Journal of Cancer Diagnosis
ISSN: 2476-2253

Mini Review Open Access

# Cervical Cancer Diagnosis Using Data Mining Algorithm

**Johnson Gurneey***

*Department of Public Health, University of Otago, Wellington, New Zealand*

**Abstract**

A class of data mining techniques can be used to accurately diagnose cervical cancer, which has significant practical implications. In particular, the beneficial information present in a sizable amount of medical data may not only subtly advance medical technology but also, in the future, aid in the detection of cervical cancer. In order to collect and analyse picture information, this study enhances the data mining algorithm and integrates image recognition and data mining technologies. Additionally, this study fully exploits the image data to segment the cervical cancer cell image, choose the feature vector in accordance with the features of the cervical cancer cell, and create the classifier using the statistical classification approach. The test results demonstrate that this system's automatic recognition and supplementary diagnosis effects are both good. As a result, it can be confirmed in clinical settings throughout the follow-up.

**Keywords:** Cervical cancer; Diagnosis; Data mining; Algorithm; Machine learning; Medical data analysis; Artificial intelligence (AI)

## Introduction

Among the disorders that endanger people's lives are those that can cause different types of malignant lesions to appear in various parts of the body. Women have long been concerned about cervical cancer, which is the fourth leading cause of mortality among females and should not be ignored. The stage of the patient's tumour typically dictates the course of cervical cancer treatment. If found early enough, surgery can be used to treat cancer in its early stages, If the tumour has gone to an advanced stage and the majority of the lymph nodes have been invaded or metastasized, radical surgery for cervical cancer combined with radiation and chemotherapy is the only treatment choice. However, the majority of women who receive a breast cancer diagnosis are already suffering from advanced stages of the illness. The stage of a patient's cervical cancer helps doctors predict the possibility of a recurrence and adjust the patient's treatment. Predictive biomarkers are therefore becoming more and more important in assessing prognosis [1].

Our ultimate objective is to mine the information present in the data, given the quick advancement of technology. Related data mining methods for analysing huge data have a wide range of applications in areas like medical treatment, corporate operations, science, and social practise networks. This is closely related to what we do every day. Certain aspects of everything from online appointment scheduling in hospitals to intelligent disease diagnosis to traffic congestion management and catastrophe prediction are influenced by data mining algorithms that examine enormous amounts of big data. These data are valuable in and of themselves, in addition to their large number, variety of kinds, and quick updating [2].

Over many years, the theory of data mining has steadily advanced in areas like categorization and grouping. The classification issue is one of the key topics of research in data mining and involves building a classification model utilising related algorithms to assess and predict categories of unknowns. The qualities of a data sample are inputs, and the category in which it is categorised is returned as an output. The categorising problem is a made-up problem designed to divert attention from actual problems. Worldwide, rates of cervical cancer mortality and morbidity are rapidly increasing, and there are numerous reasons why [3]. It represents some of the most significant cancer risk factors while also highlighting the challenge of an ageing population. Instead, a rising number of countries are ageing quickly and experiencing rapid population growth, with cancer overtaking other causes of mortality as the number one killer. Due to the complexity of cervical cancer lesions, the numerous unknowns associated with developing cervical cancer, and other factors, as the patient population grows annually, their clinical data exhibit a high volume of clinical data, a variety of data types, extremely rapid development, and implicit features of high information value [4]. The lack of particular clinical data may be caused by patients' reluctance to provide certain private information in regard to clinical data; in other instances, the clinical data that was collected may contain noise and complex information. The aforementioned problems could be fixed by applying traditional methods. The limitations are extremely rigid, and the solution to the issue is not perfect. Because of the useful information present in the substantial amount of medical data that a cancer diagnosis encourages, the adoption of a particular type of data mining algorithm for the intelligent diagnosis of cervical cancer has significant practical implications [5].

Before the condition was even recognised, the literature employed several logistic regressions to ascertain the link between late diagnosis and Medicaid eligibility. The study investigates whether women who have received Medicaid are more likely to be diagnosed with advanced cervical cancer than women who have not. The results showed that 51% of Medicaid recipients and 42% of non-Medicaid recipients had advanced disease. The elderly and women from lower socioeconomic backgrounds are more vulnerable. The findings presented above indicate that greater awareness is necessary to ensure that at-risk women receive screening services. According to the study, 370,000 new cases of cervical cancer are diagnosed each year. Approximately 10% of new cancer cases worldwide are caused by this [6].

The first step in battling cancer worldwide is to compile as much information as you can about its incidence and mortality. Despite some

***Corresponding author:** Johnson Gurneey, Department of Public Health, University of Otago, Wellington, New Zealand, E-mail: johnson@otago.ac.nz

**Received:** 01-May-2023, Manuscript No: jcd-23-98346, **Editor Assigned:** 04-May-2023, Pre QC No: jcd-23-98346(PQ), **Reviewed:** 18-May-2023, QC No: jcd-23-98346, **Revised:** 23-May-2023, Manuscript No: jcd-23-98346(R), **Published:** 30-May-2023, DOI: 10.4172/2476-2253.1000178

**Citation:** Gurneey J (2023) Cervical Cancer Diagnosis Using Data Mining Algorithm. J Cancer Diagn 7: 178.

variation in the numbers, recent global statistics suggest that results are comparable to those from ten years ago. Examination, inspection, and treatment must be done frequently and promptly. At the molecular level, the literature investigated malignant tumour markers, offered a potent tumour prevention and therapy strategy, and analysed the development of cervical cancer using implicit data from gene chips [7]. In the literature, it was examined whether transvaginal real-time ultrasound elastography technique was actually beneficial for detecting cervical cancer. In the literature, the combined use of tumour markers for the diagnosis of cervical cancer was discussed. In the literature, it has been discussed how certain anti-HPV16E6 ribozymes affect the phenotypic and gene expression of cervical cancer cells, Intelligent Diagnosis of Cervical Cancer based on Data Mining Algorithm shown in (Figure 1) [8].

As we all know, this is the option with the lowest average risk for females over the age of 30. The research assesses the procedures used by primary care practitioners, the stimuli, the impediments to using standard testing techniques, and the extended screening intervals for low-income women. It is advised to suitably increase the time interval because few of the patients described in the literature underwent the combination test of the two screening modalities [9]. The challenges brought on by lengthening the screening interval are balanced by the errors brought on by excessively frequent screening and the error brought on by misreporting injury, which are both within the budget's acceptable range. According to estimates from the World Health Organisation, cervical cancer has the second-highest prevalence among female malignant tumours. Cancer is the third leading cause of mortality, and there is a 52.91% chance that it will be the primary or second cause of death before the age of 70. Four is likely to occur in roughly 12.79% of cases. Finally, cancer is ranked fifth to tenth in the remaining 59 nations (a total of 172 nations) [10]. The literature suggested a way of detection for figuring out whether a woman has cervical cancer or is at risk for developing it. It can then be tested to see if a woman is ill or susceptible to it by looking for Brn-3amRNA or Brn-3a protein (quality) in Brn-3a cells. In order to ascertain the characteristics of the data composition, endometrial cancer and ovarian cancer were examined. In order to ascertain the prevalence of cervical cancer and compare it to women who are not infected with HIV, the literature looks at the diagnosis of ICC every six years in women with human immunodeficiency virus. An August study had women as its main subject. The study investigated whether or not the Pap test was HIV-positive.

The limitations of conventional data analysis methods become evident when confronted with similar data, resulting in subpar results and a notable loss of valuable information. As a consequence, the inherent information and patterns within the data remain untapped and undiscovered. Thus, it becomes imperative to grasp the intricacies of data processing and classification procedures in order to extract the concealed information embedded in the sample data. (Figure 2) [11].

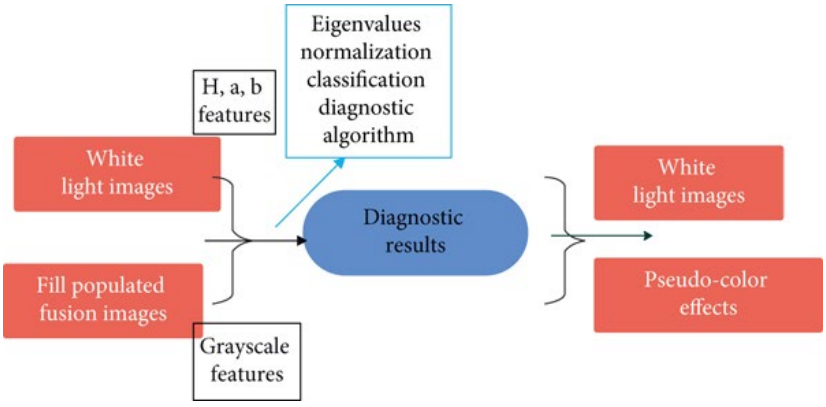


**Figure 1:** Intelligent diagnosis of cervical cancer based on data mining algorithm.

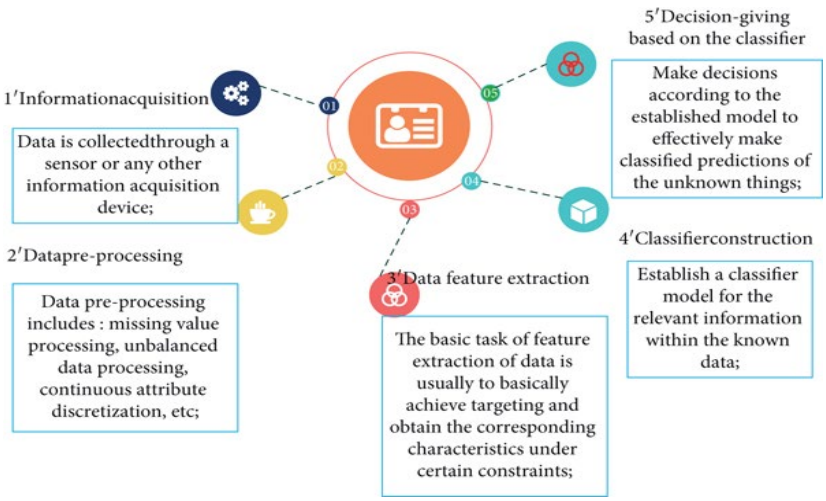


**Figure 2:** The process of processing cervical cancer diagnosis and classification problems.

## Materials and Methods

The UCI Machine Learning repository hosts the cervical cancer risk factors data set that was used in the study. It was compiled at the "Hospital Universitario de Caracas" in Caracas, Venezuela. It has 858 records, some of which have missing values because some patients choose not to respond to some questions out of respect for their privacy. 32 risk factors and 4 targets, or the cervical cancer diagnosis tests, are included in the data collection. It includes several feature sets, including recordings of behaviours, demographic data, history, and genomic medical information [12]. Age, Cancer Dx, CIN Dx, HPV Dx, and Dx characteristics don't have any missing data. Dx: CIN is an alteration in the cervix's walls that is frequently caused by HPV infection; if it is not treated effectively, it can occasionally progress to cancer. However, whether the patient has other cancer kinds or not is indicated by the Dx: cancer variable. A patient may occasionally have many cancer types. Some of the patients in the data set did not have cervical cancer, yet their Dx: cancer value was true. As a result, it is not a target variable [13].

The goal of this data is the most common diagnosis tests, with a brief description of each feature and the kind. Cervical cancer diagnosis typically involves a number of tests. Four frequently used diagnostic procedures for cervical cancer include Hinselmann, Schiller, cytology, and biopsy. The Hinselmann or Colposcopy test involves looking within the vagina and cervix with a device that magnifies the tissues in order to look for any irregularities. In the Schiller test, iodine is injected to the cervix [14], where it causes healthy cells to turn brown while leaving diseased cells uncoloured. In the cytology test, body cells from the uterine cervix are examined for any cancerous cells or other disorders. Additionally, the procedure known as a biopsy involves the microscopic examination of a small portion of cervical tissue. Most biopsy tests are capable of making important diagnoses [15].

The data collection had a significant amount of missing values; 24 of the 32 features had missing values. The characteristics with the highest percentage of missing values were initially eliminated. Since they have 787 missing values, or more than half of the data, the STDs: Time since initial diagnosis and STDs: Time since final diagnosis characteristics were eliminated. However, data imputation was carried out for the characteristics that had fewer missing value instances [16]. The remaining missing values were imputed using the most frequent value method. Additionally, there is a significant class imbalance in the data set. As shown in Figure 1, the target labels for the data set were unbalanced, with 35 for the Hinselmann, 74 for the Schiller, 44 for cytology, and 55 for biopsies out of the 858 records. To address the disparity in class, SMOTE was applied. By creating additional synthetic data for minority instances based on nearest neighbours and using the Euclidean Distance between data points, SMOTE oversamples the minority class. The number of records in the data collection for each class label [17].

One of the most efficient methods for choosing the features that enhance the performance of the supervised learning model is dimensionality reduction. For the study, we used the Firefly algorithm, which was inspired by nature, to choose the attributes that would best express the issue. Yang was the one who first suggested Firefly for the optimisation. The Metaheuristic Firefly algorithm takes its cue from fireflies and a fly's ability to flash lightning. Finding the ideal value or parameter for a target function is done using a population-based optimisation technique. Each fly is drawn out using this method by the strength of the glow from the neighbouring flies. The attraction will be waning if the gleam's intensity ever drops below a certain point [18].

Firefly follows three rules: (a) all the flies must be the same gender; (b) the glow's intensity determines how enticing something is; and (c) the firefly's gleam is produced by the target function. The flies with fewer glowing cells will approach the flies with more glowing cells. The objective function allows for the adjustment of brightness. The algorithm uses the same concept to look for the best attributes that can suit the training model. In feature selection, Firefly outperformed other metaheuristic methods like genetic algorithms and particle swarm optimisation and was more computationally efficient [19].

Breiman first suggested Random Forest (RF) in 2001. An ensemble model called random forest uses bagging as the ensemble method and decision trees as the individual model. By including more trees, it helps the decision tree operate better by reducing overfitting. Both classification and regression can be done using RF. The best option is chosen with the most votes after RF creates a random forest with decision trees in it and receives predictions from each one of them. It's critical to gauge how much each feature reduces impurities when training a tree since the impurity reduction reveals the importance of the feature. The impurity metric that is applied affects the tree classification outcome. Information gain or Gini impurity are the measurements for impurity in classification, while variance is the measure for impurity in regression. Iterative data splitting is used when training a decision tree. The formula used by the Gini impurity to determine the optimal data split [20].

## Conclusion

Computer systems have been used more frequently in the service industry thanks to advancements in information technology, particularly in hospitals. The hospital's Gynaecology Research Office is primarily in charge of evaluating cervical cancer cell images, categorising the findings of cell evaluations, and evaluating cancerous, healthy, and doubtful cells. However, the processing of these processes requires image processing software. The data centre implements front and back ends using database software and system development tools. A distinct environment will be maintained for the business processing centre. The business and data processing environment is safeguarded in this way in order to prevent serious errors brought on by catastrophic failures. This article discusses the use of image data to separate cervical cancer cell images, selecting feature vectors based on cell characteristic data, and creating a classifier using statistical techniques. This system has a great automatic recognition effect and a decent additional diagnostic effect, per the test findings. Clinical practise in follow-up could therefore confirm it.

### Conflicts of Interest

None

### Acknowledgement

None

#### References

1. Bunn PA, Dziadziuszko R, Varella-Garcia M (2006) Biological markers for non-small cell lung cancer patient selection for epidermal growth factor receptor tyrosine kinase inhibitor therapy. Clin Cancer Res 12: 3652-3656.
2. Riely GJ, Politi KA, Miller VA, Pao W (2006) Update on epidermal growth factor receptor mutations in non-small cell lung cancer. Clin Cancer Res 12: 7232-7241.
3. Yu J, Kane S, Wu J (2009) Mutation-specific antibodies for the detection of EGFR mutations in non-small-cell lung cancer. Clin Cancer Res 15: 3023-3028.

4. Mossé YP, Wood A, Maris JM (2009) Inhibition of ALK signaling for cancer therapy. Clin Cancer Res 15: 5609-5614.

5. Kadara H, Behrens C, Yuan P (2011) A five-gene and corresponding protein signature for stage-I lung adenocarcinoma prognosis. Clin Cancer Res 17: 1490-1501.

6. Austin LA, Osseiran S, Evans CL (2016) Raman technologies in cancer diagnostics. Analyst 141: 476-503.

7. Bhattacharjee T, Khan A, Maru G, Ingle A, Krishna CM, et al. (2015) A preliminary Raman spectroscopic study of urine: diagnosis of breast cancer in animal models. Analyst 140: 456-466.

8. Huang ZW, McWilliams A, Lui H, McLean DI, Lam S, et al. (2003) Near-infrared Raman spectroscopy for optical diagnosis of lung cancer. Int J Cancer 107: 1047-1052.

9. Bergholt MS, Zheng W, Lin K (2011) In vivo diagnosis of gastric cancer using Raman endoscopy and ant colony optimization techniques. Int J Cancer 128: 2673-2680.

10. Haka AS, Volynskaya Z, Gardecki JA (2006) In vivo margin assessment during partial mastectomy breast surgery using Raman spectroscopy. Cancer Res 66: 3317-3322.

11. Pectasides D, Pectasides M, Nikolaou M (2005) Adjuvant and neoadjuvant chemotherapy in muscle invasive bladder cancer. Eur Urol 48: 60-67.

12. Mitchell E, Macdonald S, Campbell NC, Weller D, Macleod U, et al. (2008) Influences on pre-hospital delay in the diagnosis of colorectal cancer. Br J Cancer 98: 68-70.

13. Singh H, Daci K, Petersen LA (2009) Missed opportunities to initiate endoscopic evaluation for colorectal cancer diagnosis. Am J Gastroenterol. 104: 2543-2554.

14. Ferrari M (2005) Cancer nanotechnology: opportunities and challenges. Nat Rev Cancer 5: 161-171.

15. Miller AD (2003) the problem with cationic liposome/micelle-based non-viral vector systems for gene therapy. Curr Med Chem 10: 1195-1211.

16. Cho K, Wang X, Nie S, Chen ZG, Shin DM, et al. (2008) Therapeutic nanoparticles for drug delivery in cancer. Clin Cancer Res 14: 1310-1316.

17. Davis ME, Chen ZG, Shin DM (2008) Nanoparticle therapeutics: an emerging treatment modality for cancer. Nat Rev Drug Discov 7: 771-782.

18. Lammers T, Hennink WE, Storm G (2003) Tumour-targeted nanomedicines: principles and practice. Br J Cancer 99: 392-397.

19. Byrne JD, Betancourt T, Brannon-Peppas L (2008) Active targeting schemes for nanoparticle systems in cancer therapeutics. Adv Drug Deliv Rev 60: 1615-1626.

20. Matsumura Y, Maeda H (1986) A new concept for macromolecular therapeutics in cancer chemotherapy: mechanism of tumoritropic accumulation of proteins and the antitumor agent smancs. Cancer Res.46: 6387-6392.