A Systematic Review of AI Techniques in Cervical Cancer Detection

Introduction

Cancer is the most fatal disease discovered in the history of medical science. It is the primary factor for premature death on a global scale(Bray et al., 2021). Of different forms of cancer, Cervical Cancer stands the fourth major form of cancer with 604,000 incidents and 342,000 mortalities globally (WHO, 2021). The burden of this dreadful illness is staggeringly high in low- and middle-income nations (Sung et al., 2021). The risk of cervical cancer is 6-fold higher in women with HIV. 20% of children who lose their mothers to cancer do so due to cervical cancer(WHO, 2023).

Human Papillomavirus, a common infection, infects most of the sexually active population at some point in their lives. The infection is normally dormant, rarely exhibits any symptoms and is normally cleared by the immune system of the body(CDC, 2022). However, persistent HPV infections of the cervices (birth canal) combined with disparities in access to right screening technologies and preventive vaccination account to 95% of cervical cancer cases. (WHO, 2023)

This review study primarily revolves around the dataset of UCI Machine learning repository. By comparing the performance of different algorithms against this data set, and with citations of other similar works in the area, this review will fulfill 3 purposes. First, to understand various aspects of AI based diagnostic methodologies of cervical cancer. Second, the paper will evaluate performance of different AI techniques being developed and tested for cervical cancer detection. Third, the paper will try to identify the gap (limitation) of the technological system, if any.

Logistic Regression: a simple and powerful statistical technique for linearly separable cases.

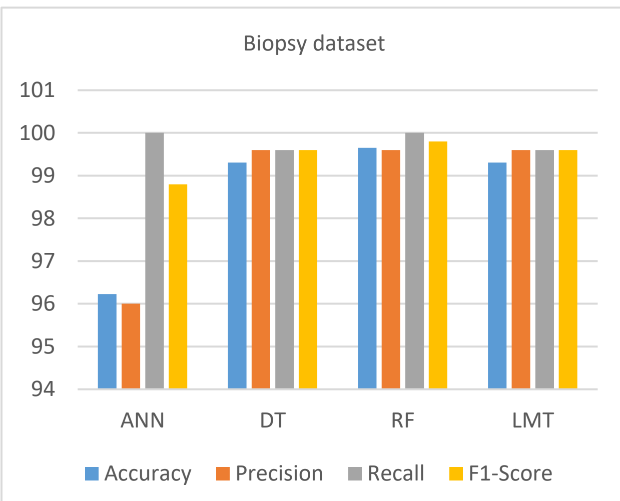
Logistic Regression is a statistical method used to predict the probability of occurrence of a particular outcome based on the given set of input, making it suitable for binary classification tasks such as cancer detection. It provides a simpler approach compared to multi-expert setting(Singh et al., 2019). This supervised learning technique has been used to create a model that identifies the predicting factors of Pap cervical cancer screening. Logistic regression models showed an accuracy of 84% to 89% in cervical cancer prediction with a sensitivity of up to 94% while analyzing 298 sets of data from cervical screening(Devi et al., 2023). Another study investigated the data of 260 patients diagnosed with reoccurrence of cervical cancer. This pool of 260 data was filtered from 4913 patients' data who were diagnosed with cervical cancer and were under checkup for possible recurrences. A total of 5 different models analyzed the data over 27 data attributes. Logistic Regression technique out-performed other models with the prediction accuracy of 96.3% for reoccurrences(Geeitha et al., 2023). A significant aspect of commonality in both the research works is that the dataset size is rather small. Logistic regression technique shows superior performance in terms of sensitivity and specificity, two of the very crucial metrics in medical diagnosis(Long et al., 1993). Despite showing significant results in linearly (or curvy linear) separable cases, Logistic regression lacks interpretability making it challenging to understand the reasons behind the predictions thus made(Rahimi et al., 2023). Logistics regression cannot effectively handle outlier datapoints and its performance depends upon independence of observations(Guedes, n.d.). The performance of logistic regression tends to decrease as the size of the dataset, in consideration, grows(Work et al., 1989). Another core dependency of Logistic Regression model is with labelled datasets. A comparative study of Machine Learning (ML) algorithms to predict cervical cancer revealed that despite having the best recall of 68.37%, logistic regression was outperformed by other ML techniques when it came to imbalanced data(Ortiz-Torres et al., 2023).

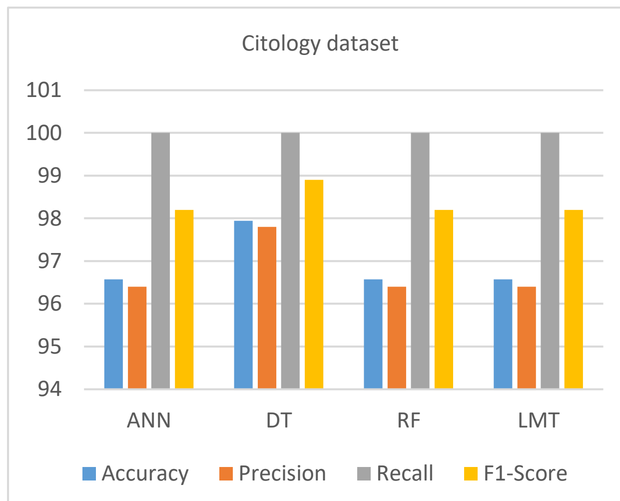
Decision Trees: A highly interpretable approach suitable for outliers and sparse datasets.

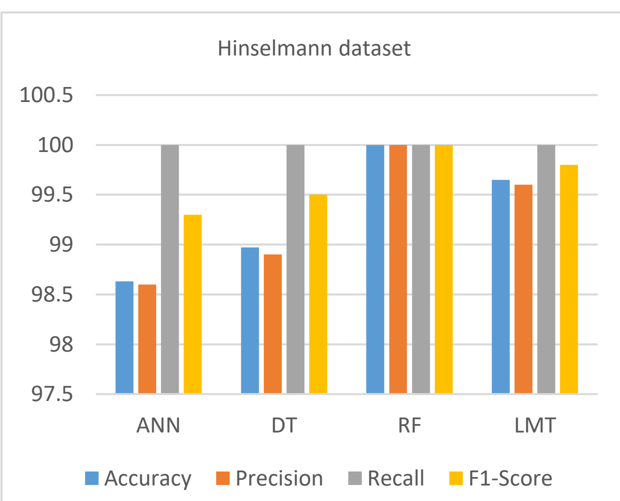
Key performance issues of logistic regression model with respect to outlier data as well as the issue of interpretability is mitigated to some extent by Decision tree approach. It implements hierarchical structure of decisions. A data, available in UCI machine learning repository, from Hospital Universitario de Caracas, Venezuela was used to create a prediction model. The dataset contains medical and behavioral records of 858 patients. The issue of imbalanced dataset was addressed using the SMOTE oversampling technique. This research used Recursive Feature Elimination approach for feature selection to enhance efficiency and measured the performance against accuracy, precision, recall and F1-score. Decision tree approach achieved the maximum performance with the measure of accuracy and recall of 98% and 80% respectively. The model had an F1-score of 95%(Cosar Sogukkuyu and Ata, 2022). Another independent study that sourced the same data from Kaggle compared performance of various machine learning technique. The performance was measured against the results of accuracy, recall, precision combined with F1-score, root mean square error value and relative absolute error value. Decision tree outperformed all other algorithms in terms of all metrics of evaluation with 96.9% accurate prediction results, while standing second to Random Forest in terms of RMSE value(Srivastav et al., 2023). The RMSE value for decision tree in this research was 0.15. Decision Tree mitigates the issues of imbalanced dataset or small sample size through data overfitting but in this process, the model has chances of attaining overly sophisticated structure(Long et al., 1993). This marks a rather serious question on its usability on larger scale datasets. The quality and consistency of results will highly depend on the capacity to create just the right level of pruning of the nodes. Despite being robust regarding outlier data sets, Decision Trees tend to exhibit limited prediction beyond the response variable limits of the training data(R, 2015).

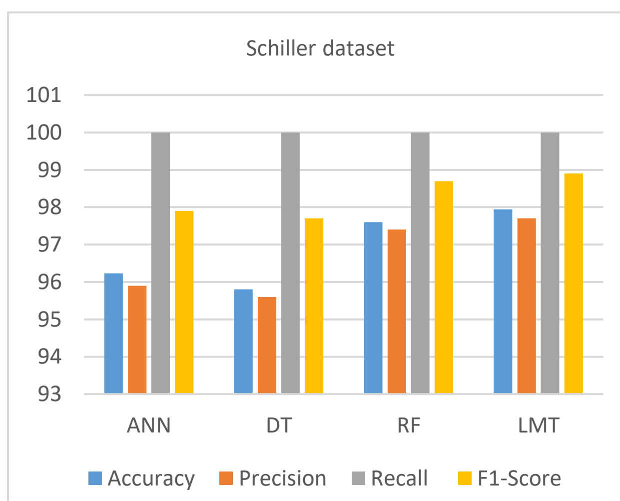
Random Forest: A model to create accurate predictions on larger datasets.

Random Forest is an ensemble to decision tree approach that accumulates several randomized decision trees and combines their predictions through a process called averaging(Biau and Scornet, 2016). Since its development in 2001, this method has received tremendous attention and adoption for its simplicity and improved accuracy. Random Forest approach could be seen as a mere extension to Decision trees with the sole purpose of improving accuracy while keeping the issue of data overfitting and other limitations in check. Random forest has already proven to be effective in early disease detection of 8 chronic diseases including diabetes, cancer and heart diseases(Yadav et al., 2023). Furthermore, Random Forest could predict Stroke related mortality based on the behavioral analysis of 9476 abstract samples in Japan, with an accuracy result of 89.94% (Yadav et al., 2023). An empirical analysis of 8 machine learning algorithms, including logistic regression and random forest, to predict 4 major cancers: lung, liver, breast and cervical showed that random forest performed best in predicting two forms of cancer, liver and cervical. The respective accuracy of 73.73% and 99.22% was observed(Mangla et al., 2023). Although authors suggest that the real-world application of the model might need additional research and exploration, this survey highlights the exceptional predictive accuracy of the random forest. An independent study, on the dataset from UCI Machine Learning repository, used 4 different classification algorithms to analyze 32 attributes against 4 target classes. The 4 target classes were namely Biopsy, Citology, Hinselmann and Schiller. Random Forest was seen to perform consistently better across each of these 4 target classes(-Ul-Islam et al., 2019). The precise calculative result of the performance of these 4 algorithms: artificial neural network , decision tree, random forest and logistic model tree is presented below:









Despite achieving a reasonable accuracy for the proposed model, the performance stats are close among different classification methods. Another independent study of the same data using 5 different ML algorithms including random forest was done in the same year. The technique had implemented Boruta algorithm and Simulated Annealing to refine the feature selection process of the model. Random Forest approach could make 100% accurate predictions with AUC of 0.91 (Nithya and Ilango, 2019). The reason for increased efficiency in this research could be because of the implementation of advanced features to generate optimal feature subsets. But again, achieving 100 % accuracy is a big challenge for an AI based system. The report acknowledges that assumptions such as factorization and variable replacement were made during the data processing stage. This highlights a need for further study and a thorough exploration before considering its real-world application. A different study in the same data set discussed above evaluated the outcomes of decision tree and random forest in prediction of cervical cancer. Random forest technique outperformed the other approach with a maximum precision of 98% (and a recall of 90%)(Cosar Sogukkuyu and Ata, 2022). Decision tree had a higher recall (of 95%), though. Similarly, in another study of 4 machine learning algorithms including random forest approach and logistic regression, the prediction accuracy of former outperformed the latter. The accuracy of the random forest was 96.49% and that of decision tree was 95.94%(Rawat et al., 2023). Despite promising results, the study accepted the possible limitation of datasets in representation of certain subtypes of cervical cancer and acknowledges the need for further research to improve on the model. Random forest technique has been found to perform exceptionally well even with oversampled unscaled and scaled data sets(Chauhan and Singh, 2022). Consistent and highly precise accuracy has been thoroughly validated for this technique. This considerable leap of accuracy over larger data sets often comes with sizable computational overhead and implementation of such should be purely context and problem driven. A reasonable justification for such consistently improved accuracy over larger data sets is because of the implementation of the ensemble learning approach.

Ensemble Approach: For improved accuracy, robustness, interpretability and versatility

Most machine learning techniques that produce only one hypothesis generally suffer either (if not all) of the three issues: the issue of representation, the computational and the statistical issues, thereby creating variances and biases in the result (Dietterich TG, 2002).Ensemble learning is a machine learning paradigm that involves combining predictions from multiple models to achieve better results in terms of accuracy, prediction and the overall performance of the system. Ensemble learning algorithms work by running a “base learning algorithm” repeatedly and forming a vote out, to devise a conclusion, using two principal approaches called ‘bagging’ and ‘boosting’(Dietterich TG, 2002). Random Forest, eXtreme Gradient Boosting (XG Boost), Gradient Boosting Machines, LightGBM and Stacking are some of them. Ensemble approach-based models have proven delivery in terms of versatility and stability. Various research and studies have shown concurrencies to this statement. The citations to all studies and research works made in the discussion of Random Forest above would naturally qualify for consideration in this context as well. Random forest technique is even, at times, superior, in terms of performance, compared to other ensemble approaches such as LightGBM and AdaBoost (Roopa, 2023). The empirical analysis report of 8 ML algorithms, also discussed in the context of random forest approach above, revealed one more ensemble learning based approach that had a levelling performance to random forest. XG Boost, a yet another ensemble learning model, also performed best with the exact accuracy of 99.22%(Mangla et al., 2023). An independent study, carried out on the UCI machine learning repository dataset, shows us the extent of capabilities that ensemble learning approach possesses as a classifier model. An ensemble model was created using 4 different classifier algorithms. The performance of the ensemble model was compared with that of the individual algorithm against identical dataset. The evaluation metrics for this evaluation included accuracy, recall and precision along with the F1-score. The ensemble approach was found to perform better than each of the comprising algorithms with an accuracy of 95.6% and superior results in specificity and sensitivity (Avanija et al., 2022). The outcome of this study is crucial because most research works on ensemble approach focus on working with decision trees (to create random forest). Now, because of how decision trees are constructed, the very compositional architecture cascades one single occurrence of wrong decision, if any, throughout the entire system, thereby creating higher variance in the result of the model(Dietterich TG, 2002) . While ensemble methods can address many challenges in machine learning, practitioners should carefully consider the trade-offs such as computational complexity, interpretability, biasness and sensitivity to outliers. It is important to note that the effectiveness of ensemble approach heavily relies on the choice of base classifiers(Avanija et al., 2022). In-depth knowledge about the characteristics of data set and a thorough understanding of the problem in hand play a pivotal role in creating the most effective ensemble model for predicting cervical cancer.

Deep Learning for image-based large-scale screening for early detection of cervical cancer

The search for a more accurate and consistently precise model for cervical cancer detection takes us to the realm of deep learning. A typical machine learning model undergoes a 4-stage process of data harmonization, representation, fitting and evaluation. These techniques often exhibit limitation in their ability to process raw data(Miotto et al., 2017). The successive advancement of technology and the degree of integration between multiple heterogeneous components in medical industry have positively fueled the availability of biomedical data on a large scale. This brings opportunities as well as challenges. Deep learning incorporates computational models composed of multiple processing layers that possess capacity to represent raw data with multiple layers of abstraction. This enables the model to directly learn from raw, abstract data in the form of representations. With this feat, Deep Learning is making significant progress in solving problems that conventional machine learning techniques were struggling with for years (Lecun et al., 2015). Research has shown high efficiency of various Deep Learning approaches in detecting cervical cancer(Alias et al., 2023). Among the prevalent deep learning models, single and multi-layered convolution neural networks happen to be used most extensively(Kavyashree et al., 2021a). Various studies and research works have shown the exceptional performance of convolution neural networks(Angara et al., 2021; Youneszade et al., 2023) with the achieved accuracy of 98%(Kavyashree et al., 2021b). A study on the UCI dataset using CNN models was able to perform with an accuracy of 97.67%(Sundarambal et al., 2022). Convolution neural networks combined with bidirectional Long-Short term memory could analyze colposcopy images to yield 98.3% accurate and 99% precise prediction results(Zhang et al., 2020). The precision accuracy and the overall performance of a convolution neural networks (CNN) can be improved by creating multiple channels into them. A comparative study of performance of 3 different models of CNN with varying channel depths concluded that CNN model with highest number of channels in the convolution layer achieves leading performance (Chauhan and Singh, 2021). The precision and accuracy of the system to classify whole slide images was tuned up to be as high as 96.89% and 93.38%. The high computational efficiency of deep learning approach demands a significant data size, is time consuming and resource intensive(Ahmed et al., 2023). We are yet to create a model that can be well replicated across diverse domains or be generalized for larger application area(Wang and Zhang, 2022).

Conclusion

Numerous research and study works have experimentally and clinically proven the strength of AI technology in the health sector. Different models and classifiers aid in the effective analysis of patient data and clinical test reports including images to make early prediction of cancer. Reports have shown that half of the cancer fatalities could be eliminated with early detection. Machine learning techniques provide direct and easy to implement models while deep learning adds large data sets handling capabilities. This collective potential of Artificial Intelligence can aid in creating effective health care systems at reduced cost per patient service. However, the world is yet to witness a wide scale implementation of such a system for diagnosis of cancer diseases. In course of the research for this review paper, in the models being suggested, proposed and tested, two distinct issues were noticed. First was the issue of *data availability*. Health care systems employ human experts to assess data and readings from heterogeneous measuring entities to diagnose (and predict) cancer. Most of these records have issues of universal representation and completeness, thus making it difficult for an intelligent system to read from and analyze further. The second issue noted was that of *data computation capabilities.* The AI community is yet to innovate a model that creates highly accurate systems at reasonable computational cost for large data sets. Some deep learning techniques, although highly accurate predictive systems, could simply fail to work if the dataset being analyzed is below a certain threshold. Although deep learning approaches offer techniques such as transfer learning, adversarial training and data augmentation, a detailed study of these could not be covered in this scope of study. Having pointed out these issues, this review acknowledges a need for a systematic and in-depth review of deep learning approaches to identify better models. A more generalized model which could be implemented irrespective of topological and demographic constraints, and that works with smaller as well as larger datasets with consistent results, is yet to be identified.

References

Ahmed, S.F., Alam, M.S. Bin, Hassan, M., Rozbu, M.R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A.B.M., Gandomi, A.H., 2023. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif Intell Rev 56. https://doi.org/10.1007/s10462-023-10466-8

Alias, N.A., Mustafa, W.A., Jamlos, M.A., Ismail, S., Alquran, H., Rohani, M.N.K.H., 2023. Pap Smear Image Analysis Based on Nucleus Segmentation and Deep Learning – A Recent Review. Journal of Advanced Research in Applied Sciences and Engineering Technology 29, 37–47. https://doi.org/10.37934/araset.29.3.3747

Angara, S., Guo, P., Xue, Z., Antani, S., 2021. Semi-supervised learning for cervical precancer detection, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems. Institute of Electrical and Electronics Engineers Inc., pp. 202–206. https://doi.org/10.1109/CBMS52027.2021.00072

Avanija, J., Sunitha, G., Madhavi, K.R., Chakravarthi, S.S., Sai Vittal, R.H., 2022. Prediction and Analysis of Cervical Cancer: An Ensemble Approach, in: 4th International Conference on Inventive Research in Computing Applications, ICIRCA 2022 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 1734–1739. https://doi.org/10.1109/ICIRCA54612.2022.9985601

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25. https://doi.org/10.1007/s11749-016-0481-7

Bray, F., Laversanne, M., Weiderpass, E., Soerjomataram, I., 2021. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer 127. https://doi.org/10.1002/cncr.33587

CDC, 2022. Genital HPV Infection – Basic Fact Sheet [WWW Document]. URL https://www.cdc.gov/std/hpv/stdfact-hpv.htm (accessed 2.15.24).

Chauhan, N.K., Singh, K., 2022. Diagnosis of Cervical Cancer with Oversampled Unscaled and Scaled Data Using Machine Learning Classifiers, in: 2022 IEEE Delhi Section Conference, DELCON 2022. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/DELCON54057.2022.9753298

Chauhan, N.K., Singh, K., 2021. Impact of Variation in Number of Channels in CNN Classification model for Cervical Cancer Detection, in: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICRITO51393.2021.9596366

Cosar Sogukkuyu, D.Y., Ata, O., 2022. Diagnosing Cervical Cancer Using Machine Learning Methods, in: HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/HORA55278.2022.9800033

Devi, S., Gaikwad, S.R., Harikrishnan, R., 2023. Prediction and Detection of Cervical Malignancy Using Machine Learning Models. Asian Pacific Journal of Cancer Prevention 24. https://doi.org/10.31557/APJCP.2023.24.4.1419

Dietterich TG, 2002. Ensemble Learning. The Handbook of Brain Theory and Neural Networks.

Geeitha, S., Renuka, P., Ramya, S., Sinduja, K., 2023. Progression of Disease Forecasting in Recurrence Cervical Cancer Using Logistic Regression, in: 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, pp. 1050–1055. https://doi.org/10.1109/ICECA58529.2023.10395205

Guedes, L.P., n.d. LIMITATIONS OF LOGISTIC REGRESSIONS TO ESTIMATE MEASURES OF ASSOCIATION FOR BINARY HEALTH OUTCOMES A HANDS-ON STUDY WITH TWO EPIDEMIOLOGICAL EXAMPLES.

Kavyashree, C., Sowmya, H.K., Reddy, N.V.U., 2021a. A Survey on the Cervical Cancer Detection using Deep Learning methods, in: 2021 International Conference on Forensics, Analytics, Big Data, Security, FABS 2021. https://doi.org/10.1109/FABS52071.2021.9702701

Kavyashree, C., Sowmya, H.K., Reddy, N.V.U., 2021b. A Survey on the Cervical Cancer Detection using Deep Learning methods, in: 2021 International Conference on Forensics, Analytics, Big Data, Security, FABS 2021. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/FABS52071.2021.9702701

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature. https://doi.org/10.1038/nature14539

Long, W.J., Griith, J.L., S er, H.P., D, R.B., 1993. A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain, Biomedical Research.

Mangla, S., Saini, P., Jayswal, A.K., Sanyal, K., Pal, S., 2023. An Ai Based Application for Cancer Diagnosis - An Emperical Analysis, in: Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2023. Institute of Electrical and Electronics Engineers Inc., pp. 231–236. https://doi.org/10.1109/Confluence56041.2023.10048847

Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T., 2017. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform 19, 1236–1246. https://doi.org/10.1093/bib/bbx044

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. https://doi.org/10.1007/s42452-019-0645-7

Ortiz-Torres, C., Reátegui, R., Valdiviezo-Diaz, P., Barba-Guaman, L., 2023. A Comparison of Machine Learning Algorithms to Predict Cervical Cancer on Imbalanced Data, in: Communications in Computer and Information Science. https://doi.org/10.1007/978-3-031-24985-3_9

R, P.T., 2015. A Comparative Study on Decision Tree and Random Forest Using R Tool. IJARCCE 196–199. https://doi.org/10.17148/ijarcce.2015.4142

Rahimi, M., Akbari, A., Asadi, F., Emami, H., 2023. Cervical cancer survival prediction by machine learning algorithms: a systematic review. BMC Cancer 23. https://doi.org/10.1186/s12885-023-10808-3

Rawat, P., Bajaj, M., Mehta, S., Sharma, V., Vats, S., 2023. A Study on Cervical Cancer Prediction using Various Machine Learning Approaches, in: International Conference on Innovative Data Communication Technologies and Application, ICIDCA 2023 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 1101–1107. https://doi.org/10.1109/ICIDCA56705.2023.10099493

Roopa, C., 2023. Detection of cervical cancer using Ensemble Approach in Machine Learning, in: 2023 International Conference on Computer Communication and Informatics, ICCCI 2023. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCCI56745.2023.10128521

Singh, S., Panday, S., Panday, M., Rautaray, S.S., 2019. Logistic Regression for the Diagnosis of Cervical Cancer, in: Data Engineering and Applications. pp. 109–117.

Srivastav, S., Guleria, K., Sharma, S., 2023. Predictive Machine Learning Approaches for Cervical Cancer Detection: An Analytical Comparison, in: Proceedings of the 17th INDIACom; 2023 10th International Conference on Computing for Sustainable Global Development, INDIACom 2023.

Sundarambal, B., Karthikeyini, C., Bommi, R.M., Subramanian, S., Jacintha, V., 2022. Comparison of Machine Learning and Deep Learning models for Cervical Cancer Prediction, in: ICDCS 2022 - 2022 6th International Conference on Devices, Circuits and Systems. Institute of Electrical and Electronics Engineers Inc., pp. 135–138. https://doi.org/10.1109/ICDCS54290.2022.9780705

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71. https://doi.org/10.3322/caac.21660

-Ul-Islam, A., H Ripon, S., Qaisar Bhuiyan, N., 2019. Cervical Cancer Risk Factors: Classification and Mining Associations. APTIKOM Journal on Computer Science and Information Technologies 4, 8–18. https://doi.org/10.11591/aptikom.j.csit.131

Wang, S.-H., Zhang, Y.-D., 2022. Advances and Challenges of Deep Learning. Recent Patents on Engineering 17. https://doi.org/10.2174/1872212116666220530125230

WHO, 2023. Cervical Cancer Fact Sheet.

WHO, 2021. Cervical Cancer .

Work, J.W., Ferguson, J.G., Diamond, G.A., 1989. Limitations of a conventional logistic regression model based on left ventricular ejection fraction in predicting coronary events after myocardial infarction. Am J Cardiol 64. https://doi.org/10.1016/0002-9149(89)90751-0

Yadav, A., Kumar, D., Hasija, Y., 2023. Behaviour Analysis Using Machine Learning Algorithms in Health Care Sector, in: 2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023. Institute of Electrical and Electronics Engineers Inc., pp. 877–880. https://doi.org/10.1109/InCACCT57535.2023.10141829

Youneszade, N., Marjani, M., Pei, C.P., 2023. Deep Learning in Cervical Cancer Diagnosis: Architecture, Opportunities, and Open Research Challenges. IEEE Access 11, 6133–6149. https://doi.org/10.1109/ACCESS.2023.3235833

Zhang, Y., Li, L., Gu, J., Wen, T., Xu, Q., 2020. Cervical Precancerous Lesion Detection Based on Deep Learning of Colposcopy Images. J Med Imaging Health Inform 10. https://doi.org/10.1166/jmihi.2020.3051