International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

COPY RIGHT

2023 IJIEMR. Personal use of this material is permitted. Permission from IJIEMR must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. No Reprint should be done to this paper, all copy right is authenticated to Paper Authors

IJIEMR Transactions, online available on 27th Mar 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 03/44

Title **DIAGNOSIS AND PREDICTION OF CERVICAL CANCER: AN INTEGRATED APPROACH USING MACHINE LEARNING ALGORITHMS**

Volume 12, ISSUE 03, Pages: 314-322

Paper Authors

Dr.M.V.L.N. Raja Rao, Venkata Narayana Battula, Vasanthi Kanna, Swathi Jetti, Sonika Nagidi

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Diagnosis and Prediction of Cervical Cancer: An Integrated Approach Using Machine Learning Algorithms

Dr.M.V.L.N. Raja Rao,
Professor, Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh.
Venkata Narayana Battula,
IV B. Tech Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh.
Vasanthi Kanna,
IV B. Tech Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh.
Swathi Jetti,
IV B. Tech Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh.
Sonika Nagidi,
IV B. Tech Department of Information Technology, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh.

Abstract
Medical professionals have labeled cervical cancer as a possibly fatal illness. Late detection and therapy pose significant challenges and put patients' lives in jeopardy. Formal screening for disease detection is hindered by factors such as high medical costs, a lack of accessible healthcare facilities, cultural attitudes, and a delay in the onset of symptoms in both established and developing countries. Several illnesses, including cervical cancer, can be diagnosed early with low processing costs using machine intelligence. Machine-intelligent solutions make early detection of cervical cancer much easier, and patients don't have to go through modern, time-consuming medical treatments. The present machine categorization techniques for illness detection suffer from over-reliance on the forecast precision of a particular predictor. Due to prejudice, over-fitting, improperly managing noise data, and anomalies, the implementation of singular categorization techniques does not guarantee the optimal forecast. In order to arrive at a correct diagnostic that takes into account the patient's medical circumstances or symptoms, the authors of this study suggest using a majority-voting Ensemble categorization technique. Different classifications, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), are tested in this investigation. A 94% improvement in forecast accuracy was recorded in the research, which is fundamentally higher than the forecast exactness's of individual arrangement strategies assessed on the equivalent bench-marked datasets. Hence, the proposed model furnishes clinical experts with a second view to aid in the early detection and management of illness.

Keywords: Machine Learning Algorithms, Classifiers, Cervical Cancer, Ensemble Classification

1. INTRODUCTION

In modern medicine, cancer has proven to be a major cause of death [1]. The illness is typified by the aberrant activity of the afflicted cells. Normal tissue is damaged by cancer cells, which disrupts their function [2]. In addition, cancer has a high risk of metastasizing, or extending to other areas of the body. Consequently, in some instances, mortality may result from cancer that was not detected early. There is a correlation between early detection and increased mortality rates for most cancers [3]. One of the most prevalent cancers in women, cervical cancer is almost always brought on by human papillomavirus (HPV) (Human Papilloma Virus). The ability to identify cervical cancer at an early stage has been shown to have a major effect on patients' therapy and healing in several studies [4, 5]. The Pap smear (Papanicolaou test) [6], HPV DNA sequencing [7], HCII (Hybrid Capture II), hybrid capture [8], [9], and Southern blot hybridization analysis [10] are all methods commonly used to identify cervical cancer. Biosensors are the most recent innovation in this area, and they hold great promise because of their cheap cost, quick findings, and user-friendliness [11], [12]. None of these methods is without flaws. Major drawbacks of the pap screening test include its high misleading negative rate [13], unfortunate explicitness [13], and high cost [13]. This is on the grounds that the southern smudge hybridization method is a tedious and heartless measure [14]. Testing for HPV DNA sequencing is tedious and exorbitant [15]. One of the most state-of-the-art techniques for spotting cervical disease in its earliest stages is HCII hybrid capture. On the other hand, it can only pick up on 13 different HPV types [16]. There are a number of limitations to biomedical sensors' usefulness in practice, including their inability to reuse electrodes, unstable performance, and short lifespan [17]. Because of these drawbacks, new approaches are needed to improve cervical cancer screening. "Computational techniques using experience to enhance performance or to make correct forecasts" is how machine learning is defined in reference [18]. Personal [19], financial [20], big business [21], government [22], military [23], and even space science [24] are just a few examples of the many fields where machine learning methods have been successfully implemented. The medical field has seen some of the most promising applications of machine learning to date [25], [26]. Many forms of cancer may benefit from the use of machine learning [27]– [29], and cervical cancer is no exception [30, 31]. Currently, most methods for detecting cervical cancer rely on only a single classification. This study suggests employing Ensemble methods for cervical cancer diagnosis using two freely accessible datasets, as they can typically yield superior outcomes in particular situations [32]. The method takes advantage of a number of models, such as the Support Vector Machine, Random ForestJ48 Tresses, and Logistic Regression. Classification of the objective characteristic is decided by a consensus vote. Following this structure is the remainder of the report. In Part 2, we examine and evaluate new literature on the subject of cervical cancer detection. The methods used in this study is discussed in greater detail in Section 3. The findings are presented in Section 4, and the study is wrapped up in Section 5.

LITERATURE REVIEW

[1], X. Li, and J. Liu, "Application of cyclodextrins in cancer treatment," J. Inclusion Phenomena Macrocyclic Chem., vol. 89, nos. 3–4, pp. 229–246, Dec. 2017, doi: 10.1007/s10847-017-0752-2.

Among the leading causes of death, cancer stands out. Chemotherapy is a common treatment that employs a variety of medicines in an effort to eliminate cancer cells or at least halt their development. Formulation challenges arise because most deadly chemotherapy medicines are impermeable in water. Use of cyclodextrins
(CDs), which have been extensively used to improve the solubility, absorption, stability, and safety of medicinal compounds by creating non covalent inclusion complexes, is an exciting new

direction. This study aims to shed light on how Discs fit into the various cancer therapy modalities currently in use. Discs have been shown to have antitumor action in laboratory and animal models, which is of particular relevance. Discs are addressed in relation to their potential use as an antitumor drug and the possible method by which they might do so. Future uses and methods for creating CDs/anti-neoplastic-drug compounds with better solubility, stability, and anti-cancer action are addressed. This study describes the benefits of using CDs in various drug delivery systems, such as liposomes, conjugates, nanoparticles, and siRNA transporters, for the therapy of cancer.

[2] L. Schwartz, C. Supuran, and K. Alfarouk, "The warburg effect and the hallmarks of cancer," Anti-Cancer Agents Med. Chem., vol. 17, no. 2, pp. 164–170, Jan. 2017, doi: 10.2174/1871520616666161031143301.

Whether cancer should be considered a single illness or a collection of very different conditions has been debated for decades. This article aims to provide convincing evidence that the Warburg effect is responsible for many, if not all, of the characteristics typically associated with cancer. There is a drop in ATP levels because of biochemical dysfunction of the oxidative phosphorylation. Massive glucose absorption, anaerobic glycolysis, and up-regulation of the Pentose Phosphate Pathway result in increased biosynthesis, which in turn leads to increased cell proliferation and local pressure to make up for the lower energy output. This heightened pressure is crucial in spreading dissemination, tumor fractalization, and collagen production by fibroblasts. Large amounts of lactic acid are secreted, increasing the pH of the external environment and stimulating the immune system. There is something like two malignant growth related carbonic anhydrase types that impact the impacts of diminished oxidative phosphorylation on CO2 levels inside and outside the cell, prompting more prominent intracellular alkalosis and the expansion of carbonic corrosive to outer sharpness.

[3] E. M. Hassan and M. C. DeRosa, "Recent advances in cancer early detection and diagnosis: Role of nucleic acid based apt sensors," TrAC Trends Anal. Chem., vol. 124, Mar. 2020, Art. no. 115806, doi: 10.1016/j.trac.2020.115806.

Cancer's complexity makes it a global health problem. Early and precise discovery of cancer and cancer indicators not only improves patient results, but also helps researchers better understand the disease's evolution and create more targeted treatments. To recognize and identify cancer cells and their indicators, apt sensors have shown tremendous potential among the many current detection techniques. Late progressions in aptamer-based frameworks for malignant growth exosome identification are summed up, alongside those in optical and microfluidic apt sensors for early disease recognition.

First, we provide an overview of how aptamers are made and what characteristics they have, and then we discuss the most cutting-edge visual apt sensors for recognizing and detecting cancer cells. In the following section, we explore microfluidic aptamer-based methods for precise biomarker and cancer cell identification. Next, cutting-edge florescent, colorimetric, and other techniques for detecting cancer exosomes are outlined. In conclusion, we offer a look into the potential of apt sensors for use in cancer diagnosis in the future.

[4] N. A. Parmin, U. Hashim, S. C. B. Gopinath, S. Nadzirah, Z. Rejali, A. Afzan, and M. N. A. Uda, "Human papillomavirus e6 biosensing: Current progression on early detection strategies for cervical cancer," Int. J. Biol. Macromolecules, vol. 126, pp. 877–890, Apr. 2019, doi: 10.1016/j.ijbiomac.2018.12.235.

Perhaps of the main test confronting the medical care framework today is deciding

the probability of early disease conclusion. Cervix malignant growth surrenders early screening, risk factors cross-response, versatility gadget, quick and free stamping framework have turned into a well-known subject of conversation among clinical trained professionals and specialists inspired by treatment strategies. Hybridization of DNA duplexes between test target and DNA sensors loans belief to the electrical biosensing based framework's cases of expanded accuracy and segregation. Because of its elite presentation, straightforward taking care of, fast framework, and potential for scaling down, the electrical DNA screen for cervical malignant growth has accumulated an excess of interest from specialists. In this article, we investigate the territory of HPV E6 gene biosensing for early Cervical Disease Recognition Procedures as well as its new headways and expected future unimportance.

5. N. Wentzensen and M. von Knebel Doeberitz, "Biomarkers in cervical cancer screening," Disease Markers, vol. 23, no. 4, pp. 315–330, 2007, doi: 10.1155/2007/678793.

The event of cervical disease has diminished essentially in evolved countries because of broad cytological observing projects utilizing the Pap test. In spite of the reasonable viability of Pap-stained cytological examples in screening programs, there are still some drawbacks to be aware of. For starters, a proportion of high-grade tumours that require urgent therapy lurk among ambiguous or slightly anomalous test results, necessitating expensive work up via repetitive retesting or direct colposcopy and excision. The expenses associated with investigating cytological results that are borderline or inconclusive account for a sizable portion of the total expenditures on cervical cancer screening. Costs could be drastically cut if handling of these data was improved. Human papillomavirus (HPV) is a malignant human virus that causes cervical cancer when it infects a woman repeatedly (HPV). Although HPV contamination is required for carcinogenesis, it is not adequate on its own. While nearly all acute HPV infections cause low-grade precursor lesions, over 90% of these lesions resolve on their own within a few months, and only 10% proceed to high-grade lesions or metastatic malignancy. Infected basal and parabasal cells display uncontrolled production of viral oncogenes E6 and E7 during progression. Lesions with a high risk of development can be better detected in primary screening and emergency situations with the help of novel indicators that enable tracking of these important molecular processes in histopathological or cytological tissues. Likely biomarkers for cervical disease screening will be talked about, with an accentuation on the clinical information that upholds their execution as new markers in superior cervical malignant growth screening programs

1. PROBLEM STATEMENT

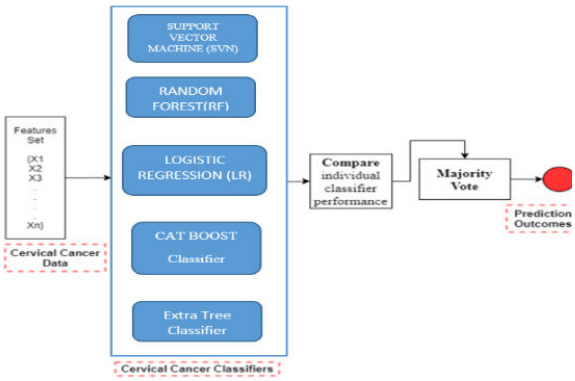
The World Health Organization ranks cervical cancer as the third leading cause of death in women globally. Most cases of this illness are brought on by HPV, which causes a growth to grow in the affected area. Medical experts say that a remedy can be found in prompt diagnosis and immunization. The majority (around 99.7%) of cervical cancer cases can be traced back to infection with human papilloma-virus. We can't learn the full extent of our health situation until we undergo all of these tests and vaccinations, which require time. Progress is painfully slow. Expenses are rather expensive. Seek Qualified Individuals.

2. PROPOSED SYSTEM

The suggested system uses a majority-voting Ensemble categorization technique to determine the cause of the patient's complaints. Different classifications, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), are tested in this investigation. A 94% improvement in forecast accuracy was recorded in the research, which is altogether higher than the forecast correctness of individual arrangement strategies assessed on the equivalent bench-marked datasets.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Accurate. Classification. There is less going on. Exceptional efficiency.



3. METHODOLOGY AND ALGORITHMS

5. 1 Logistic regression

In Machine Learning, logistic regression is a categorization method used to make predictions about a classified dependent variable. The dependent variable in logistic regression is a binary one, with values of 1 (yes, achievement, etc.) or 0 (no) (no, failure, etc.). That is, given a set of independent variables X, the logistic regression model can make a prediction about the probability that the outcome Y will be 1.

5.2. Support vector classifiers algorithm

The SVM method, short for Support Vector Machine, is a remarkably versatile Guided Machine Learning technique that can be applied to the construction of both regression and categorization models. Both linearly and non-linearly separable samples work well for the SVM method. The support vector machine method continues to impress even when given a small dataset.

Step 1: Start by bringing up the Pandas environment and loading the information.

Step 2: Specify the parameters and your end goal.

Step 3: The dataset was divided into train and test sets using sclera, and then an SVM model was constructed using the train and test sets.

Step 4: Bring in Sclera's support vector machine (SVM) module's SVC function, which performs support vector classification. Create a model of a Support Vector Machine using the SVC procedure.

Step 5: Estimate numbers with the SVM algorithm in model

Step 6: Think about how well the Support Vector Machine paradigm works.

5.3. Random Forest:

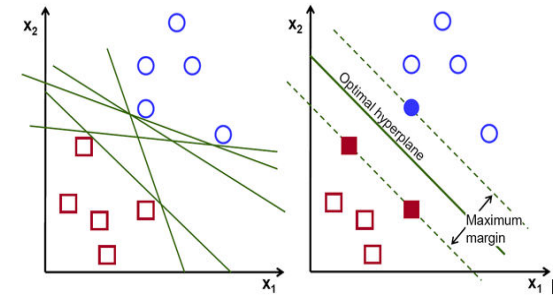
When it comes to machine learning, a random forest is a go-to method for addressing issues of regression and categorization. The method employs ensemble learning, which pools the results of multiple classifications to address difficult issues.

Multiple decision trees make up a random forest method. The random forest method produces a "forest" that is taught using bagging or bootstrap aggregation. By working as a group meta-algorithm, bagging enhances the precision of ML algorithms.

Using the information contained in the decision trees, the (random forest) method determines the final result. It makes forecasts by averaging the results of several plants. If there are more plants, the result will be more accurate.

5.4 SUPPORT VECTOR MACHINES:

The goal of the support vector machine method is to locate, in an N-dimensional space (N — the number of characteristics), a hyper plane that accurately and reliably categorizes the data points.



5.5 Support Vectors

The location and direction of the hyper plane can be affected by the presence of support vectors, which are data points that are closest to the hyper plane. By utilizing these training support vectors, we are able to increase the classifier's margin to its maximum potential. When the support vectors are removed, the

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

hyper plane shifts. Incorporating these elements into our SVM allowed us to achieve the desired results.

5.6 Large Margin Intuition

In logistic regression, we take the value returned by the linear function and, using the sigmoid function, squeeze it into the interval [0,1]. A flattened value is labelled as "1" if it is higher than a cut-off value (0.5), and "0" otherwise. In SVM, we use the linear function's result to assign a class label if the value is higher than 1, and assign a different class label if the value is less than 1. When the thresholds in SVM are adjusted to 1 and -1, we get the range of reinforcement values [-1, 1] that serves as a safety net.

5.7. Cat Boost:

To execute gradient boosting on decision trees, CatBoost is a high-performance, open-source tool. To put it another way, CatBoost is a decision tree gradient boosting method. Yandex's own academics and programmers created it, and it's already being put to use for things like search, suggestion systems, personal helper, self-driving vehicles, weather forecast, and more at Yandex and elsewhere, such as at CERN, Cloudflare, and Careem cab. The code is freely available for anyone to use. Just over a year after its introduction, CatBoost has emerged as a serious competitor to established technologies like XGBoost and LightGBM.

The greatest standard outcomes are achieved by CatBoost, which is fantastic. However, when considering datasets in which category characteristics play a crucial part, the significance and irrefutability of this enhancement becomes clear.

5.8. Extra Tree Classifier:

The Exceptionally Randomized Trees Classifier (Extra Trees Classifier) is a method of ensemble learning that uses the combined expertise of many individuals, independently trained decision trees arranged in a "forest" to make a single, unified categorization. It is conceptually very close to a Random Forest Classifier, with the primary difference lying in the specifics of how the judgement trees in the forest are built.

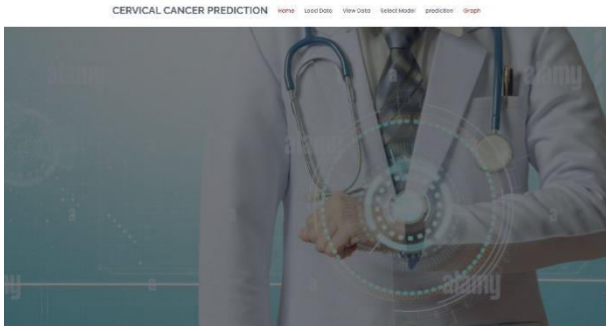
In the Extra Trees Woods, the preparation information is utilized to make individual Choice Trees. Then, at that point, at each test hub, the trees are given a free irregular subset of k elements from the list of capabilities, from which they should pick the most instructive component for information parcelling as per some factual basis (commonly the Gini File). By picking out characteristics at random, we can generate several independent decision trees.

At the time the forest is being built, the standardized total decrease in the statistical model utilized in the decision of element of parted (the Gini File, assuming that the Gini List is being utilized) is determined for each component to lead highlight determination utilizing the above backwoods structure. The feature's Gini Significance corresponds to this numerical number. When performing feature selection, features are sorted from most to least important using the Gini Index, and the customer chooses the top k features.

4. OUTPUT SCREEN SHOTS WITH DESCRIPTION.

Home Page:

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org



Load Data:

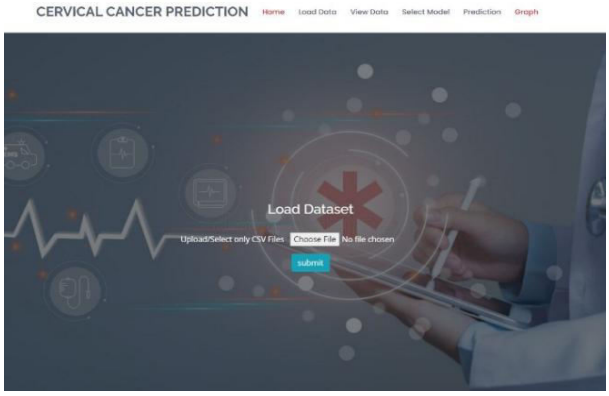


Fig3: Load Data.

ViewData:

Fig4: View Data.

Select Model:

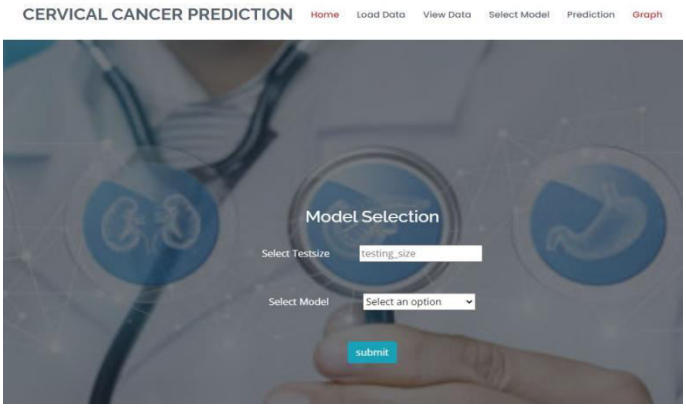


Fig5: Select Model.

Prediction:

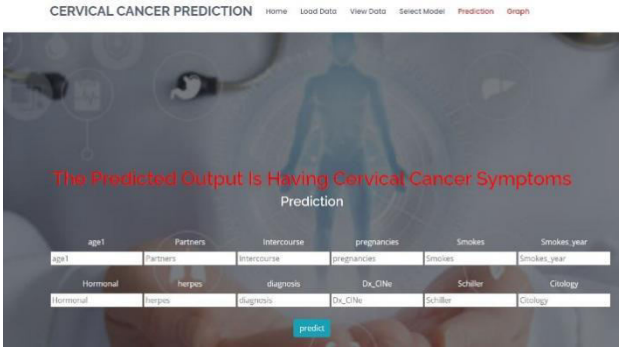


Fig6: Prediction results.

Graph:

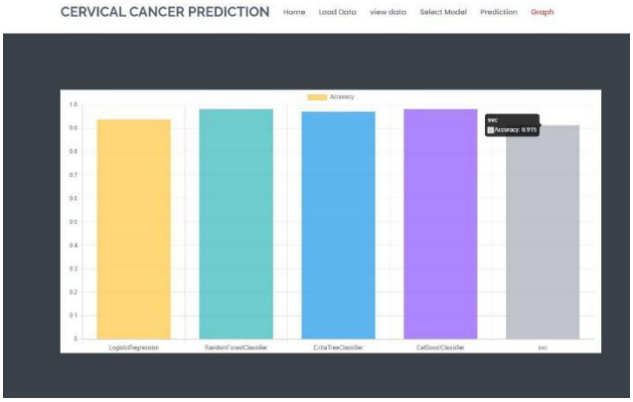


Fig7: Graph

5. CONCLUSION:

Cervical cancer detection must be quick and accurate if we're going to meet the health 4.0 criteria and keep people from

losing their lives to the disease. While there are many obstacles that prevent early illness detection, the present machine intelligent methods are deemed to be effective. Yet, the efficiency of different categorization strategies is a problem when trying to spot ovarian cancer. Since various classifiers produce wildly different outcomes when applied to the same datasets, it's important to ensure that the data being classified matches the classifier being used. Using a weighted voting strategy, the best cervical cancer forecast classification results were chosen and implemented in this study's ensemble classification technique. The research used a broad variety of models, including Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The suggested ensemble classifier beat the performance assessment of the individual classifiers by achieving the best accuracy at 94%. Practitioners can use the findings of this study to give patients with cervical cancer a second view based on professional knowledge and confidence.

8. REFERENCES:

[1] N. Qiu, X. Li, and J. Liu, "Application of cyclodextrins in cancer treatment," J. Inclusion Phenomena Macrocyclic Chem., vol. 89, nos. 3–4, pp. 229–246, Dec. 2017, doi: 10.1007/s10847-017-0752-2.

[2] L. Schwartz, C. Supuran, and K. Alfarouk, "The warburg effect and the hallmarks of cancer," Anti-Cancer Agents Med. Chem., vol. 17, no. 2, pp. 164–170, Jan. 2017, doi: 10.2174/1871520616666161031143301.

[3] E. M. Hassan and M. C. DeRosa, "Recent advances in cancer early detection and diagnosis: Role of nucleic acid based aptasensors," TrAC Trends Anal. Chem., vol. 124, Mar. 2020, Art. no. 115806, doi: 10.1016/j.trac.2020.115806.

[4] N. A. Parmin, U. Hashim, S. C. B. Gopinath, S. Nadzirah, Z. Rejali, A. Afzan, and M. N. A. Uda, "Human papillomavirus e6 biosensing: Current progression on early detection strategies for cervical cancer," Int. J. Biol. Macromolecules, vol. 126, pp. 877–890, Apr. 2019, doi: 10.1016/j.ijbiomac.2018.12.235.

[5] T. A. Kessler, "Cervical cancer: Prevention and early detection," Seminars Oncol. Nursing, vol. 33, no. 2, pp. 172–183, May 2017, doi: 10.1016/j.soncn.2017.02.005.

[6] N. Wentzensen and M. von Knebel Doeberitz, "Biomarkers in cervical cancer screening," Disease Markers, vol. 23, no. 4, pp. 315–330, 2007, doi: 10.1155/2007/678793.

[7] C. J. Meijer, P. J. Snijders, and P. E. Castle, "Clinical utility of HPV genotyping," Gynecol. Oncol., vol. 103, no. 1, pp. 12–17, Oct. 2006, doi: 10.1016/j.ygyno.2006.07.031.

[8] H. N. Luu, K. R. Dahlstrom, P. D. Mullen, H. M. Vonville, and M. E. Scheurer, "Comparison of the accuracy of hybrid capture II and polymerase chain reaction in detecting clinically important cervical dysplasia: A systematic review and meta-analysis," Cancer Med., vol. 2, no. 3, pp. 367–390, Jun. 2013, doi: 10.1002/cam4.83.

[9] H. N. Luu, K. Adler-Storthz, L. M. Dillon, M. Follen, and M. E. Scheurer, "Comparing the performance of hybrid capture II and polymerase chain reaction (PCR) for the identification of cervical dysplasia in the screening and diagnostic settings," Clin. Med. Insights, Oncol., vol. 7, Jan. 2013, Art. no. CMO.S12811, doi: 10.4137/CMO.S12811.

[10] T. A. Brown, "Southern blotting and related DNA detection techniques," in Encyclopedia of Life Sciences. Hoboken, NJ, USA: Wiley, 2001.

[11] Q. Wang, B. Zhang, X. Lin, and W. Weng, "Hybridization biosensor based on the covalent immobilization of probe DNA on chitosan–mutiwalled carbon nanotubes nanocomposite by using glutaraldehyde as an arm linker," Sens. Actuators B, Chem., vol. 156, no. 2, pp. 599–605, Aug. 2011, doi: 10.1016/j.snb.2011.02.004. 12386 VOLUME 9, 2021 Q. M. Ilyas, M. Ahmad: Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence

[12] N. Zari, A. Amine, and M. M. Ennaji, "Label-free DNA biosensor for electrochemical detection of short DNA sequences related to human papilloma virus," Anal. Lett., vol. 42, no. 3, pp. 519–

535, Feb. 2009, doi: 10.1080/00032710802421897.

[13] H. C. Kitchener, K. Canfell, C. Gilham, A. Sargent, C. Roberts, M. Desai, and J. Peto, "The clinical effectiveness and cost-effectiveness of primary human papillomavirus cervical screening in England: Extended follow-up of the ARTISTIC randomized trial cohort through three screening rounds," Health Technol. Assessment, vol. 18, no. 23, p. 1, Apr. 2014, doi: 10.3310/hta18230.

[14] J. Hoebeeck, F. Speleman, and J. Vandesompele, "Real-time quantitative PCR as an alternative to southern blot or fluorescence in situ hybridization for detection of gene copy number changes," in Protocols for Nucleic Acid Analysis by Nonradioactive Probes, vol. 353. Totowa, NJ, USA: Humana Press Inc., 2007, Ch. 15, pp. 205–226, doi: 10.1385/1-59745-229-7:205.

[15] F. Şahiner, A. Kubar, R. Gümral, M. Ardıç, N. Yiğit, K. Şener, M. Dede, and M. Yapar, "Efficiency of MY09/11 consensus PCR in the detection of multiple HPV infections," Diagnostic Microbiol. Infectious Disease, vol. 80, no. 1, pp. 43–49, Sep. 2014, doi:10.1016/j.diagmicrobio.2014.03.030