A WEB-BASED DIAGNOSTIC FRAMEWORK FOR A KNOWLEDGE - CENTRIC CLINICAL DECISION SUPPORT SYSTEM FOR CERVICAL CANCER

Temitope O. Efuwape, Yinka A. Adekunle

Babcock University, Ilisan, Ogun State, Department of Computer Science

Corresponding Author: Temitope O. Efuwape, seunwape@gmail.com

ABSTRACT: Cervical cancer related mortality in developing countries like Nigeria is alarming and coupled with the incessant industrial dispute in the medical industry, early detection and treatment of the deadly disease is arduous hence the mortality rate. In this paper, we develop a framework of a knowledge-based diagnostic support for cervical cancer using the instrumentality of predictive analytics by deploying Iris dataset for the training of four base learners. This is aimed at presenting a proactive measure towards early detection and diagnosis of the menace via a web-based use case. Experimental result returned decision tree as the best learner after the performances of K-Nearest Neighbour, Naïve Bayes and Support Vector Machine were tested. The resulting model was built adopting the spiral software engineering framework for the diagnosis system which is deployed on a web based platform.
KEYWORDS: Cervical Cancer, Machine Learning, Clinical Decision Support System, Diagnosis, Decision Tree

1. INTRODUCTION

Cancer is nowadays ubiquitous as occasioned several cases daily reported in news media and it is generally been described as an uncontrolled growth of unusual cells in the body of its host and as noted by the National cancer Institute (2015), its varieties can be categorized into Carcinomas, Sarcomas, Leukemia, Lymphomas and Melanomas cancers. Cancer is often described with reference to the part of the body it attacks in women, breast, colorectum, breast, lung, cervical(cervix uteri), thyroid, stomach, vaginal, vulvar cancers are feminine prone as breast cancer is the most diagnosed among female worldwide with 1.7 million cases established in 2012 [1]. Colorectal cancer is regarded as the second most prominent women world over but with low incidence cases in Nigeria. This is mainly attributed to the rich fiber intake and deficiency in pre-malignant colorectal lesions [2].

For cervical cancer which has been established through epidemiology studies as caused by the sexually transmitted infection Human Papilloma Virus (HPV), especially the 16 and 18 HPV which are found in over 99% of cervical cancers highly regarded with high risks [3]. Women are believed to be generally infected in their teens and 20s but takes up to 10 or 20 years after initial infection of HPV before full blown cancer [4] and symptoms include postcoital bleeding, bleeding during menstrual cycle and pregnancy, painful sex etc. [5]. The foregoing activities towards the diagnosis of cervical cancer are primarily hectic, time consuming and can be expensive. Furthermore, the vaccines for the treatment of cancer are very expensive. The organisation as well as the huge materials and human resources required for mass screening are lacking in developing countries. Cytological screenings via Pap smears, Visual Inspection with acetic acid, HPV DNA testing, Visual inspection with Lugol's iodine are known accurate methods to early detection and diagnosis for cervical cancer which have contributed in no small measure to the reduction of its prevalence as data from countries have shown that screening with cytology tames incidence and mortality rate [6]. The diagnosis procedure including (i) cervix examination with colposcope special magnifying instrument, (ii) taking sample of cervical cells, (iii) ejection of a cone-shaped jurisdiction of cervical cells (iv) imaging tests, (v) visual examination of the bladder and rectum etc. are complex making the availability of a medical expert indispensable. Consequent upon the foregoing, aforementioned approach is not only hectic and time-consuming but indeed expensive considering the huge human and material resources required for mass screening, which are obviously lacking in developing countries like Nigeria [4].

The implementation of a CDSS for the early and efficient diagnosis of cervical cancer promises a reliable expert system with a utilitarian value for the treatment of cervical cancer especially in Nigeria

where prevalence of cervical cancer is rife and un-abating. Cervical Cancer has been termed second prevalent cancer among women in Nigeria, in fact, the fourth commonest among women in the world [1]. Besides several cervical diagnostic studies in literature, these studies however have not achieved satisfactorily good and convincing results of diagnosing and abating the number of false positives and false negatives of the deadly menace. Existing studies have not taken the pain to incorporate germane machine learning techniques like training set resampling, parameter optimization, reduction of dimensionality etc. while predetermining the best learner algorithm for detecting the disease, hence the prevalent hazy diagnosis results obtainable and the consequent un-abating mortality rate. This study attempts several stages towards the development of cervical cancer diagnostic system including a first stage of thoroughbred feature engineering to obtain the optimal training set and as well as determining the most suitable learner algorithm using the Rapid Miner tool to design a diagnostic predictive analytics framework. A web-based diagnostic Clinical Decision Support System suffixes to achieve proposed early detection of cervical dysplasia; so as to reduce the incidence and mortality rate of its victims in Sub-Saharan regions of Africa. The instrumentality of predictive analytics comes handy for an efficient classification of the image data corpus which is germane for any image processing model in the machine learning use case [8]. The goal of this work is to develop a knowledge-based diagnostic support framework for cervical cancer. The work will facilitate the early detection and prevention of cervical cancer in developing countries and in particular Nigeria.

2. METHODOLOGY

2.1. Datasets

The cervical cancer datasets used for the classification experiment was acquired from the public datasets (IRIS dataset). A total of 150 instances of cervical cells were employed in this study. The datasets contained features on basic demographic information, medical historical records, sexual life history, total number of pregnancies and age of first pregnancy [8].

2.2. Analysis of the Diagnostic and Detective Framework

The proposed cervical cancer diagnostic framework incorporates the patient's clinical background, the knowledge representation/preprocessing phase, clinical knowledge respository, the classification algorithm and clinical decision support system as presented on Figure 1.

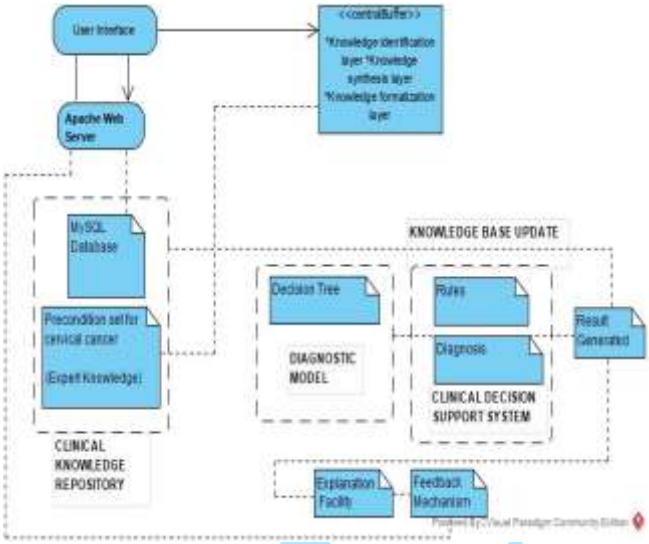

Figure 1: Activity diagram of the proposed model

Knowledge Representation/Pre-processing

The knowledge base is constructed through a sophisticated modelling process. A reasoning component for inferring ground answer to the given problem searches for relevant knowledge from the knowledge base. The knowledge base is present, the reasoning component is algorithmic and the answer derived is justifiable. Hence, the knowledge representation/pre-processing phase in the proposed framework comprises of the following layers for a robust clinical knowledge-based.

Knowledge Identification Layer

This layer identified valid sources and specific patient management knowledge as it pertains to the prevention, early detection and diagnosis of cervical cancer. The knowledge sources considered not only entailed evidence-based recommendations and treatments but also specific tasks and procedures and their scheduling information. The knowledge sources identified in this layer are best practice advice from available evidence and therapeutic intervention, institution specific drug management protocols, medical journals, clinical practice guidelines (CPGs) on cervical screening & oncology and most importantly expert knowledge.

Knowledge Synthesis Layer

This layer involves the acquisition of the clinically useful task-specific heuristics from the identified knowledge sources through the processes of selection, interpretation and augmentation of the guideline statements, strategic and logic. Where necessary, the heuristics were further decomposed into atomic tasks and temporally organized to develop the CPGs and Clinical Pathways packages for cervical cancer containing clear and relevant

evidence-based diagnostic and therapeutic plans for patient care management for clinical professionals.

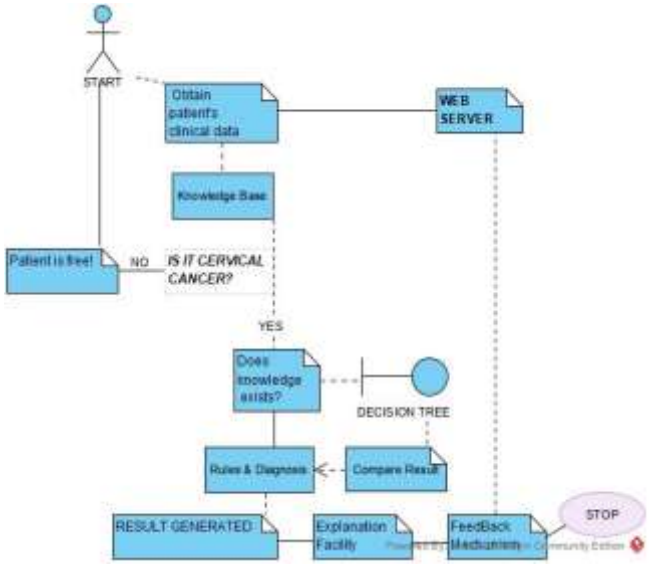

Figure 2: Communication diagram of the web-based model

Knowledge Formalization Layer

In this phase, the knowledge synthesized was modelled and formalized in terms of a dedicated Clinical Pathway ontology. Ontology is the standard knowledge representation for the Semantic Web framework. This layer captures the disease-specific knowledge inherent within the CPG, whilst maintaining the underlying semantics, avoiding ambiguities and identifying the key decisional elements in cervical cancer diagnosis. This ensured the conceptualization of the domain into an unambiguous model, thereby determining any implicit constraints on the relationships between the domain concepts, particularly to assist the alignment of concepts in handling the cancer.

Knowledge Alignment Layer

This layer involves the alignment of discrete and ontologically defined care plans in response to cervical cancer preconditions and risk factors. The alignment of cervical cancer CPs is achieved at knowledge modelling level by developing a unified ontological model that encompasses the combined knowledge of aligned CPs. Also, knowledge alignment was finalized at the ontology level. This is indeed a complex activity given the fact that the alignment needs to take into account the medical correctness and clinical pragmatics of the resultant clinical pathways for cervical cancer.

2.3. Clinical Knowledge-Based Repository

This is the repository for all the relevant domain knowledge and new expert knowledge discovered through research gotten from the knowledge precondition set pertaining to cervical cancer. This knowledge-base is also updated through the workings of the results classification/diagnostic algorithms section where the rules and diagnostic support modules are housed which makes both the knowledge-based and CDSS active.

The optimized outcome from the diagnostic algorithms prior to generating the CPs is also used to update the knowledge-base. When this happens, a similar problem can be taken care of without subjecting it to the SVM and DT algorithms, an instance of learning having taken place. The knowledge base is so built that it is capable of offering diagnosis of cases with no variation (normal) and positives to malignity which are considered as the basis for cervical cancer diagnosis. The section also consists of a sub-module named precondition set for cervical cancer which houses the knowledge set akin to the condition that sets up a need to know if the patient has or does not have the risk of contracting the cancer cell called cervical intraepithelial neoplasia (CIN).

2.4. Diagnostic Model

The classifier deployed for the framework is the Decision Trees**.** Decision trees, such as C4.5, Classification and Regression Trees (CART), Iterative Dichotomiser 3 (ID3), and newer variants (C5.0), are machine learning techniques or classifiers that predict class labels for data items. Decision trees are at their heart a fairly simple type of classifier, and this is one of their advantages. They classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. To sum up, one of the most useful characteristics of decision trees is their comprehensibility, and also it is relatively easy to interpret and implement. One can easily understand why a decision tree classifies an instance as belonging to a specific class. Decision Tree creates a model based on a tree structure. Nodes in the tree represent features, with branches representing possible values connecting features. A leaf representing the class terminates a series of nodes and branches. Determining the class of an instance is a matter of tracing the path of nodes and branches to the terminating leaf. However, many decision tree construction algorithms involve a two - step process. First, a very large decision tree is grown. Then, to reduce large size and overfitting the data, in the second step, the given tree is pruned.

The C5.0 algorithm is a new generation of Machine Learning Algorithms (MLAs) based on decision trees with a number of improvements. Decision

Trees are constructed by analysing a set of training examples for which the class labels are known, and then applied to classify previously unseen examples. If trained on high quality data, decisions trees can make very accurate predictions. C5.0 was developed as an improved version of well-known and widely used C4.5 classifier has several important advantages over C4.5 and Iterative Dichotomiser (ID3). The generated rules are more accurate and the time used to generate them is lower (even around 360 times on some data sets). In C5.0 several new techniques were introduced including:

i. boosting: several decision trees are generated and combined to improve the predictions.
ii. variable misclassification costs: it makes it possible to avoid errors which can result in a harm.
iii. new attributes: dates, times, timestamps, ordered discrete attributes. Values can be marked as missing or not applicable for particular cases.
iv. supports sampling and cross-validation
v. speed significantly faster and more memory efficient

Three other classifiers including Naïve Bayes, K-Nearest Neighbour and Support Vector Machine (SVM) were trained on the IRIS dataset to compare performances with the Decision tree C5.0.

2.5. Clinical Decision Support System

This module of the framework incorporates the rules devoted to interpreting patients' data, diagnosis and detection, tuning platform services such as development of the Clinical Practice Guidelines and Clinical Pathways via the Cervical Cancer ontology; and providing pertinent evaluation and treatment process through the generation of patient specific care plans.

Rules, Diagnosis and Detection

The rules module receives an optimized output upon which rules are generated and subjected to further coordinated analysis that becomes the yardstick for the detection and diagnosis of the cervical intraepithelial neoplasia (or cervical dysplasia). It thereby reduces the false positives and false negatives incurred in the diagnosis. Once this is established, the predicted values become the benchmark upon which generation of patient-specific care plans and treatments take place.

Web development framework

The conceptual framework was designed using Microsoft Visual Studio, an Integrated Development Environment for web application; this is to ensure the following

i. To visualize the Clinical Pathways using a user-friendly interface
ii. To assist healthcare professionals and clinicians in care planning and decision support for patients with cervical cancer
iii. To be a deployable application with minimal effort on any modern web browser, thereby ensuring its portability and accessibility

3. RESULTS & DISCUSSION

The machine learning phase returned C5.0 as the best classifier hence its adoption for this work. For the Naïve Bayes, 140 instances were classified correctly giving 66.90% of accuracy while 10 instances were incorrectly classified giving 33.10%. The build time for Naive Bayes classifer took 0.22 seconds with a Kappa statistic of 0.727 reliable followed by the C5.0 decision tree classifier which resulted in 144 instances correctly classified yielding 94.00% higher performance in terms of accuracy and 6 instances incorrectly classified with 6% performance. The build time was 0.19 seconds and a kappa statistic of 0.554 reliable while the K-NN Classifier results showed 142 instances correctly classified giving 16.27% of accuracy. The build time for the simulation was 0.82 seconds and Kappa statistic of 0.327. The Support Vector Machine (SVM) classifier resulted in 141 instances correctly classified givin 78.86% of accuracy and 9 instances incorrectly classified giving 21.14%.

Table 1: Summary result of each classification algorithm

Algorithm (Total Instances, 150)	*Correctly Classified Instances % (Value)*	*Incorrectly Classified Instances % (Value)*	*Time taken to build Model (seconds)*	*Kappa Statistic*
Naive Bayes	66.90 (140)	33.10 (10)	0.22	0.727
Decsion Tree (C5.O)	94.00 (144)	6.00 (6)	0.19	0.554
K-Nearest Neighbor	83.73 (142)	16.27 (8)	0.82	0.327
Support Vector Machine	78.86 (141)	2.14 (9)	0.51	0.480

Table 2: Classification Accuracy of Nominal Cross-Validation

	Decision Tree (C5.0)	K-Nearest Neighbor	Naïve Bayes (Naïve Bayes)	Support Vector Machine
Cross Validation	94.00	83.73	66.90	78.86

The build time for the simulation was 0.51 seconds with a kappa statistic of 0.480. All the classification models (i.e. Naive Bayes, SVM, KNN and Decision tree) generated can generally achieve good performances to classify cervical precancerous data dependent on the data input. A summary of the foregoing results and analysis presents C5.0 decision tree classification model as the best, having the highest number of instances correctly classified and corresponding percentage as observed from Table 1. This was thus followed by the KNN, SVM and Naive Bayes respectively. Thus, further analysis carried out in this study employed the decision tree classification model due to the better results and potential of the technique to build a real time system as its cross validation accuracy likewise outperforms other base learners as observed on Table 2. Consequently as explained in Figure 2, the web user interface is implemented and its output for user login is as presented on Figure 3. The proposed CCDDS is thus operational for the diagnosis of cervical cancer with its front end operational framework presented on Figure 4.

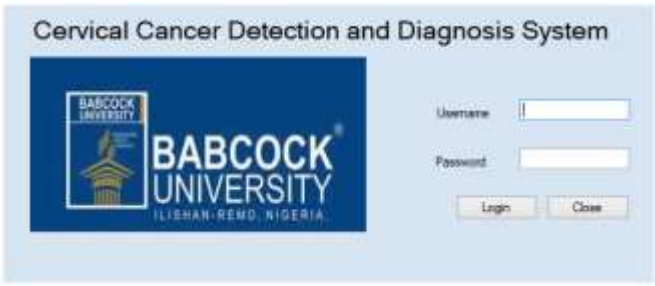

Figure 3: Login page of the web based system

Figure 4: The User Interface for the Diagnostic Framework

At the onset of the operation of the framework, patient's clinical data are elicited from various sources such as screening results which include Pap smear, liquid-base cytology, Electronic Medical Records (EMR) and other clinical databases. The clinical data is then submitted to the knowledge presentation and pre-processing module. Using the rule-base, the system endeavours to establish the presence of HPV using a decision box which is an indication of the prevalence of the cancer. If the outcome of the decision is in the contrary, the patient is declared free or is not prone to cervical cancer and control is transferred back to obtain another patient's data. If on the other hand, the outcome of decision is positive, the knowledge clinical repository queried to find out if the appropriate knowledge already exists. If not yet, the optimized classification algorithm (Decision Tree) takes over and generates results. The results are compared and generated. The outcomes are then displayed via the web user interface of the system as noticed on Figure 2. Decisions taken are now stored via the web server to the clinical knowledge repository for update.

CONCLUSION

A web based cervical cancer detection decision system is proposed in this work with the instrumentality of machine learning as the decision support system. The diagnostic model was built on the Iris dataset by training a C5.0 decision tree learner while consequently building an expert system for the implementation of the framework. C5.0 returned a 94.0% prediction accuracy on cross validation approach outperforming other learners including SVM, Naïve Bayes and K-NN. This CCDDS is implemented on the software engineering spiral model. The research work has addressed an important issue that has become a big problem to most countries of the world and especially the developing countries.

REFERENCES

[1] **H. Sung, R.L. Ferlay, M.L. Siegel, I. Soerjomataram, A. Jermal, F. Bray** - Global Cancer Statistics. GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries CA Cancer J Clin 2021 Feb 4. Epub ahead of print. PMID 33538338

[2] **S. G. Morounke, J.B. Ayorinde, A.O. Benedict, A.F. Faduyile, A.O. Fadaka, I. Oluwadaramilare, B. Adekunle, S. Sunday** - Epidemiology and Incidence of Common Cancers in Nigeria . Journal of Cancer Biology & Research 5(3), 1105, 2017

[3] **N. Munoz, X. Bosch, S. d. Sanjose, R. Herrero, X. Castellsague, K. Shah , C. Meijer** -Epidemiologic Classification of Human Papillomavirus Types Assiociated with Cervical Cancer, The New England Journal of Medicine , 348, 518-527, 2003.

[4] **P. O. Adejumo, L.Y. Ojewale** - Perception of the Relationship between Multiple Sexual Partners and Cervical Cancer by Female Undergraduates of University of Ibadan, Journal of Medicine and biomedical Research, 13(1), 146-153, 2014.

[5] **C. Sabulei, J.E. Maree** - An exploration into the quality of life of women treated for cervical cancer, Curationis 42(1), a1982.

[6] **A. Buskwofie, G. David-West, C.A. Clare** - A Review of Cervical Cancer: Incidence and Disparities. Journal of the National Medical Association, 112(2), 229-232, 2020.

[7] **T. Olaleye, O. Arogundade, C. Adenusi, S. Misra, A. Bello** - Evaluation of Image Filtering Parameters for Plant Biometrics Improvement Using Machine Learning, Patel et al. (Eds): icSoftComp 2020, CCIS 1374, 301-315, 2021.

[8] **Y. M. Al-Wesabi, Y. M. S. Choudhury, D. Won** - Classification of Cervical Cancer Dataset, Proceedings of the 2018 IISE Annual Conference at Orlando