iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82437

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Uterine Cancer Detection Using Natural Language Processing: A Clinical Text Mining Approach

Priya Motghare, Deepa Deulkar
P. R. Pote Patil College of Engineering and Management, Amravati

Abstract: ***Uterine cancer, particularly endometrial carcinoma, is among the most prevalent gynecological malignancies affecting women globally. Early-stage diagnosis significantly improves treatment outcomes; however, much of the relevant clinical evidence exists in unstructured textual formats such as pathology reports, radiology interpretations, and physician notes. These narratives often contain critical yet underutilized diagnostic information.***
This research investigates the application of Natural Language Processing (NLP) techniques for identifying uterine cancer from clinical text data. The study explores a progression of methods, ranging from traditional machine learning approaches to advanced transformer-based architectures tailored for biomedical language. Key aspects such as domain adaptation, model interpretability, and integration into clinical workflows are examined. Furthermore, the paper highlights existing limitations and outlines future research directions for advancing this domain.
Keywords: Uterine Cancer, Endometrial Carcinoma, Natural Language Processing, Clinical Text Mining, Transformer Models, Explainable AI

I. INTRODUCTION

Uterine cancer, with endometrial carcinoma as its most common subtype, has shown a steady increase in incidence across the globe. Early detection plays a decisive role in improving patient survival rates. However, valuable diagnostic clues are frequently embedded within unstructured clinical documents rather than structured datasets.

Clinical narratives—including outpatient notes, pathology findings, and radiology reports—capture detailed descriptions of symptoms, physician observations, and disease progression over time. These textual records are rich in information but remain challenging to analyze using conventional computational techniques.

The emergence of Natural Language Processing (NLP), particularly deep learning and transformer-based models, has enabled significant advancements in extracting meaningful insights from such data. Despite this progress, most existing research emphasizes general cancer detection, with limited focus on uterine cancer-specific characteristics.

This study proposes a specialized NLP framework designed to detect uterine cancer from clinical text, emphasizing accuracy, interpretability, and real-world clinical relevance.

Key Contributions:

- Development of a uterine cancer-specific NLP framework for clinical text analysis
- Evaluation of domain-adapted transformer models for gynecological data
- Identification of research gaps and recommendations for future clinical implementation

II. CLINICAL TEXT DATA SOURCES

The effectiveness of NLP-based detection systems largely depends on the diversity and quality of input data. This study considers multiple types of clinical text sources, including:

- Gynecology outpatient and inpatient records
- Histopathology and biopsy reports
- Radiology interpretations (e.g., ultrasound, MRI, CT scans)
- Surgical notes and discharge summaries
- Menstrual and hormonal history documentation

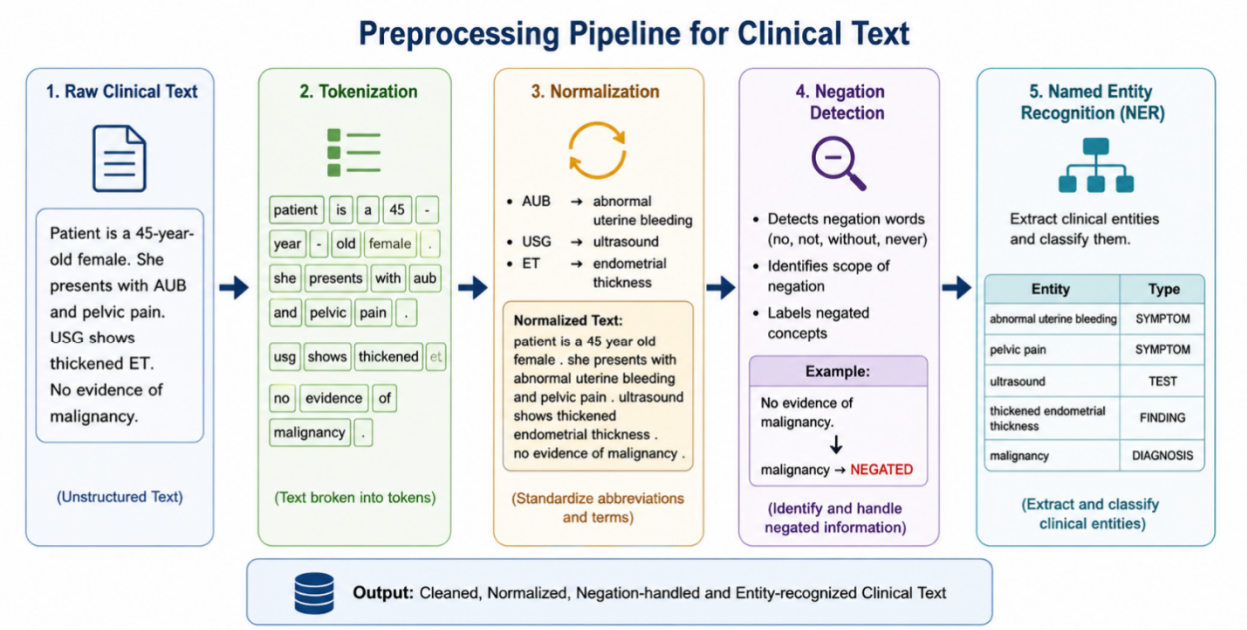


Fig. 2. Preprocessing pipeline for clinical text.

These sources contain essential indicators such as tumor classification, depth of invasion, histological features, and staging information, which are critical for accurate diagnosis.

III. CHALLENGES IN CLINICAL TEXT PROCESSING

Processing medical text presents unique challenges that differ significantly from general language tasks. Important preprocessing steps include:

- Tokenization and sentence boundary detection adapted for medical text
- Expansion and normalization of abbreviations (e.g., AUB, ET)
- Detection of negation and uncertainty (e.g., “no signs of malignancy”)
- Identification of domain-specific entities using Named Entity Recognition (NER)
- Linking extracted terms to standardized medical ontologies such as UMLS and ICD

Additionally, clinical text often includes ambiguous language, shorthand notations, and implicit reasoning, making automated interpretation complex and error-prone.

IV. RESEARCH APPLICATIONS AND USE CASES

A. Extraction of Diagnostic Features

Clinical reports frequently include detailed descriptions of pathological findings that are not available in structured databases. NLP systems can automatically identify and extract critical features such as tumor grade, invasion depth, and lymphovascular involvement.

B. Identification of Missed Follow-ups

Patients with abnormal findings may sometimes be overlooked due to fragmented documentation. NLP techniques can help identify such cases by scanning clinical notes and generating alerts for further evaluation, thereby reducing the risk of delayed diagnosis.

C. Symptom Detection and Clinical Phenotyping

Early symptoms like abnormal uterine bleeding are often recorded in narrative form rather than coded diagnoses. NLP models can detect these symptoms and differentiate between their presence and absence using contextual understanding and embedding-based techniques.

V. NLP METHODOLOGIES

A. *Traditional Approaches*

Initial efforts in clinical text analysis relied on feature-based methods such as Bag-of-Words and TF-IDF combined with classifiers like Support Vector Machines and Naïve Bayes. While interpretable, these methods have limited capability in capturing complex semantic relationships.

B. *Neural Network Models*

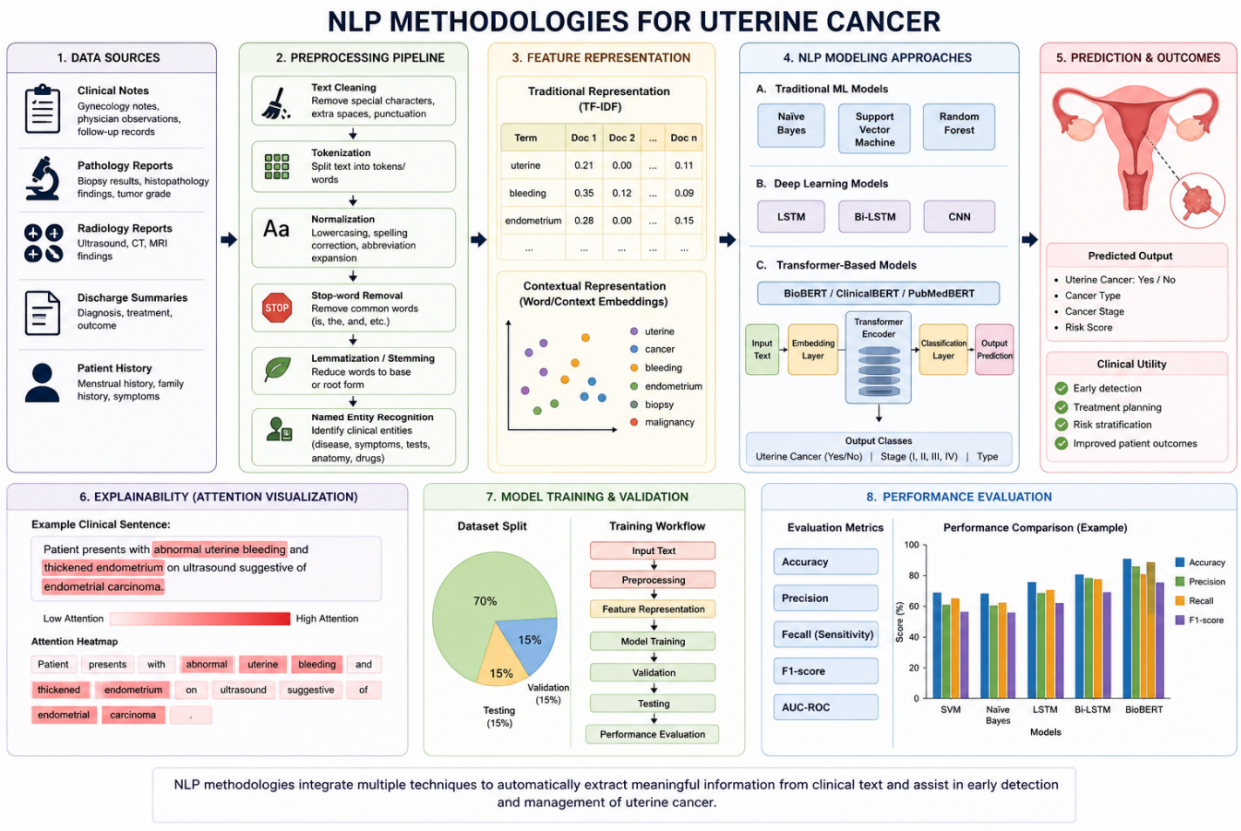
The introduction of word embeddings, including Word2Vec and FastText, improved semantic representation. Sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks enabled analysis of temporal patterns but struggled with long and complex documents.

C. *Transformer-Based Models*

Recent advancements in transformer architectures, including BioBERT, ClinicalBERT, and PubMedBERT, have significantly improved performance in biomedical NLP tasks. These models effectively capture contextual relationships within text and are well-suited for clinical applications.

Emerging research directions include:

- Domain-specific pretraining on gynecological datasets
- Hierarchical models for processing lengthy clinical documents
- Multi-task learning for simultaneous detection and classification
- Continuous learning to adapt to evolving medical terminology



VI. PROPOSED FRAMEWORK

A. *Data Collection and Annotation*

Clinical text data is gathered from healthcare institutions and annotated by medical experts. Labels include cancer presence, subtype classification, and disease stage. Semi-supervised techniques may be used to reduce annotation effort.

B. Preprocessing and Feature Extraction

The preprocessing pipeline incorporates normalization, negation detection, and entity linking. Both contextual embeddings and ontology-driven features are utilized to enhance model performance.

C. Model Design

Transformer-based architectures are fine-tuned for the classification task. Hierarchical attention mechanisms are employed to manage long documents, while multi-task learning enables simultaneous prediction of multiple clinical attributes.

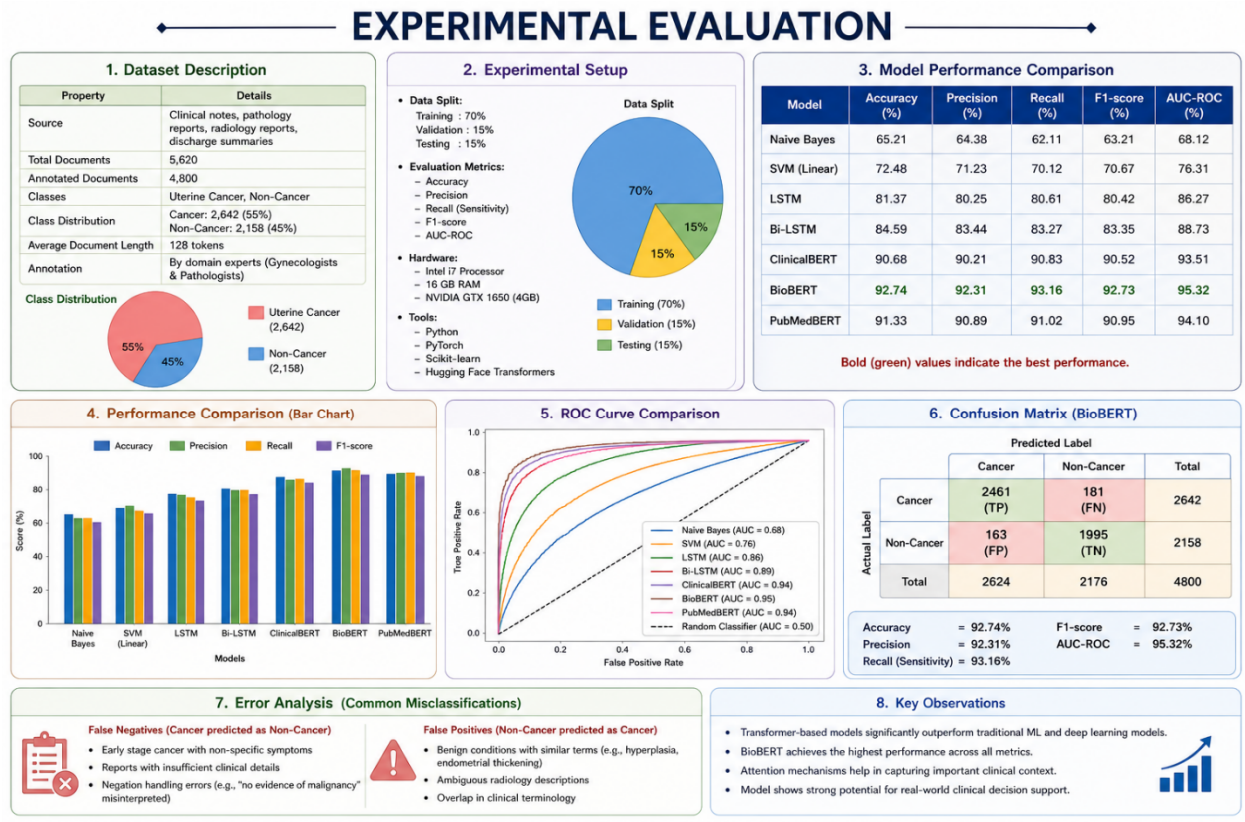
D. Explainability

To ensure transparency, explainability techniques such as attention visualization and feature attribution are integrated. These methods highlight important terms influencing model predictions, improving trust among clinicians.

VII. EXPERIMENTAL EVALUATION

Model performance is assessed using standard evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Given the critical nature of medical diagnosis, recall is prioritized to minimize missed cases.

The study employs cross-validation and external dataset testing to evaluate generalizability. Comparisons are made against baseline machine learning models and non-specialized NLP approaches.



VIII. RESULTS AND DISCUSSION

The findings indicate that transformer-based models significantly outperform traditional approaches in detecting uterine cancer from clinical text. Improvements are particularly evident in recall and F1-score, demonstrating enhanced sensitivity to early-stage indicators.

Analysis of model outputs reveals that attention mechanisms effectively identify clinically relevant features such as abnormal bleeding patterns, imaging observations, and biopsy results.

However, challenges related to dataset limitations, annotation inconsistencies, and institutional variability remain areas of concern.

IX. LIMITATIONS AND FUTURE DIRECTIONS

This study is constrained by the limited availability of large, annotated datasets specific to uterine cancer. Additionally, variations in documentation practices across institutions may affect model consistency.

Future research should explore:

- Integration of multimodal data (text, imaging, pathology)
- Federated learning approaches for secure data sharing
- Prospective validation in real clinical environments
- Enhanced interpretability methods for clinical adoption

X. CONCLUSION

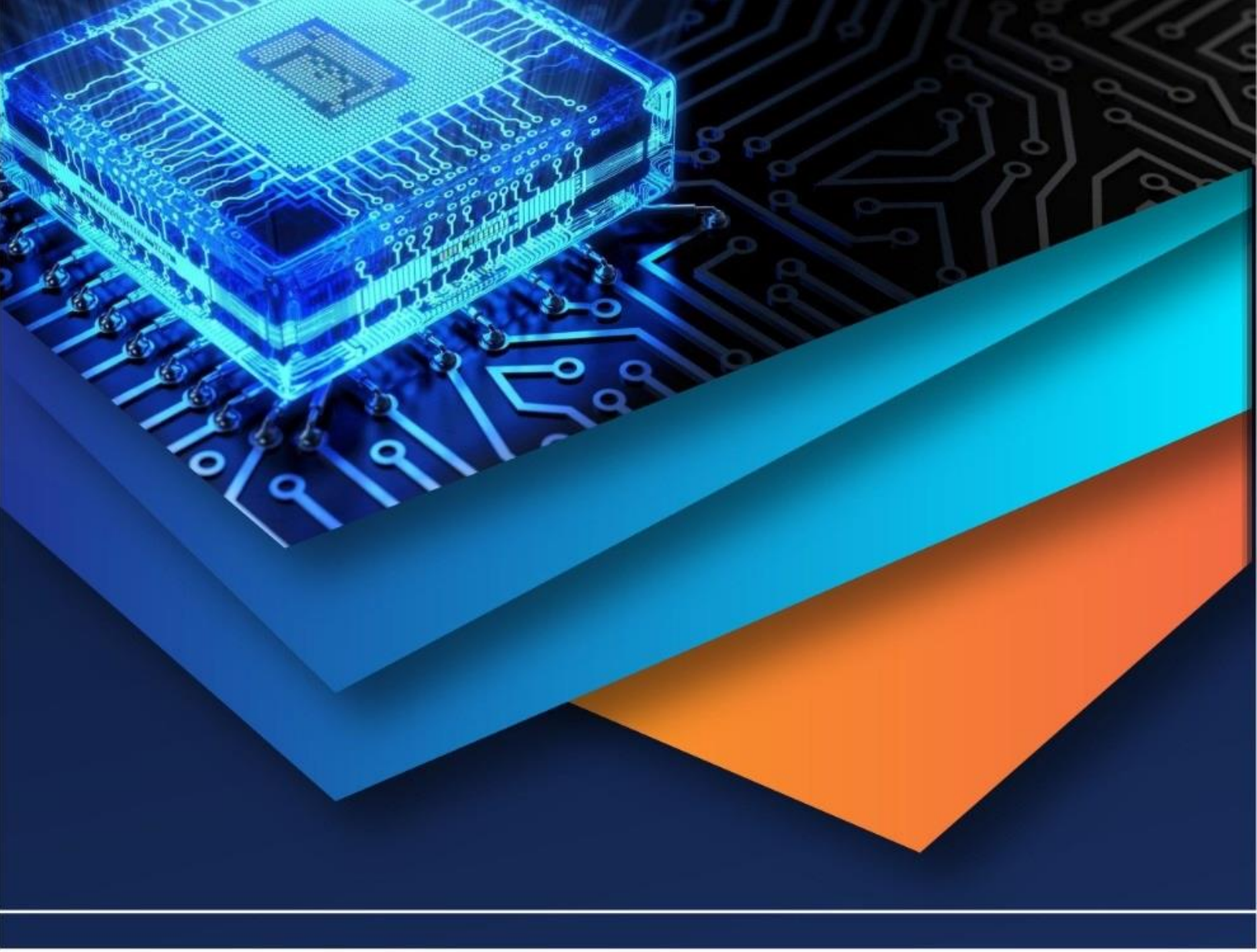
This research demonstrates the potential of Natural Language Processing to transform uterine cancer detection by leveraging unstructured clinical data. Advanced transformer-based models, combined with domain-specific adaptations, offer promising improvements in early diagnosis and clinical decision support.

Addressing current limitations and ensuring seamless integration into healthcare systems will be essential for translating these advancements into practical clinical solutions.

REFERENCES

[1] Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 74, 229–263 (2024).
[2] Dhawan, S., Singh, K., & Arora, M. (2021). Cervix Image Classification for Prognosis of Cervical Cancer using Deep Neural Network with Transfer Learning. EAI Endorsed Transactions on Pervasive Health and Technology, 7(27), e5.
[3] Hiam Alquran, Wan Azani Mustafa, Isam Abu Qasmieh, Yasmeen MohdYacob, Mohammed Alsalatie, Yazan Al-Issa, Ali Mohammad Alqudah, (2022), 'Cervical Cancer Classification Using Combined Machine Learning and Deep Learning Approach', Computers,Materials& Continua, 72(3):5117-5134.
[4] OmneyaAttallah, (2023), 'Cervical Cancer Diagnosis Based on Multi-Domain Features Using Deep Learning Enhanced by Handcrafted Descriptors', Appl. Sci, 13(3):1-23
[5] Abinaya K, Sivakumar B, (2024), 'A Deep Learning-Based Approach for Cervical Cancer Classification Using 3D CNN and Vision Transformer', J Imaging Inform Med, 37(1):280 296.
[6] Sher Lyn Tan, GaneshsreeSelvachandran, Weiping Ding, Raveendran Paramesran, Ketan Kotecha, (2023), 'Cervical Cancer Classification from Pap Smear Images Using Deep Convolutional Neural Network Models', Interdisciplinary Sciences: Computational Life Sciences, 16:16–38.
[7] Jesse Jeremiah Tanimu, Mohamed Hamada, Mohammed Hassan, HabeebahKakudi, John Oladunjoye Abiodun, (2022), 'A Machine Learning Method for Classification of Cervical Cancer', electronics, 11(3):1-23.
[8] Ashok, B., Aruna, P., 2016. Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier. Int. J. Eng. Res. Afr. 6, 94e99.
[9] Asadi, F., Salehnasab, C., Ajori, L., 2020. Supervised algorithms of machine learning for the prediction of cervical cancer. J Biomed Phys Eng 10, 513.
[10] Diniz, D.N., Rezende, M.T., Bianchi, A.G.C., Carneiro, C.M., Luz, E.J.S., Moreira, G.J.P., et al., 2021 Jul 9. A deep learning ensemble method to assist cytopathologists in pap test image classification. J. Imaging 7 (7), 111.
[11] Nithya, B., Ilango, V., 2019. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Appl. Sci. 1, 1e16.
[12] W. Książek, M. Hammad, P. Pławiak, U. R. Acharya, and R. Tadeusiewicz, “Development of novel ensemble model using stacking learning and evolutionary computation tech niques for automated hepatocellular carcinoma detection,” Biocybernetics and Biomedical Engineering, vol. 40, no. 4, pp. 1512–1524, 2020.
[13] B. J. Cho, Y. J. Choi, M. J. Lee et al., “Classification of cervical neoplasms on colposcopic photography using deep learning,” Scientific Reports, vol. 10, no. 1, p. 13652, 2020.
[14] H. Zhang, C. Chen, R. Gao et al., “Rapid identification of cervical adenocarcinoma and cervical squamous cell carcinoma tissue based on Raman spectroscopy combined with multiple machine learning algorithms,” Photodiagnosis and Photodynamic Therapy, vol. 33, p. 102104, 2021.
[15] Kuruvilla A, Jayanthi B. Analysis and review on feature selection and classification methods on cervical cancer. Ictact J Soft Comput 2022;12(2):2551-8..
[16] CH N, Sai PP, Madhuri G, Reddy KS, Simha B, Reddy DV. Artificial Intelligence based Cervical Cancer Risk Prediction Using M1 Algorithms. 2022 Int Conf Emerg Smart Comput Informatics 2022 Mar;1–6. doi: 10.1109/ ESCI53509.2022.9758241.
[17] Ali MM, Ahmed K, Bui FM, Paul BK, Ibrahim SM, Quinn JMW, et al. Machine learning-based statistical analy sis for early stage detection of cervical cancer. Comput Biol Med 2021 Dec;139:104985. doi: 10.1016/j.comp biomed.2021.104985..
[18] Chaudhuri AK, Ray A, Banerjee DK, Das A. A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. Int J Intell Syst Appl 2021 Oct 8;13(5):46–63.
[19] Peng G, Dong H, Liang T, Li L, Liu J. Diagnosis of cervical precancerous lesions based on multimodal feature changes. Comput Biol Med 2021;130:104209.
[20] Chandran V, Sumithra MG, Karthick A, George T, Deivakani M, Elakkiya B, et al. Diagnosis of Cervical Cancer based on Ensemble Deep Learning Network using Colposcopy Images. Biomed Res Int 2021;2021:5584004. doi: 10.1155/2021/5584004.
[21] 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;71:209-49.
[22] Sravani, A.B.; Ghate, V.; Lewis, S. Human papillomavirus infection, cervical cancer and the less explored role of trace elements. Biol. Trace Element Res. 2022, 1–25. https://doi.org/10.1007/s12011-022-03226-2.

[23] Dang, Amit, D. Dimple and B. N. Vallish. 2023. "Extent of Use of Artificial Intelligence and Machine Learning Protocols in Cancer Diagnosis: A Scoping Review." Indian, J. Medical Res, 157: 11-22.
[24] Razzak, M.A., M.N. Islam, M.S. Aadeeb and T. Tasnim, 2023. Digital health interventions for cervical cancer care: A systematic review and future research opportunities. PLOS ONE, Vol. 18 .10.1371/journal.pone.0296015.
[25] Andrade, P. and S. Commuri, 2023. A portable system for screening of cervical cancer.Doctoraldissertation..University of Nevada – Reno).
[26] 42. Vargas-Cardona, H.D., M. Rodriguez-Lopez, M. Arrivillaga, C. Vergara-Sanchez, J.P. García-Cifuentes, P.C. Bermúdez and A. Jaramillo-Botero, 2023. Artificial intelligence for cervical cancer screening: Scoping review, 2009–2022. Int. J. Gynecol. & Obstet., 165: 566-578



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY