International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 10 (2017), pp. 2365-2377
© Research India Publications
http://www.ripublication.com

A Probabilistic Approach to Detect Cervical Cancer using Protein Sequence

Jiji. N
*Associate Professor, Department of Computer Science and Engineering,
Younus College of Engg. & Tech., Pallimukku, Kollam, Kerala, India.*

Dr. T Mahalakshmi
*Principal, Sree Narayana Institute of Technology,
Kollam, Kerala, India.*

Abstract

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have more attention in cancer research and this will provide a great facilitate the subsequent clinical management of patients. Cervical Cancer is one of the most commonly affected cancers for most of the Women between age group of 31 to 45. Cervical cancer is a cancer arising from the cervix. It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body. The cervical cancer affected patients are classified into high or low risk groups have led many research teams, from the biomedical and the bioinformatics field. Most of the research peoples are working towards the cancer classification fields in the computational biology. In this paper, we develop a model for Cervical Cancer detection using Probabilistic Approach. In the proposed method, we have taken simple population based probabilistic approach for effective detection of Cervical Cancer by using proteinsequence. The proposed scheme is implemented with the created dataset from available information in CCDB and this dataset is used for training the system. We have presented the result and detection accuracy of the proposed scheme. The initial stage of evaluation provides good result for the early detection of cervical cancer using protein sequence.

Keywords: Cervical cancer, probabilistic model, Human Papilloma Virus HPV

I. INTRODUCTION

Cervical cancer is one of the most commonly affected cancer type in American, African and European ladies. Recently, In India marginal amount of women's are affected by cervical cancer. One of the most important virus for creating cervical cancer is Human Papilloma Virus (HPV). Persistent infection by certain DNA viruses, *i.e.*, the high-risk (carcinogenic) types of the genus alpha papillomavirus of the papillomaviridae family, commonly known as Human Papilloma Virus (HPV), is a necessary factor in the pathologic process which may lead to cervical cancer development [8–11]. Human Papilloma Virus (HPV) is one of the most common virus groups in the world today affecting the skin and mucosal areas of the body. Over 140 different strains of HPV have been identified. Papilloma viruses are highly species specific and do not infect other species, even under laboratory conditions. Humans are the only known reservoir for HPV. Different types of the human Papilloma virus are known to infect different parts of the body. The most visible forms of the virus produce warts (papillomas') on hands, arms, legs, and other areas of the skin.

Cervical cancer is highly preventable disease if detected at its precancerous stages and treated by ablative procedures. Since there are no treatments to cure the persistent HPV infections, prevention of cervical cancer is accomplished by physical removal of the cancer-susceptible transformation zone of the uterine cervical epithelium when a biopsy shows the presence of high-grade cervical intraepithelial neoplasia (CIN3), representing pathologic changes with a high frequency of progressing to invasive cancer. In the United States, the widespread use of Papanicolaou (Pap) smear screening for detection followed by treatment of these precancerous lesions reduced the incidence of cervical cancer from 44 in 100,000 women in 1947 to 8.8 in 1970 [13]. About 79 million American women are infected with HPV with about 14 million becoming newly infected each year [14]. In 2010, there were 11,818 American women diagnosed with cervical cancer, and 3939 cervical cancer deaths [15]. The age-standardized mortality rate is 1.7 per 100,000 [12]. Cervical cancer is primarily a disease among unscreened or rarely screened women [16].

The cervical cancer detection is one of the important research areas in the field of computational biology. Clinical data is recorded as a database and this may be used for early detection of cervical cancer by applying the test results. In recent years, most of the computer science research peoples are providing or extending their research towards multidisciplinary area. Biological based computational area is one such kind of research filed. We have large amount of research papers for detecting cancers in early stage by applying data mining tools, image analysis tools, prediction system and artificial intelligent applications.

In this paper, we have proposed a model for detecting cervical cancer in early stage by using probabilistic approach. The proposed scheme is modeled with two phases, training and testing phase. In the training phase, we have developed a dataset which contains cancer affected protein sequence as pattern. This dataset is usedto identify the affected or possibility of cancer affected protein sequence during the testing phase.

The remaining section of papers are organized as follows, section 2 provides some related works involved in the field of cervical cancer detection. In section 3, we proposed a probabilistic approach for detection of cervical cancer from the protein sequence. Section 4 provides result and discussions and section 5 provides conclusion and future work.

II. RELATED WORK

HPV structure

Human Papilloma viruses (HPV) belong to the Papovaviridae family. They consist of 72–capsomerecapside containing the viral genome. Capsomers are composed of two structural proteins: the 57 kD late protein L1, which accounts for 80% of the viral particle, and the 43-53 kD minor capside protein L2. The HPV genome consists of eight kilobasepairs(Kbp) and is a double-stranded DNA molecule. Arrangement of the 8-10 open reading frames (ORFs) within the genome is similar in all papilloma virus types and partly overlapping ORFs are arranged on a sole DNA strand. The genome can be divided into three regions: the long control region (LCR) without coding potential; the region of early proteins (E1E8), and the region of late proteins (L1 and L2)

Human papilloma virus infection and the disease, cervical cancer, is one of the leading cancers among women, which affects approximately 500,000 women each year, resulting in approximately 230,000 deaths worldwide. 85% have been reported from developing countries including India.

Worldwide, cervical cancer is both the fourth-most common cause of cancer and the fourth-most common cause of death from cancer in women [6]. In 2012, an estimated 528,000 cases of cervical cancer occurred, with 266,000 deaths [6]. This is about 8% of the total cases and total deaths from cancer [15]. About 70% of cervical cancers occur in developing countries [6]. In low-income countries, it is the most common cause of cancer death [13]. In developed countries, the widespread use of cervical screening programs has dramatically reduced rates of cervical cancer [16]. In medical research, the most famous cell line known as HeLa was developed from cervical cancer cells of a woman named Henrietta Lacks.

The inhibitory effect of natural compounds against E6 protein has been studied by few researchers. Advancements in computational chemistry and bioinformatics are very useful for the investigation of novel inhibitors from natural sources. The E6 protein of HPV-16 inactivates p53, therefore, the process of gene regulation is disturbed, which is a fundamental cause of cervical cancer.

Detection of HPV risk types can be a protein function prediction even though functions are described at many levels, ranging from biochemical function to biological processes and pathways, all the way up to the organ or organism level [17]. Many approaches for protein function prediction are developed by using similarity search between proteins with known function. The similarity among proteins can be defined in a multitude of ways [18], sequence alignment, structure match by common surface clefts or binding sites, common chemical features, or certain motifs comparison. However, the existing prediction systems cannot guarantee for providing good performance. Eom et al. [19] proposed a method using sequence comparison for HPV classification. They use DNA sequences to discriminate risk types based on genetic algorithms. Joung et al. [20] combined several methods for the risk type prediction from protein sequences. Protein sequences are first aligned, and the subsequences in high-risk HPVs against low-risk HPVs are selected by hidden Markov models. Then a support vector machine is used to determine the risk types. The biomedical literature can be used to predict HPV risk types [21].

A commercial HPV DNA detection kit was first introduced to detect high-risk HPV genotypes, *i.e.*, HPV-16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68 for cervical screening [11, 12]. Since 2006, two protein-based vaccines containing virus-like particles (VLPs) as the active ingredient have been available for prevention of the infection by the two most common high-risk HPV genotypes, *i.e.*, HPV-16 and HPV-18, with the aim to prevent cervical cancer in the subjects.

Preliminary Study of Probability

a. Probability of Event Occurrence

A sample space S in which all outcomes are equally likely is called a uniform sample space.

If S is a finite uniform sample space and E is any event, then the probability of E, P(E), is given by:

$$P(E) = \frac{Number\ of\ Ways\ E\ occurs}{Total\ Number\ of\ possible\ outcomes\ in\ S} = \frac{n(E)}{n(S)}$$

b. Mean Calculation

To find the mean value for the N number of events with different probability values, the following formula is used to find the mean value

$$\mu = \frac{\sum_{i=1}^{N} P(E_i)}{N}$$

Problem Statement: Cervical Cancer detection has become one of the most important area of research in the field of computational biology/bioinformatics. Because cervical cancer is highly treatable when detected early, researchers are developing better ways to detect pre cancer and cervical cancer. In this paper, we have proposed a method for the early detection of Cervical Cancer using probabilistic approach with event occurrence.

Outcome: The proposed method provides the result of probability value for the given input protein sequence by comparing it with the trained data set.

III. PROPOSED METHOD

In this paper, we have proposed a probabilistic model for cervical cancer detection using event occurrence. This scheme takes supervised protein sequencesfrom normal and abnormal category as the trained data set and is compared against a protein sequence from the user. This scheme has two phases, the Training Phase and the Testing Phase.

1. In *Training Phase*, the system will be trained with well studied and identified protein sequence affected by cervical cancer. This phase takes 500 samples as the training data set with four fields (protein sequence, protein ID, HPV Type and Status) Training is carried out with supervised learning method. The status field in the training dataset indicates the normal and abnormal protein sequence.
2. The *Testing Phase* takes an input protein sequence and compares it with the trained data set. Each trained protein sequence is split into a number of small blocks and calculates the probabilityvalue for the occurrence of a particular block in the given input test data. This process is carried out for both Normal and Abnormal categories.The following algorithm explains the steps involved in the testing phase.

Algorithm for Testing Phase

```
1. Initially both normal and abnormal threshold value is set as 0
2. For each protein sequence from the training data set,
   a. Split into K number of blocks with same size
   b. For each block,  calculate the probability of occurrence of the block in the
      given test input protein sequence by using the following formula
```

$$P(E) = \frac{Number\ of\ Times\ a\ block\ occurred\ in\ the\ Test\ Data\ Set}{Total\ Number\ of\ \ blocks\ in\ the\ Test\ Data} = \frac{n(E)}{n(S)}$$

   c. Find the mean value for the given test instance
   d. If the training data set is a normal and the mean value is greater than the normal threshold value then change the normal threshold as mean value; continue the process from step2
   e. If the training data set is an abnormal and the mean value is greater than the abnormal threshold value then change the abnormal threshold value as mean value; continue the process from step2

3. Compare the normal probability value and abnormal probability value.
   a. If normal probability value is high then the probability of Cervical Cancer is low.
   b. If abnormal probability value is high then the probability of Cervical Cancer is high.

The detailed algorithm for the testing phase is given,

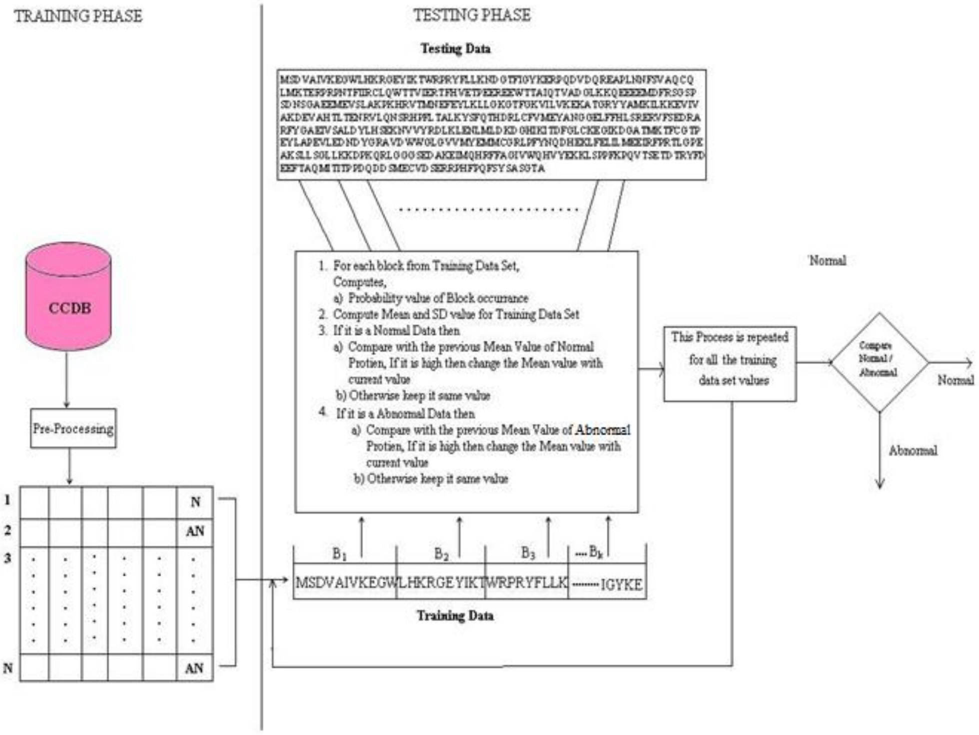
TRAINING PHASE
TESTING PHASE
Testing Data
CCDB
Pre-Processing
1
2
3
N
N
AN
AN
1. For each block from Training Data Set, Computes,
a) Probability value of Block occurrance
2. Compute Mean and SD value for Training Data Set
3. If it is a Normal Data then
a) Compare with the previous Mean Value of Normal Protien, If it is high then change the Mean value with current value
b) Otherwise keep it same value
4. If it is a Abnormal Data then
a) Compare with the previous Mean Value of Abnormal Protien, If it is high then change the Mean value with current value
b) Otherwise keep it same value
This Process is repeated for all the training data set values
Compare Normal / Abnormal
Normal
Normal
Abnormal
B_1
B_2
B_3
.... B_k
MSDVAIVKEGWLHKRGEYIKTWRPRYFLLK .... IGYKE
Training Data

Figure 1: Architecture for the proposed Approach

```
Input: Preprocessed Trained Data Set$TrainData[\ ][\ ]$ and Test Data Set$TestData[\ ][\ ]$
Output: Normal/Cancer Affected

$Normal_{threshold} = 0$
$AbNormal_{threshold} = 0$
$for\ (i = 1\ to\ N)$
Begin
   $for\ (j = 1\ to\ K)$
Begin
   $PE_j = \frac{Number\ of\ Times\ block\ TrainData[i][j] occurred\ in\ the\ TestData[\,]}{Total\ Number\ of\ blocks\ in\ the\ TestData[\,]}$
                              $Sum = Sum + PE_j$
End
$Mean = \frac{Sum}{K}$
     $if(TrainingData[i].Status == Normal\ \&\&\ Mean \geq Normal_{threshold})$
          $Normal_{threshold} = Mean$
     $if(TrainingData[i].Status == AbNormal\ \&\&\ Mean \geq AbNormal_{threshold})$
          $AbNormal_{threshold} = Mean$
End
$if(Normal_{threshold} \geq AbNormal_{threshold})$
       Protein Sequence Normal
Else
       Protein Sequence Abnormal
```

IV. IMPLEMENTATION AND RESULT DISCUSSIONS

The Cervical Cancer gene Database (CCDB) is a database of genes involved in the cervical carcinogenesis [1]. The Cervical Cancer Database is the first database that has been manually curated [2]. The database serves as an entity for clinicians and researchers to examine basic information as well as advanced information about the genes that differentiates into cervical cancer [3]. There are 537 genes that have been cataloged into the CCBD [4]. The genes that have been cataloged affect the polymorphism, methylation,amplification of genes, and the change in how the gene is expressed [4]. Science investigators have examined data that compared normal cervical cells with malignant cervical cells which has been used to study the different gene expressions that result in cervical cancer. Of the 500,000 women that have succumbed to cervical, most are from developing countries as well as of the low socioeconomic level in developed countries.

We have generated a dataset with minimum fields from CCDB and we have measured the cancer detection accuracy with different size of sampling test input with minimum normal and maximum abnormal value.The table1 illustrate the sample value in the trained dataset from the CCDB database. The training dataset is equipped with 200 instances for testing purposes.

Table 1: Sample Training Dataset

HPV Affected Protein Sequence	**HPV Type**	**Risk**	**Status**
-SAT.....EPH-DQ-...-T-A--.....--VC--PM-S---P-A----TVC	HPV54	?	?
---T......SASSQ-...ST-YQ-.....--DFGLT-RN--IC-IW--NH--	HPV32	Low	normal
-SGT......SASSQ-...-T-YQ-......---FGLT-RN--IS-IW---H--	HPV42	Low	normal
--V-M...SM-ANCPK...N-F-.-....-RNTG-GFD--R-H-I--T-Q--	HPV3	Low	normal
-D.........DQRPK...N-F-.-.............-RDSG--FD--R-H-I--A-V--	HPV28	Low	normal
-SM.......GAQ-PR...N---.-............-RNCG-P-E--R-C-I--T-Q--	HPV10	Low	normal
-SR.......GDGYPK...N-F-.-..........-RDSGVPFE--R-Q----T----	HPV29	Low	normal
-G.........-CNPT...N-F-.-................--DYEVDFE--R-T-----N---	HPV61	High	abnormal
-PM.......GLHNPT...N-W-.-.........---IEVD-E--RIT-I---N---	HPV72	High	abnormal
-HTRA....GMSE-N-CPRN-F-.-...---YGLE-E--R----W--RP-S	HPV2a	Low	normal
-RTRA....GMSE-N-CPRN-F-.-....--QYGLE-E--R----Y-RRA-S	HPV27	Low	normal
-SE.......ENPCPR...N-F-.-........-R-YGLE-E--RI---Y--RP-S	HPV57	?	?
-FE.......--R---...-T-HE-.........-ES--TT-QN--VQ--Y--ET-Q	HPV26	?	?
-FE.......-KR---...-T-HE-..........-EA--V-M-NI-VV--Y-----C	HPV51	High	abnormal
--FKF....ENTG---...-TVHH-.....-EVQET--LE---Q--Y-----S	HPV30	Low	normal

During the testing phase, the testing protein sequence is divided into K number of blocks form 1,…,K. The comparison process is carried out for each block and find the number of times occurred in the testinstance. The threshold values for both normal and abnormal instances are calculated. Initially the $Normal_{threshold}$ and$AbNormal_{threshold}$is set as a 0 and the value changes during the testing phase execution. Finally, the $Normal_{threshold}$ and$AbNormal_{threshold}$ value is compared and the status of the test instant is considered as which one is greater.

For the result comparison, the time taken for finding the given test instance as a Normal or Abnormal is considered. The time comparison phase is done with 2 different samples. In the first sample, out of500 samples for the CCDB data base and 50 are taken as the test instances and aretested during the testing phase. In the second sample, 100 test instances are takenas the test instances for testing. The proposed scheme is implemented using simple Core JAVA with MySQL as backend for training dataset storage. We have used simple computational facility for implementing the proposed scheme. This scheme needs minimum computational cost and minimum time duration for Cervical Cancer identification. This scheme can be used for any kind of hospitals or testing clinics. The following table 2, illustrate the time comparison for two different samples

Table 2: Accuracy and Execution Time taken for the Proposed Approach with different Sample Size (50, 100, 150 and 200)

Sample Size	**Number of**		**Number of Accurately Identified**		**Accuracy %**	**Execution Time in ms**
	Normal	**Abnormal**	**Normal**	**Abnormal**		
50	18	32	16	30	92%	$\approx 12\,ms$
100	34	66	31	63	94%	$\approx 16\,ms$
150	54	96	49	89	92%	$\approx 20\,ms$
200	74	126	69	119	94%	$\approx 25\,ms$

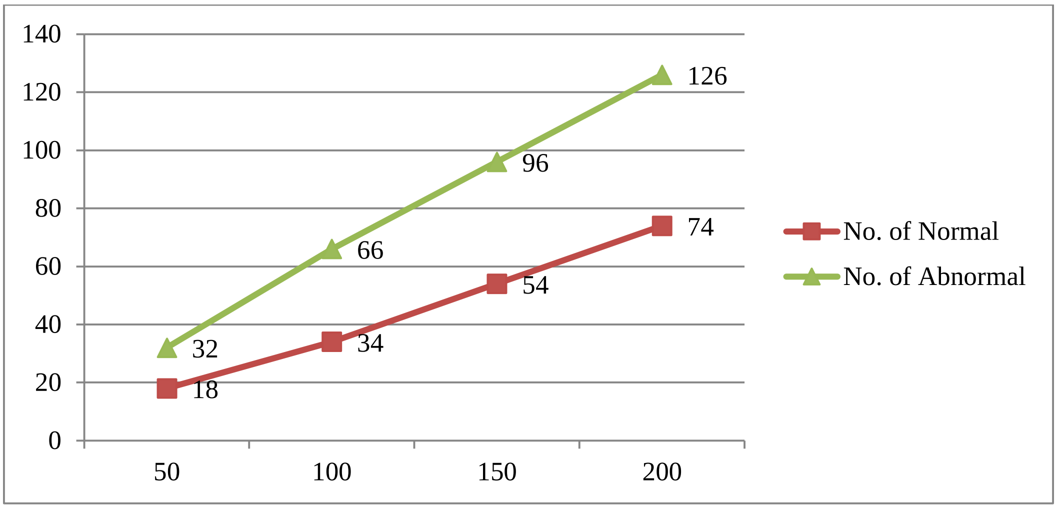

Figure 2: Number of Test Samples with Normal and Abnormal Instances

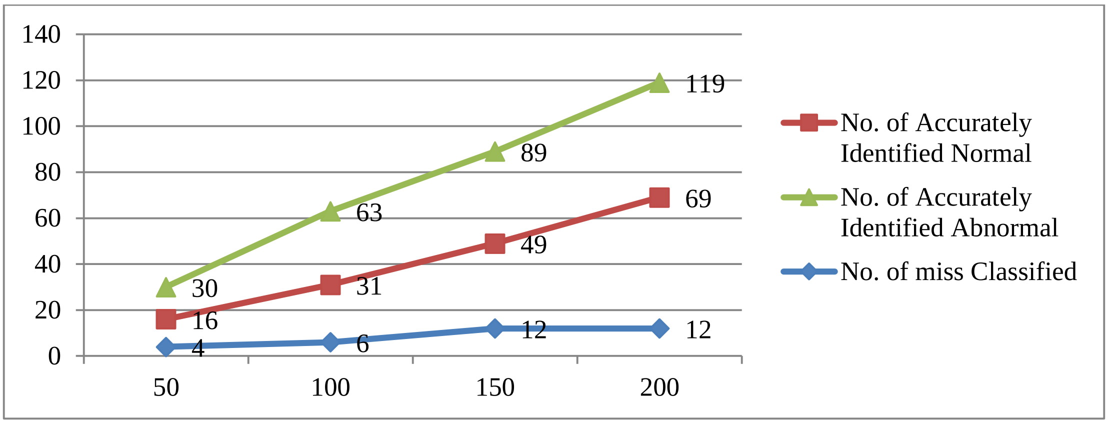

Figure 3: Accurate and Miss Classification

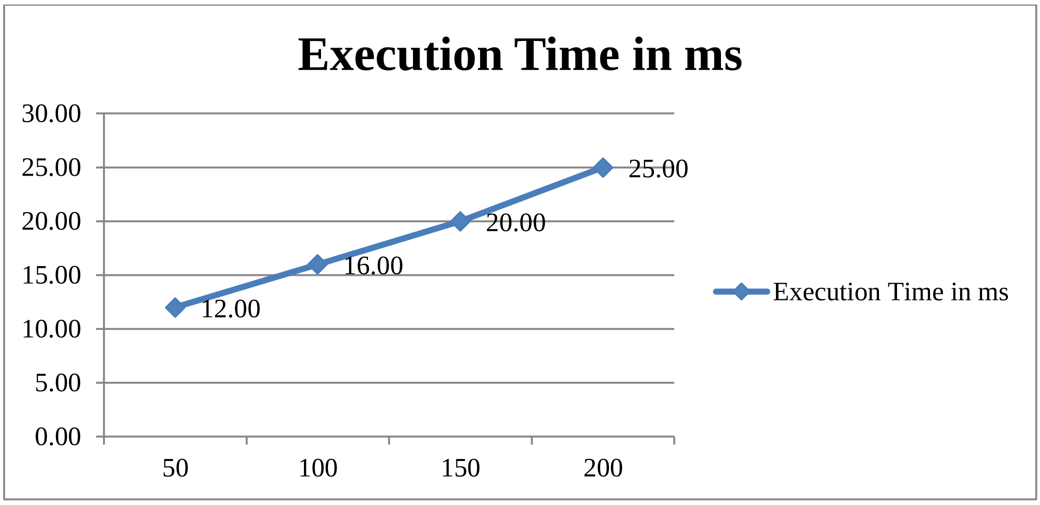

Figure 4: Execution Time taken for different sample size

V. CONCLUSION AND FUTURE WORK

Infection by the Human Papillomavirus (HPV) is associated with the cervical cancer affection. HPV can be classified into high and low-risk type according to its malignant potential. The detection of the risk type is important to understand the mechanisms and to diagnose potential patients. The risk identification is one of the most wanted research area in the field of bio-computation. In this paper, we have proposed a probabilistic approach for cervical cancer detection in early stages using protein sequence. The proposed scheme can detect the cervical cancer by using the HPV affected protein sequence. We have used the protein sequence with normal and abnormal instances as the trained dataset. Test instances are classified into normal or abnormal by comparing it with the training dataset. Threshold value is used for finding the normal and abnormal test instance. The threshold value will be changing for every new probability value. The proposed scheme needs minimum computational facility and it takes minimum time for classification. In future, a proper dataset can be developed from the available information of the CCDB database and clustering/classification algorithms can be used for getting better result.

REFERENCE

[1] Agarwal, Subhash M. RaghavDhwani, Singh Harinder, Raghava G P S (Jan 2011). “CCDB: a curated database of genes involved in cervix cancer”, Nucleic Acids Res. England. 39 (Database issue): D975–9.

[2] Agarwal, S M, Agarwal, S.M., Raghav, D., Singh, H., Raghava, G. P. S. (2010). “CCDB: a curated database of genes involved in Cervix Cancer Nucleic Acids Res” GPS.

[3] Hakim, AA, Lin PS; Wilczynski S, Nguyen K, Lynes B, Wakabayashi MT. (2010). “Indications and efficacy of the human papillomavirus vaccine” Current Treat Options Oncol. vol. 8, No.6, pp. 393–401.

[4] Phongsavan, K, Phengsavanh A,Wahlström R, Marions L. (2010) “Women's perception of cervical cancer and its prevention in rural laos”. **20** (Int. J. Gynecol. Cancer): pp. 821–826.

[5] Agarwal, S.M., Raghav, D., Singh, H. and Raghava, G. P. S.. (2011) CCDB: a curated database of genes involved in Cervix Cancer Nucleic Acids Research (2011, 39, D975-D979)

[6] ZurHausen H.: Papillomaviruses in human cancers. Proc Assoc Am Phys 1999; 111: 1-7.

[7] Howley PM, Lowy DR. Papillomaviruses and their replication. In: Knipe DM, Howley PM, eds. Fields Virology. Vol. 2. Philadelphia: Lippincott, Williams

and Wilkins; 2001; 2197-2230

[8] Wallin, K.L.; Wiklund, F.; Angstrom, T.; Bergman, F.; Stendahl, U.; Wadell, G.; Hallmans, G.; Dillner, J. Type-specific persistence of human papillomavirus DNA before the development of invasive cervical cancer. *N. Engl. J. Med.*1999, *341*, 1633–1638.

[9] Kjaer, S.K.; van den Brule, A.J.; Paull, G.; Svare, E.I.; Sherman, M.E.; Thomsen, B.L.; Suntum, M.; Bock, J.E.; Poll, P.A.; Meijers, C.J. Type specific persistence of high risk human papillomavirus (HPV) as indicator of high grade cervical squamous intraepithelial lesions in young women: Population based prospective follow up study. *BMJ*2002, *325*, 572–676.

[10] Cuschieri, K.S.; Cubie, H.A.; Whitley, M.W.; Gilkison, G.; Arends, M.J.; Graham, C.; McGoogan, E. Persistent high risk HPV infection associated with development of cervical neoplasia in a prospective population study. *J. Clin. Pathol.*2005, *58*, 946–950.

[11] Brummer, O.; Hollwitz, B.; Bohmer, G.; Kuhnle, H.; Petry, K.U. Human papillomavirus-type persistence patterns predict the clinical outcome of cervical intraepithelial neoplasia. *Gynecol. Oncol.*2006, *102,* 517–522

[12] WHO Regional Office for Africa. Cervical Cancer. Available online: http://www.afro.who.int/en/ clusters-a-programmes/dpc/non-communicable-diseases-managementndm/programme-components/ cancer/cervical-cancer/2810-cervical-cancer. (accessed on 2 May 2014).

[13] Guzick, D.S. Efficacy of screening for cervical cancer: A review. *Am. J. Public Health*1978, *68*, 125–134.

[14] CDC. Genital HPV Infection-Fact Sheet, Page last updated: 20 March 2014. Available online: http://www.cdc.gov/std/hpv/stdfact-hpv.htm (accessed on 2 May 2014).

[15] CDC. Cervical Cancer Statistics, Page last updated: 20 December 2012. Available online: http://www.cdc.gov/cancer/cervical/statistics/ (accessed on 2 May 2014).

[16] Janerich, D.T.; Hadjimichael, O.; Schwartz, P.E.; Lowell, D.M.; Meigs, J.W.; Merino M.J.; Flannery, J.T.; Polednak, A.P. The screening histories of women with invasive cervical cancer, Connecticut. *Am. J. Public Health***1995**, *85*, 791–794

[17] Watson, J. D., Laskowski, R. A., and Thornton, J. M.: Predicting Protein Function from Sequence and Structural Data. Current Opinion in Structural Biology 15 (2005) 275–284.

[18] Borgwardt, K. M., Ong, C. S., et al.: Protein Function Prediction via Graph

Kernels. In Proceedings of Thirteenth International Conference on Intelligenc Systems for Molecular Biology (2005) 47–56.

[19] Eom, J.-H., Park, S.-B., and Zhang, B.-T.: Genetic Mining of DNA Sequence Structures for Effective Classification of the Risk Types of Human Papillomavirus (HPV). In Proceedings of the 11th International Conference on Neural Information Processing (2004) 1334–1343.

[20] Joung, J.-G., O, S.-J, and Zhang, B.-T.: Prediction of the Risk Types of Human Papillomaviruses by Support Vector Machines. In Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence (2004) 723–731.

[21] Park, S.-B., Hwang, S., and Zhang, B.-T.: Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost. In Proceedings of the 14th International Conference on Database and Expert Systems Applications (2003) 403–412.