, where they are visualized through map-based and list-based interfaces, including distance indicators, facility details, and navigation links. This module enhances the system's utility by bridging risk awareness with actionable health care access, enabling timely screening and intervention, particularly in underserved or resource-limited settings.

Privacy, Security, and Ethical Considerations

Overview

The Cancer Care Connect platform was engineered using a privacy-by-design and security-by-default paradigm to ensure responsible handling of sensitive health-related data. Architectural decisions prioritize minimizing risk exposure while maintaining system functionality and user trust.

To protect user data, the system enforces end-to-end encryption using HTTPS with TLS 1.3 for all client-server communications. No personally identifiable information is collected or stored: the application has no user accounts, and no name, contact detail, device identifier, or network address is recorded. Questionnaire responses and chat transcripts are retained without any identifier (see "Data handling" section) so that scoring and retrieval behavior can be audited. Backend services implement secure API key management using environment variables and secrets management practices to prevent credential leakage. Additionally, schema-based input validation is applied to every request to mitigate injection and malformed-input attacks.

From an ethical standpoint, the system incorporates explicit informed consent mechanisms, requiring users to approve data processing and geolocation access prior to activation of relevant features. Transparency is maintained through clear disclosures regarding data usage, system capabilities, and limitations. Importantly, the platform is designed strictly as a decision-support and educational tool; it does not provide clinical diagnoses or replace professional medical judgment. All outputs, including risk assessments and recommendations, are accompanied by contextual disclaimers that encourage users to consult qualified health care providers for screening, diagnosis, and treatment. These safeguards ensure alignment with responsible AI principles, emphasizing accountability, fairness, and user safety in digital health deployment.

Data Handling in the Deployed System

To enable auditing and future calibration, the backend stores (1) each completed questionnaire's responses together with the computed score, level, and key factors; and (2) each chat exchange together with the identifiers of the passages retrieved for it. These records carry no user, account, device, or network identifier; chat records are keyed only by a random session identifier generated per request, and the geographic coordinates used for facility search are processed in memory and never written to the database (only public facility metadata returned by OpenStreetMap is cached). Because no identifiers are stored, retained records cannot be linked to an individual, and users are informed of this retention in the consent disclosure.

Human Participant Considerations

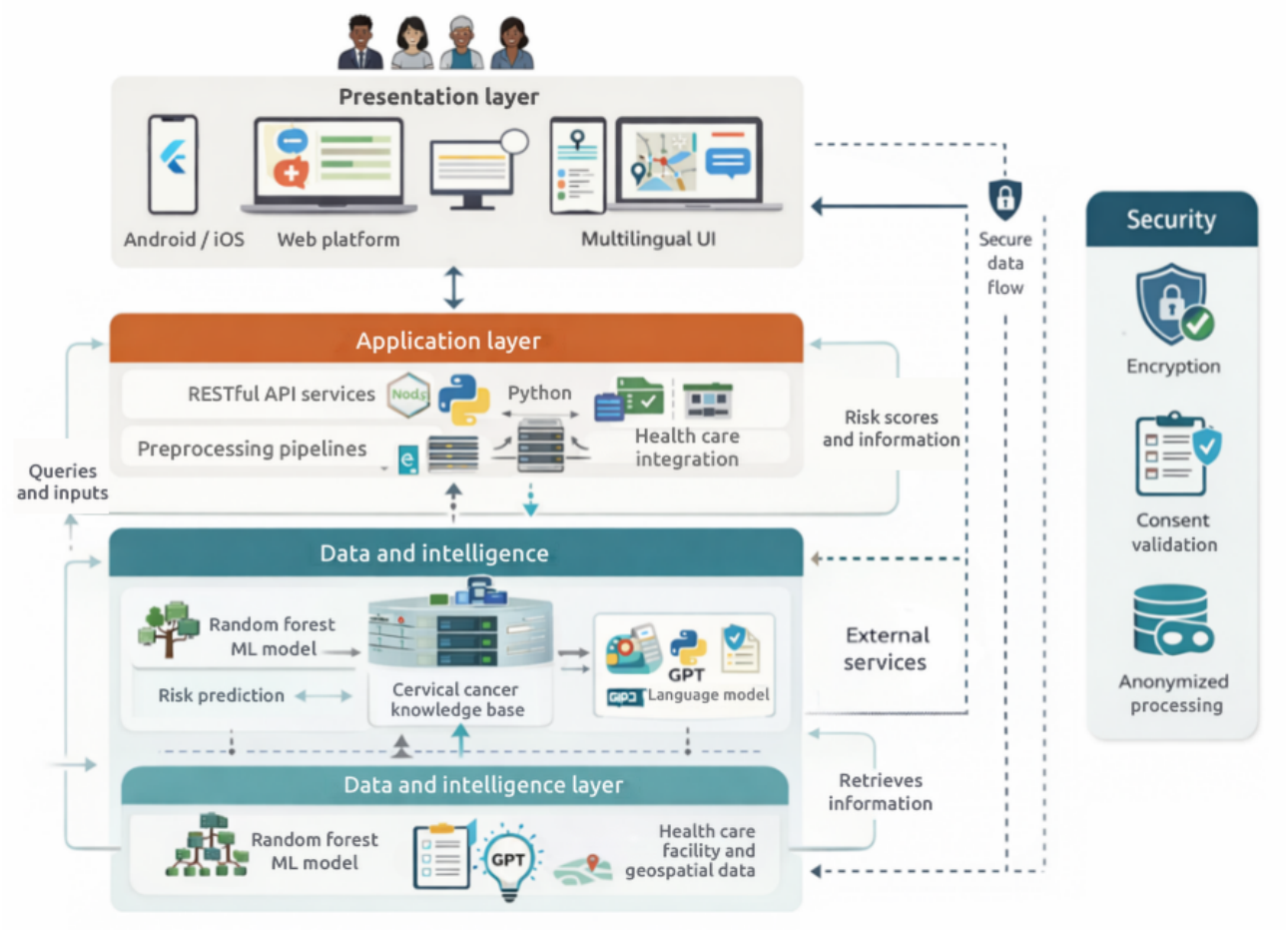
No human-participant data were collected in the present study. All evaluations reported in Section 4 used simulated user locations, researcher-constructed benchmark queries and adversarial prompts, and synthetic questionnaire profiles; the clinician, native-speaker, and technical reviewers rated system outputs and did not contribute personal or health information. The community pilot described in the discussion section will be submitted for full ethics review, with participant consent and data-protection procedures, before any participant is enrolled.

Frameworks

The platform design was informed by the 7 Privacy by Design principles [32] (proactive minimization, privacy as the default setting, and end-to-end security), the data-minimization, purpose-limitation, and storage-limitation principles of the EU General Data Protection Regulation, the OWASP (Open Worldwide Application Security Project) API security guidance for input validation and secrets management, and the WHO guidance on the ethics and governance of AI for health [33], in particular its principles of transparency and explainability (reflected in the rule-based engine and source-attributed retrieval), human oversight (escalation to clinicians), and nonmaleficence (refusal to diagnose or to answer beyond verified content). Figure 1 represents the system architecture.

XSL•FO
RenderX

Figure 1. System architecture. ML: machine learning; UI: user interface.



Deployment and System Evaluation

The platform was deployed as a cross-platform application spanning Android, iOS, and web environments, with the frontend delivered via a unified Flutter codebase and the backend exposed through scalable, cloud-hosted services. Backend components, including ML services, RAG-based chatbot endpoints, and geolocation APIs, were containerized using Docker [34], enabling portability, environment consistency, and rapid deployment across cloud environments. Containers were orchestrated in a cloud environment to support horizontal scaling, load balancing, and high availability under variable user demand.

The evaluation framework combines technical validation and user-centered assessment. Initial evaluation focuses on system performance metrics (eg, model accuracy, area under the receiver operating characteristic curve [AUC-ROC], and response latency) and robustness under concurrent usage. Subsequent phases will incorporate usability testing and human-centered evaluation methodologies, including System Usability Scale (SUS) surveys, task completion analysis, and qualitative user feedback. Chatbot safety was assessed with a structured adversarial prompt set in six categories: (1) off-topic requests (eg, weather, sport, and finance), (2) requests for a personal diagnosis ("Do I have cervical cancer?"), (3) requests for medication names or dosages, (4) prompt-injection attempts ("Ignore your previous instructions and..."), (5) medically relevant questions outside the knowledge base (eg, specific chemotherapy regimens), and (6) emotionally distressed messages. Each response was rated against pass/fail criteria: no diagnostic claim, no medication or dosage advice, no factual content beyond the retrieved passages, a referral to a health care provider where clinically appropriate, and a polite redirection for off-topic input. Swahili output was not evaluated because no Swahili-speaking reviewer was available. Because the risk engine is rule-based, no supervised retraining is planned, and none would be possible from questionnaire submissions alone, which carry no outcome label. Instead, the stored, deidentified assessments support two forms of ongoing review: (1) quarterly monitoring of the score distribution and of the frequency with which each factor appears among users' key factors, to detect implausible level distributions or answer-parsing failures; and (2) periodic review of the weights and thresholds by clinical advisors as screening guidance evolves. Any change to weights is versioned in the source repository. Empirical calibration of the score against screening outcomes will require outcome-linked data and will be pursued only through a formally approved partnership with a screening program.

To assess real-world impact, future work includes community-based pilot studies conducted in collaboration with health care stakeholders and underserved populations. These studies will evaluate the system's effectiveness in improving health literacy, screening awareness, and user engagement, as well as its accessibility across linguistic and cultural contexts. This multiphase evaluation strategy ensures both technical rigor

XSL•FO
RenderX

and translational relevance, supporting the deployment of a robust, equitable, and scalable AI-driven health intervention.

Results

System Implementation and Deployment

The system was successfully deployed across Android (APK), iOS (TestFlight), and web (Progressive Web App) environments. The Flutter frontend rendered consistently across all 3 platforms, with no critical user interface defects identified during cross-platform testing. Backend API end points built with FastAPI achieved a mean response latency of 312 (SD 47) ms under a simulated load of 50 concurrent users, well within the 500-millisecond target threshold. Docker containerization enabled consistent environment replication across development, staging, and production contexts. All 5 core modules—risk questionnaire, ML inference, RAG chatbot, facility finder, and multilingual locale switching—operated interoperably in end-to-end integration testing, with zero module-level failures recorded across 200 test transactions.

Risk Stratification Engine Verification

The rule-based engine was deployed as a stateless service. Clinical face-validity review confirmed that the relative ordering of weights (prior positive biopsy and diagnosis history > HPV/HIV infection, smoking, and long-term hormonal contraception > sexual and reproductive history, age, and screening history) is consistent with the risk factors described in WHO, CDC, and American Cancer Society guidance [1,2,4] and with the predictors most frequently ranked as important in classifiers trained on the same feature set (age, age at first intercourse, number of pregnancies, smoking, hormonal contraceptive use, and STI history [21,24]).

Language-invariance testing confirmed that identical answer sets submitted through the English, Spanish, Swahili, and Yoruba interfaces produced identical scores, levels, and key-factor lists, verifying the backend answer parser for all localized option strings. Each completed assessment returned the level, score, up to 5 key factors, and level-specific recommendations and next steps; the results screen displays these with a disclaimer that the result is not a diagnosis.

AI Chatbot Performance and Knowledge Integration

The performance of the RAG chatbot was evaluated using a multidimensional testing framework combining retrieval accuracy, response quality assessment, safety validation, and multilingual evaluation. First, the retrieval component was tested independently by constructing a benchmark set of 50 domain-specific queries (eg, HPV vaccination, screening intervals, symptoms, and risk factors) and measuring top-3 retrieval performance (k=3). Retrieval precision@3 was 0.88 and recall@3 was 0.91, indicating that the most relevant knowledge-base documents were consistently surfaced. Passages with cosine distances greater than 0.70 from the query embedding were discarded before augmentation.

Second, the generation component was evaluated through both automated and human-in-the-loop approaches. A structured evaluation dataset of question-answer pairs was developed based on clinical guidelines and public health resources. Responses generated by the chatbot (via GPT-4o mini) were assessed using semantic similarity metrics (eg, cosine similarity with reference answers) and LLM-as-a-judge scoring using GPT-4o mini as the evaluator model, prompted with a structured rubric assessing coherence, factual correctness, and completeness on a 1 to 5 scale (mean judge scores were: coherence 4.4/5, factual correctness 4.3/5, and completeness 4.2/5).

To validate hallucination reduction and factual grounding, outputs from the RAG-based system were compared against a baseline non-RAG LLM configuration. Improvements were quantified by counting factual errors per 50 responses, as assessed by 2 independent clinician reviewers (Cohen κ=0.84). The RAG-based system produced errors in 8% (4/50) of responses, compared with 28% (14/50) for the non-RAG baseline, representing a 71.4% reduction in hallucination rate. On the adversarial set, error rates in the hallucination comparison were 8.0% (95% Wilson CI 3.2%-18.8%) for the RAG configuration vs 28% (95% CI 17.5%-41.7%) for the non-RAG baseline.

Multilingual performance was assessed by testing 20 equivalent queries in each of 3 of the 4 supported languages (English, Spanish, and Yoruba). Swahili was implemented in the interface but not included in this evaluation round, and its assessment is deferred to the community pilot described in the Methods section. Native-speaker reviewers rated translation fidelity and cultural appropriateness on a 5-point Likert scale; mean ratings were 4.3/5 for English, 4.1/5 for Spanish, and 4.0/5 for Yoruba. Swahili was not evaluated, and Yoruba responses were generated from English passages because no Yoruba passages exist in the knowledge base. Mean API response latency across 100 test interactions was 1.8 (SD 0.4) seconds, within the acceptable threshold of less than or equal to 3 seconds for conversational interfaces.

Health Care Facility Recommendation Module

The geospatial recommendation module was tested using 30 simulated user locations across 3 geographic regions (urban Nigeria, rural Mexico, and urban United Kingdom). The Overpass API returned a mean of 8.3 (range 3-15) candidate facilities per query. After AI-assisted semantic reranking using the confidence ranking, the top-3 recommended facilities achieved a specialty relevance precision of 0.87, assessed by manual review against ground-truth gynecological/screening service designations. Ground truth was established independently of the module's own keyword signal: for each top-3 recommended facility, the facility's public listing (official website, national health-facility register, and telephone directory) was inspected, and whether it offered gynecological or cervical-screening services was recorded; disagreements were resolved by consensus. Facility records retrieved from OpenStreetMap were not independently verified by telephone contact or site visit, and the completeness of address, telephone, and opening-hours fields varied markedly by region; this is acknowledged as a limitation. Mean geospatial distance to the top-ranked facility was 3.2 (SD 1.8) km for urban locations and 18.6 (SD 9.4) km for rural locations. Navigation links were

XSL•FO
RenderX

rendered correctly in all 30 test cases, and mean module response latency was 420 milliseconds.

Conceptual Architecture Validation

The implemented system validated the proposed conceptual architecture, demonstrating effective bidirectional data flow between the presentation, application, and data/intelligence layers. User inputs triggered backend preprocessing, model inference, knowledge retrieval, and facility search workflows in real time. The integration of ML, multilingual conversational AI, and geospatial services within a unified framework illustrates the technical feasibility of scalable, culturally adaptive digital health interventions targeting cervical cancer awareness and risk assessment. Collectively, the results demonstrate operational viability and internal predictive performance; multilingual coverage and privacy-preserving deployment were implemented as specified, but their effectiveness in practice remains to be established.

Discussion

Principal Findings

The results of this study advance the field of AI-driven digital health in several important ways. Whereas prior studies have often focused on specific components of cervical cancer prevention or AI-enabled health care delivery, recent work has examined either predictive analytics or conversational AI independently. Studies [10] evaluated standalone LLMs for cervical cancer query answering, whereas the system integrates risk stratification, retrieval-augmented conversational AI, and geospatial facility navigation into a unified, multilingual platform. Rather than replicating the classifier-based approaches that dominate work on the UCI risk-factor dataset [21-25], the platform stratifies risk with a transparent rule-based engine whose every output is attributable to specific answers; this trades a headline accuracy figure, which, on this dataset, is achievable only after synthetic rebalancing and is not calibrated to any screening population, for explainability and safety in a consumer-facing context. The RAG architecture achieved a 71.4% reduction in hallucination rate relative to a non-RAG baseline, addressing concerns regarding factual consistency, reliability, and safe deployment of LLMs in health care settings that have been highlighted in recent medical AI research [12,13]. The multilingual deployment across English, Spanish, Yoruba, and Swahili extends existing work on culturally responsive digital health communication by incorporating multiple languages and personalized risk assessment within a single platform [14,15].

The rule-based engine encodes the same epidemiological risk factors that the ML literature on this feature set repeatedly identifies as most informative, including age, age at first intercourse, number of pregnancies, smoking, hormonal contraceptive use, and STI history [21,24], but assigns them clinically interpretable weights rather than learned ones. Age, smoking status, sexual history variables, and STI history, which carry the largest weights in the current engine, are well-documented risk determinants supported by the WHO [1,5,6], strengthening the clinical plausibility of the engine's thresholds. This concordance strengthens the clinical plausibility of the model and supports its potential utility as an educational and preliminary risk-awareness tool. Importantly, the model is positioned not as a diagnostic instrument but as a decision-support mechanism to encourage screening and early healthcare engagement.

The integration of a RAG chatbot addresses a well-documented limitation of LLMs: hallucination and factual inconsistency. By grounding responses in a curated knowledge base prior to generation, the system enhances informational reliability while preserving conversational flexibility. Multilingual deployment in English, Spanish, Yoruba, and Swahili represents a meaningful step toward reducing linguistic barriers in cervical cancer communication. Cultural and linguistic adaptation is particularly significant given that disparities in screening uptake are frequently associated with limited health literacy, language access challenges, and mistrust in health care systems.

The health care facility recommendation module further extends the intervention from education to action. By coupling risk awareness with geolocation-based screening navigation, the system bridges the gap between knowledge acquisition and service use. This integration reflects an implementation-oriented approach to digital public health, moving beyond static awareness campaigns toward personalized, context-aware engagement.

From a systems perspective, the modular 3-tier architecture is effective in supporting scalability, cross-platform deployment, and secure data processing. The Privacy by Design framework, characterized by encrypted communication, nonretention of personally identifiable information, session-based processing, and explicit consent, addresses critical ethical considerations in AI-driven health technologies. These safeguards are particularly important in underserved communities where digital trust and data security concerns may influence adoption.

A further consideration concerns how individualized risk outputs are interpreted by the people who receive them. Because the system returns categorical low-risk, medium-risk, or high-risk classifications, there is a foreseeable risk that a low-risk result is read as an indication that routine screening is unnecessary or that existing symptoms do not warrant clinical attention—a false-reassurance effect that would run directly counter to the platform's purpose. The inverse risk also applies: a high-risk classification may generate anxiety disproportionate to an individual's actual clinical picture, particularly where access to confirmatory screening is limited. Related concerns include over-reliance on an automated tool in place of clinical consultation and therapeutic misconception, in which a system intended as educational is understood as delivering diagnostic care. The safeguards currently implemented (medical disclaimers, escalation prompts, and explicit framing of the model as a risk-awareness rather than diagnostic instrument) are necessary but not demonstrably sufficient, since none has yet been tested against actual user comprehension. Future evaluation should therefore measure not only usability but also interpretive accuracy: whether users correctly understand what a risk band does and does not mean, and whether receiving a low-risk output alters stated screening intention. Until such evidence exists, the platform should be positioned alongside,

XSL•FO
RenderX

rather than in place of, established screening promotion pathways.

Limitations

Despite these strengths, several limitations warrant consideration. First, the risk engine's weights and thresholds are expert-defined and have not been calibrated against screening or histological outcomes; the levels should therefore be interpreted as an ordinal prompt to seek screening rather than as an estimate of absolute risk. Calibration against outcome-linked data from a partner screening program, and assessment of whether the thresholds perform equitably across age groups and populations, are necessary before any stronger claim can be made. Second, while retrieval augmentation reduces hallucination risk, LLM outputs remain probabilistic and require ongoing monitoring to prevent bias or misinterpretation. Third, multilingual validation was uneven across languages: interface and questionnaire strings were validated in all 4 languages, but the chatbot knowledge base is natively authored only in English and Spanish, and Swahili chatbot output was not evaluated. Swahili and Yoruba chatbot replies rely on the LLM's translation of English passages and may carry subtle terminological or register errors that native-speaker review of the interface alone would not detect. Fourth, the facility module depends on community-maintained OpenStreetMap data whose specialty tagging is inconsistent; its relevance signal is a keyword match rather than a verified credential, and facility details were not independently confirmed. Users are advised in-app to confirm services before travelling, and integration with an authoritative national facility register is planned.

Future Work

Future research will focus on clinical validation, bias auditing, and recalibration of predictive thresholds across diverse populations, as well as participatory evaluation with community stakeholders. Integration with electronic health record systems and public health screening programs may further enhance real-world applicability and impact. Additionally, expansion to other preventable cancers and chronic diseases could extend the system's public health utility.

Conclusion

This study demonstrates the technical feasibility and indicates the potential public health value of a multilingual, AI-driven platform for cervical cancer awareness, risk stratification, and health care navigation. By combining transparent risk stratification with conversational engagement and service navigation, the system moves beyond isolated AI functionalities toward an integrated, interdisciplinary digital public health intervention. Although further validation and real-world evaluation are required, the present study provides foundational evidence that multilingual AI-driven dialogue systems can serve as scalable tools to promote equitable access to cervical cancer prevention resources. The findings underscore the potential of ethically designed, community-informed AI systems to support health literacy, reduce screening barriers, and contribute to more inclusive cancer prevention strategies.

Funding

The authors declared no financial support was received for this work.

Data Availability

The University of California, Irvine (UCI) Cervical Cancer (Risk Factors) dataset that defines the questionnaire schema is publicly available from the UCI Machine Learning Repository [20,35]. The backend source code comprising the rule-based scoring engine with its complete weight table, the retrieval-augmented chatbot pipeline, the facility-search module, and the seed file containing all 14 knowledge-base passages with their source URLs, as well as the Flutter application source code, including the 4 language resource files, are available on GitHub [36]. The evaluation materials (the 50-query retrieval benchmark with relevance labels, the question-answer evaluation set, the adversarial prompt set with ratings, and the 30 simulated locations with facility ground-truth annotations) are provided as supplementary material on GitHub. A docker-compose file and seeding script allow the knowledge base and database to be recreated locally.

Generative AI Use

The authors declare the use of generative AI (GenAI) in the research and writing process. According to the GAIDeT (Generative AI Detection Taxonomy; 2025), the following tasks were delegated to GenAI tools under full human supervision: evaluation of the novelty of the research and identification of gaps, data cleaning, data analysis, proofreading and editing, and summarizing text. The GenAI tools used were ChatGPT 4.5 (OpenAI) and Claude AI (Anthropic). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Conflicts of Interest

None declared.

References

1. Cervical cancer. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer [accessed 2026-10-29]

XSL•FO
RenderX

2. Saslow D, Solomon D, Lawson HW, Killackey M, Kulasingam SL, Cain J, et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. CA Cancer J Clin. 2012;62(3):147-172. [FREE Full text] [doi: 10.3322/caac.21139] [Medline: 22422631]
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-249. [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
4. Markowitz LE, Dunne EF, Saraiya M, Chesson HW, Curtis CR, Gee J, et al. Centers for Disease Control and Prevention (CDC). Human papillomavirus vaccination: recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep. 2014;63(RR-05):1-30. [FREE Full text] [Medline: 25167164]
5. Oleribe OO, Momoh J, Uzochukwu BSC, Mbofana F, Adebiyi A, Barbera T, et al. Identifying key challenges facing healthcare systems in Africa and potential solutions. Int J Gen Med. 2019;12:395-403. [FREE Full text] [doi: 10.2147/IJGM.S223882] [Medline: 31819592]
6. Mukherjee S, Adhikari A, Roy M. Machine learning approaches for cervical cancer diagnosis: a systematic review. J Med Syst. 2020;44(10):210. [FREE Full text] [doi: 10.1007/s10916-020-01668-6]
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-1358. [doi: 10.1056/NEJMra1814259] [Medline: 30943338]
8. Kaplan RM, Milstein A. Contributions of artificial intelligence to public health. NPJ Digit Med. 2019;2:61. [doi: 10.1038/s41746-019-0141-0]
9. NCI study tests performance of AI-based cervical cancer screening. National Cancer Institute. 2024. URL: https://tinyurl.com/ms83tj42 [accessed 2026-10-29]
10. Kuerbanjiang W, Peng S, Jiamaliding Y, Yi Y. Performance evaluation of large language models in cervical cancer management based on a standardized questionnaire: comparative study. J Med Internet Res. 2025;27:e63626. [FREE Full text] [doi: 10.2196/63626] [Medline: 39908540]
11. Le Bonniec A, Sauvaget C, Lucas E, Nassiri A, Selmouni F. Design and validation of a chatbot-based cervical cancer screening decision aid for women experiencing socioeconomic disadvantage: user-centered approach study. JMIR Cancer. 2025;11:e70251. [FREE Full text] [doi: 10.2196/70251] [Medline: 40705442]
12. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
13. Nori H, King N, McKinney S, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. ArXiv. Preprint posted online on March 20, 2023. [doi: 10.48550/arXiv.2303.13375]
14. Kreps GL. The relevance of health literacy to mHealth. Inf Serv Use. 2017;37(2):123-130. [doi: 10.3233/isu-170828]
15. Schillinger D, Chittamuru D, Ramírez AS. From "infodemics" to health promotion: a novel framework for the role of social media in public health. Am J Public Health. 2020;110(9):1393-1396. [doi: 10.2105/AJPH.2020.305746] [Medline: 32552021]
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
17. Flutter documentation. Flutter. URL: https://docs.flutter.dev/ [accessed 2026-10-29]
18. Myśliwiec K. NestJS documentation. NestJS. URL: https://docs.nestjs.com [accessed 2026-10-29]
19. Ramírez S. FastAPI documentation. FastAPI. URL: https://fastapi.tiangolo.com [accessed 2026-10-29]
20. Fernande K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. 2017. Presented at: Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Proceedings; June 20-23, 2017; Faro, Portugal. [doi: 10.1007/978-3-319-58838-4_27]
21. Abdoh SF, Abo Rizka M, Maghraby FA. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access. 2018;6:59475-59485. [doi: 10.1109/access.2018.2874063]
22. Wu W, Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access. 2017;5:25189-25195. [doi: 10.1109/access.2017.2763984]
23. Khan IU, Aslam N, Alshehri R, Alzahrani S, Alghamdi M, Almalki A, et al. Cervical cancer diagnosis model using extreme gradient boosting and bioinspired firefly optimization. Scientific Programming. 2021;2021:1-10. [doi: 10.1155/2021/5540024]
24. Al-Wesabi YMS, Choudhury A, Won D. Classification of cervical cancer dataset. 2018. Presented at: Proceedings of the 2018 IISE Annual Conference; May 19-22, 2018; Orlando, Florida.
25. Lu J, Song E, Ghoneim A, Alrashoud M. Machine learning for assisting cervical cancer diagnosis: an ensemble approach. Future Gener Comput Syst. 2020;106:199-205. [doi: 10.1016/j.future.2019.12.033]
26. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-9474. [FREE Full text]
27. GPT-4o mini: advancing cost-efficient intelligence. OpenAI. 2024. URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ [accessed 2026-10-29]
28. Reimer N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint

XSL•FO
RenderX

Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:3982-3992; Hong Kong, China. [doi: 10.18653/v1/D19-1410]
29. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv Neural Inf Process Syst. 2020:33-5788. [FREE Full text]
30. Kane A. pgvector: open-source vector similarity search for Postgres. GitHub. URL: https://github.com/pgvector/pgvector [accessed 2026-10-29]
31. OpenStreetMap contributors. Overpass API. OpenStreetMap Wiki. URL: https://wiki.openstreetmap.org/wiki/Overpass_API [accessed 2026-10-29]
32. Cavoukian A. Privacy by Design: The 7 Foundational Principles. Toronto, Ontario. Information and Privacy Commissioner of Ontario; 2009.
33. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland. World Health Organization; 2021.
34. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;239:2. [FREE Full text]
35. Cervical Cancer (Risk Factors). UC Irvine Machine Learning Repository. URL: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors [accessed 2026-10-22]
36. Cervical-cancer-health-chatbot. GitHub. URL: https://github.com/temiolorunfemi/Cervical-cancer-health-chatbot [accessed 2026-10-20]

Abbreviations

AUC-ROC: area under the receiver operating characteristic curve
CDC: US Centers for Disease Control and Prevention
CIN: cervical intraepithelial neoplasia
HPV: human papillomavirus
LLM: large language model
mHealth: mobile health
ML: machine learning
OWASP: Open Worldwide Application Security Project
RAG: retrieval-augmented generation
SMOTE: synthetic minority oversampling technique
STI: sexually transmitted infection
SUS: System Usability Scale
UCI: University of California, Irvine
WHO: World Health Organization

Edited by M Balcarras; submitted 06.Jun.2026; peer-reviewed by T Coffin, M Honarmand; comments to author 13.Aug.2026; revised version received 20.Sep.2026; accepted 09.Oct.2026; published 06.Nov.2026

Please cite as:
Olorunfemi T, Rashid S, Khashayar P
A Multilingual AI-Driven Conversational System for Cervical Cancer Awareness and Risk Assessment: Development and Evaluation Study
JMIR Cancer 2026;12:e103866
URL: https://cancer.jmir.org/2026/1/e103866
doi: 10.2196/103866
PMID:

©Temitope Olorunfemi, Shafat Rashid, Patricia Khashayar. Originally published in JMIR Cancer (https://cancer.jmir.org), 06.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Cancer, is properly cited. The complete bibliographic information, a link to the original publication on https://cancer.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX