Cervical Cancer Detection: Logistic Regression vs. Neural Network

Malliboina Guru Dinesh and S. Dinakar Raj
Saveetha University, Chennai, Tamil Nadu, 602105, India

Keywords: Logistic Regression Algorithm, Artificial Neural Network Algorithm, Cervical Cancer, HPV, Analysis, Disease, Significance, Health.

Abstract: The research aims to investigate and diagnose cervical cancer conditions by comparing the Logistic Regression algorithm with the Artificial Neural Network approach. Materials and Methods: The dataset for HPV is sourced from Kaggle, utilized for the screening and treatment of cervical cancer, and split into two subsets. Subset 1 employs the Logistic Regression algorithm, while Subset 2 employs the Artificial Neural Network method. Results: The Independent Sample T-test, applied to diagnose cervical cancer, yielded a significance level of p = 0.23 (which exceeds the widely accepted threshold of 0.05). When evaluating performance, the logistic regression approach displayed higher accuracy at 83.066% in contrast to the artificial neural network, which achieved an accuracy of 80.200%. Significantly, the superiority of the Logistic Regression technique over the Artificial Neural Network method is apparent. Conclusion: Within the realm of performance analysis, the Logistic Regression (LR) Algorithm outperforms the Artificial Neural Network (ANN) Algorithm.

1 INTRODUCTION

Cervical cancer originates when the normal cells within the cervix transform into cancerous cells. While this process usually takes several years, rapid changes can also occur in a shorter period. The detection of cervical cancer involves examining surface cells from the cervix through a cytological examination, commonly known as a pap smear. Annually, around 15,000 women in the United States receive a diagnosis of cervical cancer ("Automated Image Analysis within a Multispectral System for the Diagnosis of Cervical Cancer" n.d.) (AS et al 2013). Unfortunately, one woman loses her life to cervical cancer every two minutes, a disease that is largely preventable and treatable. Most of these tragic outcomes occur in low- and middle-income countries. The progression from precancerous changes to cervical cancer usually spans 10 to 20 years. These precancerous changes can be identified through a liquid-based cytology Pap test for cervical cancer screening (Jha, Gupta, and Saxena 2021). However, challenges such as socioeconomic barriers, a shortage of skilled cytopathologists, and low inter-annotator agreements contribute to the inefficient diagnosis of cervical cancer patients in low- and middle-income countries (Piyush Kumar Pareek. et al. 2022). Thus, developing automated techniques to support pathologists in screening, based on whole slide images generated by scanning glass slides, becomes crucial to reduce subjectivity and enhance productivity ("Region of Interest Identification for Cervical Cancer Images" n.d. 2021). The discussed study's topic revolves around the development of a methodology based on the analysis of fluorescence images acquired at various excitation wavelengths. Our investigation encompasses selecting informative elements in different types of images, quantitatively assessing them, estimating the effectiveness of diverse fluorescence images, and creating specialized analytical techniques for identifying pathological regions. To delve deeper into cervical tissue changes, we devised a methodology encompassing the exploration of various image types ("Cervical Cancer Prediction Based on Risk Factors Utilizing Ensemble Learning" n.d.).

In recent times, a plethora of articles and literary works have emerged in the domain of cervical cancer detection. Around 30 research papers have been published on IEEE Xplore, while Google Scholar has witnessed the publication of nearly 100 articles. In 2019, Pikala et al. introduced a machine learning approach that employs language networks mapped via fMRI to distinguish between individuals with and without health issues. Al Mudawi and Alazeb (2022) introduced a supervised machine learning classifier

Dinesh, M. and Raj, S.
Cervical Cancer Detection: Logistic Regression vs. Neural Network.
DOI: 10.5220/0012537400003739
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 25-30
ISBN: 978-989-758-661-3
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

aimed at predicting treatment outcomes for HPV infection in recently diagnosed cervical cancer patients (Vickram, A. S. et al. 2021). The pursuit of early-stage cervical cancer diagnosis has led to proposals focusing on accuracy, precision, recall, the F1-score, and the appropriate application of Logistic Regression algorithms (Song et al. 2015).

A challenge with current studies is the inability to accurately detect cervical cancer disease. This study aims to enhance the initial detection of cervical cancer in patients and improve overall accuracy by utilizing the Logistic Regression algorithm. A comparison with the technique involving the Artificial Neural Network is also presented.

2 MATERIALS AND METHODS

The study took place within the Department of Electronics and Communication Engineering at Saveetha School of Engineering. The analysis was divided into two subsets, each comprising 15 samples. Subset 1, containing 15 samples, was dedicated to Logistic Regression, while Subset 2 was assigned to Artificial Neural Network. The sample sizes for each subset were determined through a power calculation, employing 80.200% pretest power, an alpha error of 0.94, a threshold of 0.05, and a confidence level of 83.066%.

A cervical cancer dataset was employed to determine the presence of the disease. The dataset was obtained from the Cervical Cancer.cv website and consists of 5 categories, 186 attributes, and 11,650 instances. Each subset was sampled individually, resulting in a total of 30 unique samples for the test dataset. This set of 30 samples was subsequently divided into training and testing datasets. Once the dataset was partitioned, the algorithms were applied to the training and testing sets to predict accuracy values.

2.1 Logistic Regression

A logistic regression model predicts a dependent variable by analyzing the connection between existing independent variables. It can predict binary outcomes, such as the success of a political candidate or the acceptance of a high school student into a specific institution. Logistic regression's significance has grown in machine learning, enabling the classification of new data based on historical information. As new data becomes available, algorithms improve their accuracy in classifying data points.

Algorithm for sample 1 preparation

1. Collect data from patients diagnosed with cervical cancer, including factors like age, family history, HPV status, etc. Clean the data by addressing missing values, outliers, and scaling features.
2. Divide the dataset into two subsets: training and testing.
3. Develop a logistic regression model using relevant libraries like scikit-learn or TensorFlow. Train the model on the training data to predict the likelihood of cervical cancer based on independent factors.
4. Evaluate the model's performance on the testing data using metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve.
5. If needed, fine-tune the model by adjusting hyperparameters, conducting feature engineering, or exploring alternative algorithms.
6. Deploy the trained model in a production environment to predict cervical cancer outcomes for new data.
7. Regularly monitor the model's performance and update it as necessary to maintain accuracy and relevance.

2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are constructed from interconnected units called artificial neurons, mimicking the structure of neurons in the human brain. Similar to synapses, these connections enable the transmission of signals between neurons. Upon processing incoming signals, an artificial neuron can transmit signals to its connected neurons. The neuron's output is determined by a non-linear function applied to the sum of its inputs, and the "signal" transmitted across a connection is represented by a real number. These connections are referred to as edges. Neuron and edge weights are adaptable and change during the learning process. Weight adjustments influence the strength of a connection's signal by either increasing or decreasing it. Additionally, some neurons may possess a threshold that incoming signals must exceed before transmission occurs.

Algorithm for sample 2 preparation

1. Gather patient information, medical history, clinical presentation, lab results, imaging studies, and pathology reports from cervical cancer patients.

2. Clean and preprocess the collected data to rectify errors, inconsistencies, and missing values. This stage may involve feature selection, feature scaling, and data normalization.
3. Thoroughly clean and preprocess the data to rectify errors, inconsistencies, and missing data points. This process might encompass feature selection, feature scaling, and data normalization.
    1. Train the artificial neural network using the pre-processed data. This involves inputting data into the network, adjusting weights and biases to minimize the disparity between expected and actual outputs.
    2. Validate the trained neural network's performance using a separate data set not utilized in training. This step ensures the network's ability to generalize to new data.
    3. Test the neural network on a fresh data set to assess accuracy, sensitivity, specificity, and other performance metrics.
    4. Enhance neural network performance by adjusting hyperparameters like learning rate, batch size, and epochs.
    5. Implement the neural network in a clinical environment, enabling cervical cancer diagnosis or risk estimation. Integration with electronic health records or medical systems may be part of this deployment.

The system operates on a 64-bit version of Windows 11, employing an Intel i5, 11th Generation processor, along with 8GB of RAM. The implementation is carried out using Python within Jupyter via Anaconda. Independent variables consist of data from external beam radiation therapy, utilized for cervical cancer detection and diagnosis. Corresponding to the provided dependent variables, improved accuracy metrics are obtained from EBRT frequency signals. Alterations to the independent factors directly influence the dependent variables.

2.3 Statistical Analysis

The comparison between the Logistic Regression and Artificial Neural Network algorithms was conducted using the IBM SPSS statistical tool and an independent sample T-test. The analysis involved both dependent and independent variables, where the accuracy was considered the dependent variable. The independent variables included the mean and variance, extracted from the dataset.

3 RESULT

When comparing the proposed and existing methodologies, the accuracy rate of detection is a key consideration for both approaches. The ability of the two systems to predict sleepiness is used for research and comparison purposes. Notably, the accuracy rate of the proposed method, at 83.066%, is higher than that of the current algorithm.

Figure 1 illustrates the Logistic Regression training accuracy, validation accuracy, and other relevant parameters. During training, the recall stands at 83.066%, indicating the proportion of patients with the disease correctly identified. Precision is noted at 92%, representing the accuracy of disease identification. In the validation phase, the macro average of recall is 50%, indicating how accurately patients with the disease are identified.

Figure 2 presents the confusion matrix, showcasing the performance and classification outcomes of the Artificial Neural Network algorithms. This matrix displays the prediction of actual values by the algorithm.

Figure 3 provides a comparison of the mean accuracy between the Logistic Regression and artificial neural network approaches. The mean accuracy of the Logistic Regression algorithm is 83.066%, demonstrating its superiority over the artificial neural network algorithm, which achieves an accuracy of 80.200%.

Table 1: Presents a comparison between the Logistic Regression algorithm and the Artificial Neural Network.

SL.no	Test	ACCURACY RATE	
		Logistic regression	Artificial neural network
1	Test1	98	81
2	Test2	97	80
3	Test3	96	76
4	Test4	94	78
5	Test5	95	77
6	Test6	90	79
7	Test7	91	75
8	Test8	89	71
9	Test9	88	73
10	Test10	85	72
11	Test11	86	74
12	Test12	84	70
13	Test13	83	69
14	Test14	80	68
15	Test15	79	67
Mean Average (In Percentage)		83.066	80.200

Table 2 displays the mean and standard errors for both the Logistic Regression and artificial neural network methods. The mean accuracy for Logistic Regression is 75.036%, and the standard error is 1.48474, while for the artificial neural network, the mean accuracy is 74.857%, and the standard error is 1.85471. An independent sample test indicates a statistically significant difference in accuracy between the two algorithms. Notably, the Logistic Regression algorithm exhibits a standard error of 1.48474 when compared to the artificial neural network algorithm Table 3 provides a comparison between the subsets and accuracy of the Logistic Regression and Artificial Neural Network algorithms. Utilizing an independent sample test with a significance level of 0.05, a statistically significant difference in point increment accuracy between the two algorithms is observed. The accuracy of the Logistic Regression method is recorded at 83.066%, whereas the Artificial Neural Network approach achieves an accuracy of 80.200%.

Table 2: Represents Mean and standard error of the logistic Regression algorithm and algorithms with Artificial neural network respectively.

	Subset	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Logistic Regression algorithm	15	83.06	5.75036	1.48474
Accuracy	Artificial neural network	15	80.20	7.18331	1.85472

4 DISCUSSION

The study demonstrated that the Logistic Regression algorithm outperformed the ANN algorithm with a higher detection rate of 83.066% accuracy. This outcome was determined through an independent sample T-test. The proposed method simplifies the complexity of detecting point rates by employing the Logistic Regression algorithm. The analysis indicates that the Logistic Regression method surpasses the Artificial Neural Network in addressing cervical cancer identification challenges ("Experiments with Large Ensembles for Segmentation and Classification of Cervical Cancer Biopsy Images" n.d.). Notably, these references support the study's findings, as the Logistic Regression method displayed improved identification and achieved higher accuracy than the Artificial Neural Network approach. Additionally, a machine learning classifier built on the Artificial Neural Network accurately predicts the likelihood of HPV infection for a patient, achieving an accuracy of 80.200%. On the other hand, a method utilizing HPV-based features and Logistic Regression achieved a high accuracy of 83.066% in identifying cervical cancer ("Preprocessing for Automating Early Detection of Cervical Cancer" n.d.2021).

Upon implementation, the suggested models and algorithms revealed several limitations. Notably, the Artificial Neural Network method displayed instability, making it susceptible to even minor changes in data, which significantly influenced its configuration and rendered it unreliable. Furthermore, various predictors demonstrated better performance with similar data ("Cervical Cancer Single Cell Image Data Augmentation Using Residual Condition Generative Adversarial Networks" n.d.2021). Secondly, grappling with the dataset posed significant challenges due to the extensive data listed and assessed during the data pre-processing stage. Optimal results were achievable only when a substantial array of data-processing techniques were employed. Lastly, the study aimed to conduct sentiment analysis on cervical cancer through machine learning using survey data.

Table 3: Compares Subset and accuracy of Logistic Regression Algorithm and Artificial neural network algorithms. According to an independent sample test with a p=0.23(P<0.05) value, there is a statistically 2-tailed significant difference between two algorithms' point increment accuracy.

		Levene's Test for Equality of variances		T-test for Equality of Means						
		F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std. Error Difference	95% confidence interval of the Difference	
									Lower	Upper
Accuracy	Equal variances assumed	1,116	.300	1.207	26	.238	2.86667	2.3758	-1.99995	7.733
	Equal variances, not assumed			1.207	26.720	.238	2.86667	2.3758	-2.01047	7.743

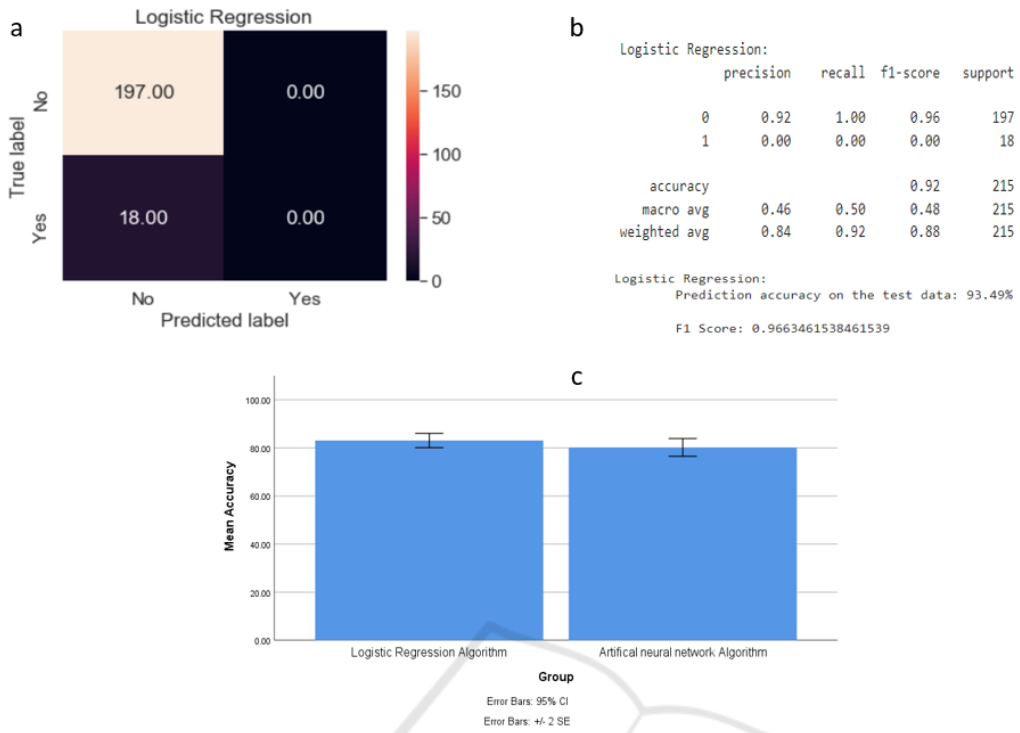

Figure 1: (a) confusion matrix shows performance and classification of Artificial neural network algorithms (b) Logistic Regression training accuracy, validation accuracy, and other parameters (c) Mean accuracy.

However, due to constraints, the researchers couldn't implement diverse data-processing techniques to apply the ML models as intended ("Application of Support Vector Based Methods for Cervical Cancer Cell Classification" n.d.2020).

5 CONCLUSION

The performance of diagnosing cervical cancer disease was analyzed using the Logistic Regression and Artificial Neural Network methods. The findings indicate that the Logistic Regression algorithm achieved an accuracy of 83.066%, while the ANN achieved an accuracy of 80.200%. In terms of diagnostic analysis, the Logistic Regression Algorithm outperforms the Artificial Neural Network Algorithm.

REFERENCES

Al Mudawi, Naif, And Abdulwahab Alazeb. (2022). "A Model For Predicting Cervical Cancer Using Machine Learning Algorithms." *Sensors* 22 (11). Https://Doi.Org/10.3390/S22114132.

"Application Of Support Vector Based Methods For Cervical Cancer Cell Classification." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/7482239.

AS, Vickram, Raja Das, Srinivas MS, Kamini A. Rao, and Sridharan TB. "Prediction of Zn concentration in human seminal plasma of Normospermia samples by Artificial Neural Networks (ANN)." Journal of assisted reproduction and genetics 30 (2013): 453-459.

"Automated Image Analysis In Multispectral System For Cervical Cancer Diagnostic." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/8071332.

"Cervical Cancer Single Cell Image Data Augmentation Using Residual Condition Generative Adversarial Networks." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/9137494.

"Experiments With Large Ensembles For Segmentation And Classification Of Cervical Cancer Biopsy Images." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/6974021.

Jha, Manika, Richa Gupta, And Rajiv Saxena. (2021). "Cervical Cancer Risk Prediction Using Xgboost Classifier." In *2021 7th International Conference On Signal Processing And Communication (Icsc)*. Ieee. Https://Doi.Org/10.1109/Icsc53193.2021.9673474.

"Prediction Of Cervical Cancer Basing On Risk Factors Using Ensemble Learning." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/9144026.

Piyush Kumar Pareek. et al. (2022). "Predicting the spread of vessels in initial stage cervical cancer through radiomics strategy based on deep learning approach" Advances in Materials Science and Engineering vol. (2022), Article ID 1008652, 13 pages, (2022). https://doi.org/10.1155/2022/1008652

"Preprocessing For Automating Early Detection Of Cervical Cancer." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/6004107.

Ramalakshmi, M., & Vidhyalakshmi, S. (2021). GRS bridge abutments under cyclic lateral push. Materials Today: Proceedings, 43, 1089-1092.

"Region Of Interest Identification For Cervical Cancer Images." N.D. Accessed December 26, (2022). Https://Ieeexplore.Ieee.Org/Abstract/Document/9098587.

Song, Dezhao, Edward Kim, Xiaolei Huang, Joseph Patruno, Hector Munoz-Avila, Jeff Heflin, L. Rodney Long, And Sameer Antani. (2015). "Multimodal Entity Coreference for Cervical Dysplasia Diagnosis." *IEEE Transactions On Medical Imaging* 34 (1): 229–45.

S. K. Sarangi, Pallamravi, N. R. Das, N. B. Madhavi, P. Naveen, and A. T. A. K. Kumar, "Disease Prediction Using Novel Deep Learning Mechanisms," *J. Pharm. Negat. Results*, vol. 13, no. 9, pp. 4267–4275, 2022, doi: 10.47750/pnr.2022.13.S09.530

V. P. Parandhaman, (2023) "A Secured Mobile Payment Transaction Handling System using Internet of Things with Novel Cipher Policies," International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10200255.

Vickram, A. S., Srikumar, P. S., Srinivasan, S., Jeyanthi, P., Anbarasu, K., Thanigaivel, S., ... & Rohini, K. (2021). Seminal exosomes–an important biological marker for various disorders and syndrome in human reproduction. Saudi journal of biological sciences, 28(6), 3607-3615

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS