THE PREDICTION OF CERVICAL CANCER OCCURENCE USING GENECTIC ALGORITHM AND SUPPORT VECTOR MACHINE

*Abisoye, O.A[1], Abisoye,B.O[2], Ekundayo Ayobami[3] & Kehinde Lawal[4]
[1]Computer Science Department, SICT, Federal University of Technology, PMB 65 Minna Niger State, Nigeria
[2]Computer Engineering Department, SEET, Federal University of Technology, PMB 65 Minna Niger State, Nigeria
[3]Computer Science Department, SICT, Federal University of Technology, PMB 65 Minna Niger State, Nigeria
[4]Computer Science Department, SICT, Federal University of Technology, PMB 65 Minna Niger State, Nigeria
*Corresponding author email: o.abisoye@futminna.edu.ng, +23460546074

ABSTRACT

Cervical cancer is a malignant neoplasm arising from cells originating in the cervical uteri. Cervical cancer can be treated using Human Papilloma virus vaccine and carrying out regular pap test. The manual system contains large amount of errors by virtue of human decision, the visual screening is very demanding, tedious, and expensive in terms of labor requirements. This paper proposed machine learning algorithm; Support Vector and Genetic algorithm to predict the occurrence of cervical cancer. Evaluation results show the effectiveness of the proposed approach with the overall Precision, Recall, F1 score, Sensibility, Sensitivity, Accuracy values 96%, 95%, 95%, 89%, 96%and 95% respectively for Biopsy and 97%, 96%, 96%, 50%, 97% and 96% for Hinselmann.

In this study cervical cancer was predicted with Support vector machine classifier and Genetic algorithm optimization tool. The prediction was found to have acceptable performance measures which will reduce future incidence of the outbreak in the world and aid timely response of medical experts.

Key Words: Cancer, Classification, Extraction, Human papillomavirus, Prediction,

1 INTRODUCTION

Human papillomavirus is a virus responsible for the cause of Cervical cancer. Symptoms of cervical cancer can include painful sex, vaginal bleeding, and discharge (CDC, 2016). A Cervical cancer risk factor is any means that changes thE possibility of contacting cervical cancer. **Cervical cancer** is a type of cancer that develops from the cervix. Other types of cancer includes breast cancer, lungs cancer, sarcoma cancer, leukaemia cancer, liver cancer (Idikio, 2011). It is due to the abnormal growth of cells that have the ability to spread to other parts of the cervix. Early on, typically no symptoms are seen. Some of the symptoms that precede later in the growth of abnormal cells include vaginal discharge of blood, pelvic pain. In as much as bleeding after sex Is not really considered a symptom of cancer because it can be caused by several diseases, it can also indicate the presence of cervical cancer (Zhang & Liu, 2004).

This cancer is preventable by screening for premalignant lesions but this is rarely provided and hardly utilized. Considering the main risk of cervical cancer result from human papillomavirus, the vaccine for this virus is required to prevent against it. People with weak immune system are at risk of developing some types of cancer. Weak immune systems include system that have had organs transplant, HIV, have consistently take drugs (UICC). Unlike the developed countries other screening methods can be explored in countries with low resources.

The diagnosis of cervical cancer using an artificial intelligent system with statistics by (Chandraprabha & Singh, 2016) reveals that the percentage of cervical cancer rising in developing countries is 70%. It is the major cause of death in low-income countries. The considered techniques in this research work are image processing and classification techniques. The research work reviewed various algorithms required for the diagnosis of primary features required for classification of cervical cancer (Chandraprabha & Singh, 2016).

2 RELATED WORKS

The research work by (Athinarayanan, Srinath, & & Kavitha, 2016) uses a screening method which is the pap smear test and several classification techniques for detecting cervical cancer like support vector machine (SVM), fuzzy based techniques and texture classification to differentiate the benign and cancerous cells.

This research reveals an expert system designed for predicting cervical cancer using data mining techniques.it explains the different methods of controls and prevention of cervical cancer which includes Pap smear, Human Papilloma virus screening and vaccination against Human Papilloma virus, liquid-based cytology. With data processing and manipulation of data taking the lead in our system as a result of large volume of data, data mining techniques are emerging. The data mining techniques used include features selection and classification (Benazir & Nagarajan, 2018).

A paper reviewed the classification of cervical cancer using artificial neural networks.it uses artificial neural network to detect cervical cancers by classifying the cells either as normal or abnormal cells in the cervix area. This classification produces an accurate result compared to Pap smear test which is a manual screening method. Analysis was made using all the network architecture which includes single neural network, deep and shallow neural network (Devi, Ravi, Vaishnavi, & S, 2016).

A research work reviewed the manual method of screening cervical cancer using Pap smear test, the need for carrying out feature selection and the classification of the features. The classification are done using K-NN and Artificial neural network. The analysis shows that the classification accuracy for K-NN was 88.04% and 54% for artificial neural network (Malli & Nandyal, 2017).

2.1 STATEMENT OF THE PROBLEM

There is a general understanding by the public that precise tool called pap smear is responsible for cancer detection has led many to believe that cancer after a normal Pap smear must imply malpractice. This understanding is a costly assumption hence pervasive. The limited number of experts and the large number of patients resulted in a long queue for the screening process. The manual system contains large amount of errors by virtue of human decision. The visual screening is very demanding, tedious, and expensive in terms of labour requirements.

3.0 PROPOSED METHODOLOGY

Several papers have revealed the need for early detection of cervical cancer owing to the fact that it is one of the deadliest cancer in the world. This paper discusses how cervical cancer data is pre-processed to eliminate noise, selected suitable features from the data collected and prepare the data for classification, and how SVM is used in training the system using the prepared data in classifying cervical cancer.

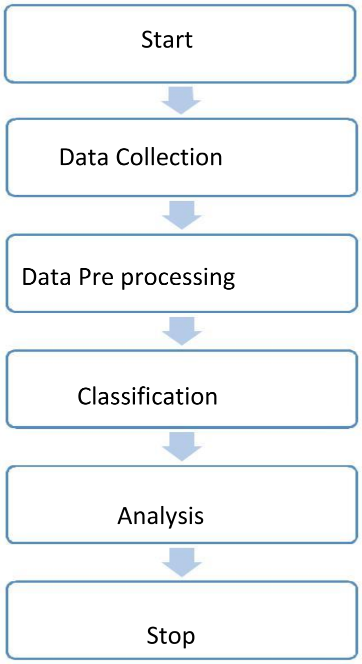

Figure 1: Block diagram for cervical cancer prediction

3.1 Data Collection

Data collection refers to the process of gathering data on a particular are of interest that are measurable and can be statically fashioned. The main goal of data collection is to capture data that is of high quality and its analysis would lead to the formation of answers that are convincing and credible. Figure 1 explains the methodology required for the predicting cervical cancer which includes data collection, preprocessing, classification and analysis.

The data collected for this research was gotten from UCI Machine Learning Repository. UCI Machine Learning Repository is a database or data generator that collects and store data relevant to the machine learning and statistics. The attributes of the cervical cancer data set can be viewed in the table 1.

Table 1: Attributes of the cervical cancer data set

S/N	Attributes	Description/ Data types
1	Age	Integer value
2	Number of sex Partners	Integer value
3	First sexual Intercourse	Integer value
4	Number of Pregnancies	Integer value
5	Smokes	Integer value
6	Smokes(year)	Integer value
7	Smokes(packs/year)	Integer value
8	Hormonal Contraceptives	Boolean value
9	Hormonal Conctraceptives (years)	Boolean value
10	Condylomatosis	Boolean value
11	Cervical Condylomatosis	Boolean value
12	Vaginal Condylomatosis	Boolean value
13	Vulvo- perineal Condylomatosis	Boolean value
14	Syphilis	Boolean value
15	Pelvic inflammatory Disease	Boolean value
16	Genital herpes	Boolean value
17	Molluscum contagiosum	Boolean value
18	AIDS	Boolean value
19	HIV	Boolean value
20	Hepatitis B	Boolean value
21	Number of Diagnosis	Integer value
22	Time since first Diagnosis	Integer value
23	Time since last Diagnosis	Integer value
24	Dx:CIN	Boolean value
25	Dx:HPV	Boolean value
26	Dx	Boolean value
27	Dx:Cancer	Boolean value
28	bool) IUD	Boolean value

29	IUD (years)	Integer value
30	STDs (number)	Integer value
31	Dx:Cancer	Boolean value
32	HPV	Boolean value
33	Hinselmann	Boolean value
34	Schiller	Boolean value
35	Cytology	Boolean value
36	Biopsy	Boolean value

3.2 Data Pre processing

Data pre-processing is a technique which involves the transformation of raw data into a more understandable and correct format. Data processing in a broad sense prepares the raw data for further processing. It aims to remove inconsistencies, missing values and error which are all contained in various real world data (El-Halees, 2008).

Data pre-processing is a very important in data mining because the value of the data rest on the feature decision.to improve the feature of medical diagnosis it is very important we improve the medical database.to check the accuracy of the diagnosis, it is required to check the attributes because the computation time is dependent on the number of attributes (Khare & Burse, 2016).

3.3 Feature selection

The data acquired from the data collection process is subjected to an optimization techniques called Genetic Algorithm to select the best fit for the classification process. Feature selection is a key tool to a successful data mining.

Genetic algorithm is a prominent example of evolutionary computation techniques. The evolutionary systems use optimization tool like evolution to solve engineering problems (Singh, 2013). This algorithm is a random search technique that is guided by genetics in natural evolution.

3.4 (Data Scaling)

Data normalization provides a better model and avoids numerical problems. Any statement problem required for pre-processing cannot eliminate the normalization stage because the data are manipulated or scaled before considered for the next stage. In this pre-processing stage, we can use the existing range to find the new range (Patro & Sahu, 2015). The normalization technique used was this research Min-Max normalization other types of normalization techniques includes, Z-score normalization, Decimal scaling normalization and Integer normalization. In this study a total number of thirty six(36)attributes were collected with 858 instances.

3.6 GENETIC ALGORITHM

Genetic algorithm is a search optimization technique that is based on Darwinian principle of evolution and natural selection solution generated by genetic algorithm is called a chromosome, while collection of chromosome is referred as a population The chromosome is made up of genes whose value can either be numerical, binary, symbols or characters. The processes in genetic algorithm includes; population initialization, selection, two point cross over, mutation, modification etc. as viewed in Figure 2

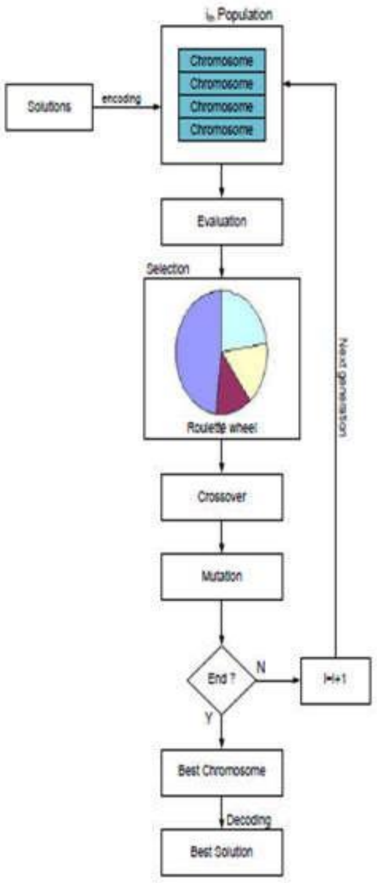

Figure 2: Genetic algorithm flowchart

In the genetic search the configurations were in Table 2 were put into considerations to select best features needed for good prediction.

Table 2: Genetic Search Features

Search Method: Genetic search.
Population size
Number of generations
Probability of crossover
Probability of mutation
Report frequency
Random number seed
Attribute Subset Evaluator

3.4 Training and testing the model.

Table 2 shows the attributes of cervical cancer that was selected from the genetic algorithm process. The selected features was used at the training and testing phase. SVM (Support Vector Machine) was used to create a model that will classify a trained dataset as either cancerous or non-cancerous.

Table 3: The Attributes of Cervical Cancer data set

S/n	Attributes	Data Types
1.	Hormonal Contraceptives	Boolean
2.	STDs: vaginal condylomatosis	Boolean
3.	STDs: genital herpes	Boolean
4.	STDs: molluscum contagiosum	Boolean
5.	STDs:Hepatitis B	Boolean
6.	STDs:HPV	Boolean
7.	STDs: Number of diagnosis	Boolean
8.	Dx:Cancer	Boolean
9.	Dx	Boolean
10.	Biopsy:	target
11.	Cytology:	target
12.	Schiller:	target

Table 3 is a list of some of the attributes in the cervical data set with its values. The IDE (Integrated Development Environment) used for the Genetic Search Attribute Subset Evaluator in WEKA. Out of existing 36 features, 12 features were selected as the best features for prediction.

3.7 SUPPORT VECTOR MACHINE

The support vector machine was invented in 1995 by Vapnik and Cortes for binary classification. It is a supervised learning model and although it has undergone many modification in its basic form the support vector machine classifies labelled data into two classes. Using separating hyper plane. The hyper plane is such that it is the farthest away from both classes of data to be classified. (Russell S & P, 2010) describes three characteristics that the Support vector machine attractive:

1. It constructs a decision boundary with the largest possible distance to example points (maximum margin separator) which helps it generalize well.
2. The possess the ability to embed data into a higher-dimensional space using something called the kernel trick
3. They combine the advantages of parametric and non-parametric models and have flexibility to represent complex functions.

Support vector machine is used for both classification and regression problem. The support vector machine is usually trained using a learning algorithm.

SVMs usually employ a learning algorithm for classification; our approach to classifying our dataset will involve the use of genetic algorithm for the extraction and selection of robust features for the SVM classifier

4.0 Results and Analysis

4.11 Genetic search Result for Biopsy

Search Method: Genetic search.
Population size: 20
Number of generations: 20
Probability of crossover: 0.6
Probability of mutation: 0.033
Report frequency: 20
Random number seed: 1
Attribute Subset Evaluator (supervised, Class (numeric): 36 Biopsy):CFS Subset Evaluator
Selected attributes for Biopsy:

- Hormonal Contraceptives (years)
- STDs:vaginal condylomatosis
- STDs:genital herpes
- STDs:molluscum contagiosum
- STDs:Hepatitis B
- STDs:HPV
- STDs: Number of diagnosis
- Dx:Cancer
- Dx
- Hinselmann
- Schiller
- Citology

4.2 Result for Biopsy from Classification

This section presents the result when the model was tested to classify a total of 223 instances.

Table 4: The confusion matrix of the SVM for Biopsy

	Predicted class	
	Non-Cancerous	Cancerous
Non-Cancerous	195	9
Cancerous	2	17

Table 4 shows the confusion matrix from the SVM classification on the Biopsy test, with 195 true negative classification, 17 true positive classification, 9 false positive calculation and 2 false negative calculation.

Table 5: The performance metrics of the classification for the Cancerous and Non- Cancerous class

Metrics	**Non-Cancerous**	**Cancerous**	**Average**
Precision	0.650	0.99	0.96
Recall	0.890	0.960	0.950
F1-score	0.760	0.97	0.95

Classification has a high precision, recall and F-score for the Cancerous class compared to the Non-Cancerous class.

Table 6 The sensitivity, specificity and accuracy.

Metrics	Cancerous
Sensitivity	0.890
Specificity	0.960
Accuracy	0.950

The sensitivity in Table 6 shows that the model recognizes 89% of the Cancerous instances in the test dataset. Specificity 96 % on the other hand shows how the model was able to differentiate among Cancerous and Non-Cancerous instances in the entire test Dataset. The total accuracy of the model is 95%. Which is moderately acceptable for the imbalanced dataset.

4.3 Genetic search Result for Hinselmann

Search Method: Genetic search.
Population size: 20
Number of generations: 20
Probability of crossover: 0.6
Probability of mutation: 0.033
Report frequency: 20
Random number seed: 1
Attribute Subset Evaluator (supervised, Class (numeric): 33 Hinselmann): CFS Subset Evaluator

Including locally predictive attributes
Selected attributes: 2, 24, 34, 36: 4

- ❖ Number of sexual partners
- ❖ STDs: Hepatitis B
- ❖ Schiller
- ❖ Biopsy

4.4 Result for Hinselmann

This section presents the result when the model was tested to classify a total of 220 instances.

Table 6: The confusion matrix of the SVM for Hinselmann

	Predicted class	
	Non-Cancerous	**Cancerous**
Non-Cancerous	208	6
Cancerous	3	3

Table 6 above shows the confusion matrix from the SVM classification on the Hinselmann test, with 208 true negative classification, 6 true positive classification, false positive calculation and 3 false negative calculation.

Table 7: The performance metrics of the classification for the Cancerous and Non-Cancerous class

Metrics	Non-Cancerous	Cancerous	Average
Precision	0.330	0.99	0.970
Recall	0.500	0.970	0.960
F1-score	0.400	0.980	0.960

Classification has a high precision, recall and F-score for the Cancerous class compared to theNon-Cancerous class.

Table 8: The sensitivity, specificity and accuracy.

Metrics	Cancerous
Sensitivity	0.500
Specificity	0.970
Accuracy	0.960

The sensitivity 50% shows tht the model could detect only averagely cancerous cases in the entire dataset. Specificity 97 % on the other hand shows how the model was able to differentiate among Cancerous and Non-Cancerous instances in the entire test Dataset. The total accuracy of the model is 96% which is not acceptable for the imbalanced dataset.

5.0 Conclusion

Machine Learning techniques have proven to be of great tools in various sects, there has been quite a number of research works in cervical cancer owing to the fact that is one of the deadliest cancer disease in the world that have taken hold of rural areas and developing countries particularly. In this study cervical cancer was predicted with Support vector machine classifier and Genetic algorithm optimization tool. The prediction was found to have acceptable performance measures which will reduce future incidence of the outbreak in the world and aid timely response of medical experts.

REFERENCES

Athinarayanan, S., Srinath, M. V., & Kavitha, R. (2016). Detection and Classification of Cervical Cancer in Pap Smear Images using EETCM , EEETCM & CFE methods based Texture features and Various Classification Techniques,. *2*(5), 533 - 549.

Benazir, B., & Nagarajan, A. (2018). An Expert System for Predicting the Cervical Cancer using Data Mining Techniques, 118(20), 1971–1987. *118*(20), 1971–1987.

CDC. (2016, January 2). www.Cdc.Gov/Cancer/Knowledge 800-CDC-INFO,.

Chandraprabha, R., & Singh, S. (2016). Artificial Intelligent System for Diagnosis of Cervical Cancer : a Brief Review and Future Outline. 38-41.

Devi, M. A., Ravi, S., Vaishnavi, J., & S, P. (2016). Classification of Cervical Cancer using Artificial Neural Networks. 465-472. doi:https://doi.org/10.1016/j.procs.2016.06.105

El-Halees, A. (2008, February). Mining Students Data to Analyze Learning Behavior: a Case Study Educational Systems. Work. doi:https://doi.org/10.1504/IJTEL.2012.051816

Idikio, H. A. (2011). Human cancer classification: A systems biology-based model integrating morphology, cancer stem cells, proteomics, and genomics. *Journal of Cancer, 2*(1), 107 - 115. doi:htps://doi.org/10.7150/jca.2.107

Khare, P., & Burse, K. (2016). Feature Selection Using Genetic Algorithm and Classification using Weka for Ovarian Cancer. *International Journal of Computer Science and Information Technologies (IJCSIT), 7*(1), 194–196.

Malli, P. K., & Nandyal, S. (2017). Machine learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network. *6*(4).

Patro, S. G., & Sahu, K. K. (2015). Normalization, A preprocessing stage. *Iarjset*, 20-22. doi:https://doi.org/10.1017/SO269888900007

Singh, S. R. (2013). Genetic Algorithms for Staging Cervical Cancer.3(Ii), 39–43. 39-43.

UICC. (n.d.). *Cancer Classification.* Media, Factsheet. doi:https://doi.org/10.1002/0471684228.egp01697

Zhang, J., & Liu, Y. (2004). Cervical Cancer Detection Using SVM Based . *2*, 873– 880.