Personalized Medicine Universe 4 (2015) 54–62

Contents lists available at ScienceDirect

Personalized Medicine Universe

journal homepage: www.elsevier.com/locate/pmu

Original article

Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis

Abid Sarwar [a, *], Vinod Sharma [a], Rajeev Gupta [b]

[a] Department of Computer Science & IT, University of Jammu, Jammu, India
[b] Department of Medical Oncology & Hematology, Fortis Escorts Hospital, Amritsar, India

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form
18 September 2014
Accepted 1 October 2014

Keywords:
Ensemble technique
Pap smear
Artificial neural networks
Machine learning
Cervical cancer

ABSTRACT

Objective: This paper presents an innovative idea of applying a hybrid ensemble technique i.e. ensemble of ensemble methods for improving the predictive performance of Artificial intelligence based system for screening of cervical cancer by characterization and classification of Pap smear images.
Methodology: Papanicolaou smear (also referred to as Pap smear) is a microscopic examination of samples of human cells scraped from the lower, narrow part of the uterus, called the cervix. A sample of cells after being stained by using Papanicolaou method is analyzed under microscope for the presence of any unusual developments indicating any precancerous and potentially precancerous changes. Abnormal findings, if observed are subjected to further precise diagnostic subroutines. Examining the cell images for abnormalities in the cervix provides grounds for provision of prompt action and thus reducing incidence and deaths from cervical cancer. It is the most popular technique used for screening of cervical cancer. Pap smear test, if done with a regular screening programs and proper follow-up, can reduce cervical cancer mortality by up to 80% [1]. The contribution of this paper is that we have pioneered to apply hybrid ensemble technique to screen cervical cancer by classification of Pap smear data. The hybrid ensemble designed in this work has also presented an idea to use an ensemble of ensemble techniques. Using such a technique, the classification potentials of individual algorithms are fused together to gain greater classification accuracy. In addition to this we have also presented a comparative analysis of various artificial intelligence based algorithms for screening of cervical cancer.
Results: The results indicate that hybrid ensemble technique is an efficient method for classification of Pap smear images and hence can be effectively used for diagnosis of cervical cancer. Among all the algorithms implemented, the hybrid ensemble approach outperformed & expressed an efficiency of about 96% for 2-class problem and about 78% for 7-class problem. The results when compared with the all the standalone classifiers were significantly better for both 2-class and 7-class problems.

1. Introduction

Over the last few decades artificial intelligence techniques have been increasingly used in solving problems in medical domains such as in Oncology [2–5], Urology [4], Liver Pathology [6], Cardiology [7,8], Gynecology [9], Thyroid disorders [10,11], Perinatology [12] etc. The primary concern of AI in medicine is construction of artificial intelligent systems that can assist a medical doctor in performing expert diagnosis. These artificial intelligent systems support the clinical decision making by anticipating the diagnostic results, after being trained using previously acquired training data followed by expending specific information of some patient case. The use of Artificial intelligence in medicine has shown substantial progress in achieving timely, reliable diagnosis and more precise treatment of many diseases.

1.1. Artificial intelligence

Artificial intelligence (AI) is a subpart of computer science, concerned with, how to give computers the sophistication to act intelligently, and to do so in increasingly wider realms. The field was founded on the claim that a central property of humans,

* Corresponding author.
E-mail address: sarwar.aabid@gmail.com (A. Sarwar).

http://dx.doi.org/10.1016/j.pmu.2014.10.001

Intelligence—can be simulated by a machine. Artificial Intelligence has now days become an essential part of the technology industry, providing the heavy lifting for many of the most difficult problems from all walks of daily life. These days, many efforts are being laid upon the development of models of diseases, using the synthetic intelligence to overcome the difficulties faced using the traditional rule based modeling techniques. Such intelligent models of diseases have resulted in significant progress in our understanding of the progression of various disorders, and thus helped in gaining more clinical expertise. Efforts to develop such programs have led to substantial progress in our understanding of clinical expertise, in the translation of such expertise into cognitive models, and in the conversion of various models into promising experimental programs. Of equal importance, these programs have been steadily improved through the correction of flaws shown by confronting them with various clinical problems.

1.2. Cervical cancer

Cervical cancer is a malignant tumor that occurs when cervical tissue cells begin to grow and replicate abnormally without controlled cell division and cell death. In such a state, the body is unable to use and manage such cells for carrying out their usual function resulting these cells transforming into a tumor. If the tumor is malignant, its cell flow through the blood stream and spread to other parts of body, as a result those parts also become infected. Usually the cervical cancer takes number of years to develop. These infected cells are then distinguished as cervical intra-epithelial neoplasia (CIN) or cervical dysplasia. The cells over the surface of cervix that show unusual changes & potentially precancerous developments are called CIN. In most of the cases CIN remains stable, or these are eliminated by host's immune system response. Although, a small percentage of cases progress to become cervical cancer, if not treated. Studies have found that CIN usually results from a virus called human papillomavirus (HPV) which is generally sexually transmitted. Although there are more than 120 types of HPV known [13], only 15 are classified as high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82) [14], 3 as probable-high-risk (26, 53, and 66), and 12 as low-risk (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and CP6108). In many cases even after getting infected with HPV, it is generally eliminated by the response of the host's immune system, but in many cases, where HPV is not done away with by the immune system, it may develop into cervical cancer. The common risk factors linked with cervical cancer include first intercourse at an early age, pregnancy at early age, having sex with multiple partners, weak immune system, smoking, use of oral contraceptives, improper menstrual hygiene etc. At an early stage the cervical cancer may be completely asymptomatic. The early stages of the cervical cancer are usually asymptomatic but symptoms do appear with the progression of pre-cancer to invasive cancer and typically shows abnormal vaginal bleeding, vaginal discharge, pain during vaginal intercourse. New bleeding may be experienced by the women who have had their menopause. Cervical cancer is the second most commonly diagnosed [15] and fifth deadliest [16] cancer in women throughout the world. In developing countries cervical cancer has a major share in cancer mortality [17]. Every year about 500,000 new cases are diagnosed and among which about 250,000 patient die. Because of poor access to screening and treatment services, approximately 80% of this disease occurs in women living in low- and middle-income countries [18].

1.3. Papanicolaou test

The Papanicolaou test (Pap smear) has been the widely used method in cervical cancer screening for many decades and has showed a dramatic lowering of incidents of cervical cancer and hence in related mortality rates in many countries [1] [17]. In taking a Pap smear, cells are scraped from the outer opening of the cervix for microscopic examination and to lookup for irregularities. The aim of the test is to detect any pre-cancerous or potentially pre-cancerous alterations called cervical intraepithelial neoplasia (CIN) or cervical dysplasia. Pap test is also used to detect endocervix and endometrium abnormalities and infections. In many developed countries, regular Pap smear screening is highly recommended for females who have had frequent sex with multiple partners [19]. If any unusual findings are observed the test may need to be repeated within a year. If the abnormality observed requires closer examination, a detailed cervical inspection by colposcopy may be done. HPV DNA testing may also be suggested to such patients, which acts as a supplementary to Pap smear testing. Once the sample is obtained, Papanicolaou technique is used to stain it. Staining using this technique helps to differentiate the cells in smear preparation from various other bodily secretions as unstained cells cannot be seen under a simple compound microscope. Most of the abnormal results are mildly abnormal (called low-grade squamous intraepithelial lesion (LSIL)) which indicates HPV infection. Most low-grade cervical dysplasia relapse on their own without usually causing cervical cancer, but presence of dysplasia can act as a warning that greater monitoring is needed. Generally, some Pap results are high-grade squamous intraepithelial lesion (HSIL), and very few of them indicate cancer.

2. Database for analysis

The database used in the study is Pap smear Benchmark database. The database consists of data about 1417 (500 + 917) Pap smear images and was prepared by the Herlev University Hospital, Denmark using a microscope and a digital camera. The university hospital prepared two databases, an old one in 2003 and a new one in 2005 containing 500 and 917 Pap smear images respectively. The database can be obtained from http://labs.fme.aegean.gr/decision/downloads. The Pap smear images in the database are unevenly classified into 7 different classes. Each of these samples is described using twenty (20) features obtained from images of each cell. These features include size of nucleus, size of cytoplasm, ratio of size of nucleus and cytoplasm, area of nucleus, area of cytoplasm, nucleus-cytoplasm area ratio, nucleus brightness, cytoplasm brightness, nucleus shortest diameter, nucleus longest diameter, nucleus elongation, nucleus roundness, cytoplasm shortest diameter, cytoplasm longest diameter, cytoplasm elongation, cytoplasm roundness, nucleus perimeter, cytoplasm perimeter, relative position of nucleus, maxima in nucleus, minima of nucleus, maxima in cytoplasm and minima of cytoplasm. Table 1 summarizes the number of different types of Pap smear images of each class in the database. Fig. 1 shows a sample images from the database. These Pap smear images were manually classified into seven (7) different classes by skilled cyto-technicians and doctors. Among these seven classes the first three i.e. 1, 2 & 3 are classified as normal cells whereas the last four i.e. 4, 5, 6 & 7 are classified as abnormal cells. The screening of these Pap smear images can be regarded as classification into normal and abnormal cells i.e. two-class problem and also as an elaborated classification into seven respective classes.

To ensure the accuracy each cell included in the database was inspected by two cytotechnicians, and complex samples were also subjected to examination by a doctor. In case of any difference of opinion in Pap smear reporting the sample was excluded from database. As such the final database so prepared contains diagnoses which are most precise and accurate. Ideally the specimens considered in the database have been taken from different portions of the cervix. The specimens mostly contain samples from the