FIREFLY ALGORITHM BASED OPTIMIZED SUPPORT VECTOR MACHINE SYSTEM FOR DETECTING CERVICAL CANCER

MICHAEL FAVOUR EDAFEAJIROKE,

18/27/MCS009

DECEMBER, 2020.

SCHOOL OF POSTGRADUATE STUDIES (SPGS)

FIREFLY ALGORITHM BASED OPTIMIZED SUPPORT VECTOR MACHINE SYSTEM FOR DETECTING CERVICAL CANCER

A M.Sc. THESIS SUBMITTED

BY

MICHAEL FAVOUR EDAFEAJIROKE,

18/27/MCS009

In Partial Fulfilment of the Requirements for the Award of Master of Science In Computer Science

DEPARTMENT OF COMPUTER SCIENCE,
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY,
KWARA STATE UNIVERSITY, MALETE,
KWARA STATE, NIGERIA.

December, 2020.

DECLARATION

I, **EDAFEAJIROKE, Michael Favour** with Matriculation Number **18/27/MCS009** hereby declare that this Master Dissertation, "**AN OPTIMISED SUPPORT VECTOR MACHINE-FIREFLY ALGORITHM BASED SYSTEM FOR DETECTING CERVICAL CANCER**", was submitted to the Department of Computer Science, Kwara State University, Malete, Nigeria, consists entirely of my work produced from research undertaken under the supervision of Prof. Kazeem Alagbe Gbolagade and that no part of it has been submitted for award of a degree here or elsewhere, except for the permissible references from other sources, which have been duly acknowledged.

_____ _____

EDAFEAJIROKE, MICHAEL FAVOUR Signature/Date

CERTIFICATION PAGE

This Master thesis by Edafeajiroke, Michael Favour has been read and certified as meeting the requirements of the Department of Computer Science and School of Postgraduate Studies, Kwara State University, Malete, Nigeria. For the award of Master of Science (MSc) Degree in Computer Science.

Prof. Kazeem Alagbe Gbolagade Main Supervisor	Signature/ Date
Dr. (Mrs) Ronke Seyi Babatunde Co-Supervisor	Signature/ Date
Prof. Kazeem Alagbe Gbolagade Head of Department	Signature/ Date
Prof. Hamza Abdulraheem Dean School of Postgraduate Studies (SPGS)	Signature/ Date
Dr. (Mrs.) Oluwakemi Christiana Abikoye External Examiner	Signature/ Date

DEDICATION

This project is dedicated to God almighty and to every Cervical Cancer Patients who are made to pass through pains for reason beyond them, and those that lost their lives due to lack of early detection and treatment of this deadly disease.

ACKNOWLEDGEMENTS

My profound gratitude goes to God almighty the giver of life and knowledge that made this project a success. To my lovely parent, Mr and Mrs. Edafeajiroke, Moses my Uncle Mr Diei Isikwei for their labour of love, support and prayers and to my brothers and sisters and to Miss Abiola Timilehin Ife, God bless you all richly and I love you all.

I am also very grateful to my ever supporting supervisor Prof. K. A. Gbolagade, for strict guidance, support, and motivation in the process of conducting this study. God bless you sir. My special appreciation goes to the Head of the department Prof. K. A. Gbolagade, Dr. R. M. Isiaka, Dr. (Mrs.) J. F. Ajao, Mr. Sulaiman O. Abdulsalam, Dr. (Mrs.) R. S. Babatunde, Mrs. S. R. Yusuff, Mrs. B. F. Balogun, Dr. A. N. Akinbowale, Mr D. Popoola, and Mr P. D. Oyinloye and all the non-Academic staff in the Department of Computer Science, Kwara State University, Malete, Nigeria. I appreciate them all for their advice, encourage and correction given to me in the course of this project and during my program in the department.

This section is not complete if I don't but acknowledge the Engr. Dr. Ajao's and Family, and Mr S. O. Abdulsalam and Family, for the support, advice, encouragement, prayers and so on. God bless you all, I appreciates support from the MSC and PHD set for 2015, 2016, 2017, and 2018/2019 Computer Science God bless you all. I will not forget to appreciate Dr B. A. Adeyemi and Dr. I. K. Kolawole who assisted me with the subject domain and also Mr. O. Muyiwa and his team for his guidance with the coding aspect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Abstract

Cervical cancer is one of the most cancerous disease caused by the tumor in the cervix. No symptoms are observed at early stage thus impairing healthy living of women above age of 30 round the world, Thereby, escalating their morbidity and mortality rate. Although Pap smear is most preferred cervical cancer control technique, however the difficulties involved in screening and analyzing numerous Pap smear images have led to human errors and loss of results. In recent time's attention have been drawn to Support Vector Machine (SVM) for cervical cancer prognosis. Due to the several computational complexities involved in SVM mapping of high dimensional space, hence the need for Firefly Algorithm (FA) Based Optimized SVM system for Detecting cervical cancer. In this study, FA was used to optimize the cervical cancer dataset obtain from kaggle.com, feature subset Obtained were partitioned to training and testing set of 75% and 25% respectively. Optimal feature subset were capitalized on for optimizing Cost (C) and Gamma (γ) parameters of Radial Bias Function (RBF) and then for developing the system. The Developed System was evaluated using Machine Learning statistical parameters and was implemented on the MATLAB R2016a (9.0.0.341360). Result showed that FA selected 15 cervical cancer risk factors at an accuracy of 95.7%, at a value of C=2 and γ =0.9, which yielded an optimization Accuracy of 96.48%. The developed System had a classification accuracy of 97.20%, Sensitivity of 97.10 and Specificity of 100%. The study was compared with related work for cervical cancer detection. However, it was discovered that the developed system outperformed existing system for cervical cancer detection in terms of classification accuracy, Sensitivity and Specificity. Hence the developed System is strongly recommended for detecting cervical cancer patient since high mortality and morbidity rate can be reduced with or without the assistance of a medical practitioners.

Keywords: FA (Firefly Algorithm), cervical cancer, RBF (Radial Basis Function), RBF parameter optimization, cervical cancer detection, SVM-FA algorithm, cervical cancer Risk Factors.

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Healthy living has been a major concern to human race this is due to the alarming rate of several cancerous, non-cancerous diseases and viruses faced by humans. Some of these diseases are breast, cervical, liver, colorectal cancer and for non-cancerous diseases are hepatitis, kidney stone, liver diseases whooping cough and yellow fever (Abisoye & Jimoh, 2015; Fatima & Pasha, 2017). In all these diseases, Cervical Cancer (CC) has been identified as the most common after breast cancer and the fourth most deadly disease that causes high mortality and morbidity rate of women above 30 years of age in the world (William, Ware, Basaza-Ejiri, & Obungoloch, 2019). Abnormal growth of cell in the cervix of women is traced to be the major cause of CC (Alam, Khan, Iqbal, Wahab, & Mubbashar, 2019).

Cytology (Pap smear), Liquid-Based Cytology, Human Papilloma Virus (HPV) screening, and vaccination against HPV are control technique to CC (Benazir & Nagarajan, 2018). Although simplicity and economical convenience makes Pap smear (PS) technique most pronounced (Kurniawati, Permanasari & fauziati, 2016). However increasing population, high illiteracy rate, unavailability of skilled manpower, lack of equipment, limited acceptability have led to difficulty in screening and analyzing of countless PS images these in turns have resulted to human errors and loss of results (Nithya1 & Ilango1, 2019; Nordqvist, 2017). Hence several health researchers and

gynecologists have affirmed that regular check-up is the only way to CC detection since it is not always obvious at early stage (Babatunde & Muhammad-Thani, 2018). At critical stage, symptoms experienced are fatigue, leg pain, bone fractures, virginal bleeding, weight loss and back pains, as well as spreading and deteriorating other organs like lungs, abdomen, kidney etc (Wu, & Zhou, 2017).

In this regards Mashhour, Houby, Wassif Tawfik and Salah, (2018) opined that low cost technology enhanced with robust machine learning algorithms is a favorable option to PS. Devi, Ravi, Vaishnavi and Punitha, (2016) and Nithya1 and Ilango1, (2019) observed that health researchers and gynecologists still hope for better approach to CC detection. Several data mining tools have been explored to CC detection in the likes of Neural Network (NN), Bayesian Belief Network (BBN), Support Vector Machine (SVM), Adaptic Neuro Fuzzy Inference System (ANFIS), Decision Tree (DT), Random Forest (RF) but SVM proves to be suitable to CC prognosis (Wu & Zhou, 2017). This is due to the tremendous contribution in the field of classification, brought about by simplicity in training, absence of local optimal and its ability to handle complex non-linear data point (Wu, & Zhou, 2017; Thendral & Lakshmi, 2019).

However, over-fitting, pair-wise classification and regularization of parameters and determining the SVM kernels to employ for a given data mining task still remains open issues to SVM researchers (Styawati, & Mustofa, 2019). As such, several algorithms and meta-heuristic algorithms have been proposed as solution to the limitation of SVM

leaving computational complexities issues unresolved (Kisi *et al*., 2015; Mashhour, *et al*., 2018).

Consequently, the uniqueness in the flashing behavior of a more recent robust swarm intelligent algorithm called Firefly Algorithm (FA) was explored for enhanced performance of the SVM (Sharma, Zaidi, Singh, Jain & Anita, 2013; Styawati & Mustofa , 2019). FA have been applied to solve optimization, classification as well as engineering problems in practice (Ashok & Aruna, 2016). Multi-modal problems have be handled with fast convergence rate and also a global problem solver. Yang and He (2018), affirmed that FA is a special case of other meta-heuristic algorithm and performs better when subjected to use. Therefore The need to ensure accurate and reliable diagnosis, low access screening to all, quick and appropriate medical intervention among women in Nigeria, creating awareness, enhancing robust tools for medical analysis and diagnosis as such eliminating human error and reducing mortality and morbidity rate of CC patients (Mashhour *et al*., 2018), aroused the interest of using FA to determining best features to be capitalize on for obtaining optimal parameters of Radial Bias Function (RBF) kernels of the SVM, suitable enough to give high classification accuracy for CC detection process as compared to existing studies.

1.2 Statement of the Problem

In spite of the fact that Cervical Cancer (CC) can be prevented, healthy living have been crippled, which sadly has escalated to high morbidity and mortality of women above the

age of 30 worldwide (National Cancer Institute (NCI), 2015; Nithya1 & Ilango1, 2019). Simplicity, economic convenience makes Pap Smear (PS) control technique most preferred to other CC control methods (Nordqvist, 2017). However PS have become ineffective due to Increasing population, high illiteracy rate, unavailability of skilled manpower, lack of equipment, limited acceptability all these have led to difficulty in screening and analyzing of countless PS images which in turn has resulted to human errors and loss of results (Babatunde & Muhammad-Thani, 2018).

As a result of PS inefficient process, several Machine learning Techniques (MLT) such as NN, DT, RF, SVM, and ANFIS, have been proposed as solution, living open issues that boils down to the computational complexity that should not be handled with levity. In all these MLTs, Wu, & Zhou, (2017) and Thendral and Lakshmi, (2019) opined that SVM simplicity in training, no local optimal and its ability to handle complex non-linear data point, makes attention to be drawn to SVM in the field of classification. Nevertheless, over-fitting, pair-wise classification, parameters regularization and ascertaining the best kernel for a given task is still a major concern to researchers (Oluyinka & Ayobami 2016; Styawati & Mustofa, 2019).

Nithya1 and Ilango1, (2019); Mashhour, et al., (2018); Devi, et al., (2016) and Kisi et al., (2015), re-affirm that, despite the work that have been carried out in this area, researchers are still in need for an enhanced performance of the SVM. Hence the need for a FA based optimized SVM system for detecting cervical cancer that will lead to a non-complicated

interface for both medical and non-medical practitioners, as such reducing the mortality and morbidity rate of CC patients.

1.3 Aim and Objectives

The aim of the study is to develop Firefly Algorithm (FA) based optimized Support Vector Machine (SVM) system for detecting cervical cancer

The specific objectives are to:

i. Perform feature selection using Firefly Algorithm
ii. Optimize Radial Bias Function (RBF) kernel parameters using the selected feature subset in item (i) above
iii. Develop SVM classifier using the feature selected and optimized RBF SVM kernels obtained from objective (i) and (ii).
iv. Evaluate the performance of the developed system using Accuracy, computational time, sensitivity and specificity parameters.

1.4 Scope of the Study

This study majorly focus on the detecting CC amongst women using FA-SVM algorithm based system. FA was used to obtain the risk factors of CC, vital features derived were then used for optimizing RBF kernel parameters of SVM. Similarly RBF and SVM was utilized for classifying the cells into cancerous and non-cancerous cells. This way, non-

complicated user-interface was developed for accurate, effective and efficient prediction of patient with CC.

1.5 Significance of the Study

This work provides an alternative solution to the conventional PS technique for CC detection with a non-complicated graphic user interface that outperform existing state of art. Only few work have be done regarding performing feature selection and Parameter optimization simultaneously on SVM kernels. Hence better means to maximize the capability of SVM towards detecting patient with CC was provided. CC risk factor were narrowed down (with aid of the Biopsy test) that are responsible for CC. CC risk factors will be determined using FA on the Biopsy target variable. Hence this in turn will provide a base for other researcher, as researchers have been silent in this aspect. The study was able to determine the appropriate parameter value of RBF kernel at which it yielded higher prediction accuracy that outperform the existing work for CC detection and the study will be an eye opener to other researchers interested in this area and the entire computing society.

1.6 Justification of the Study

The SVM-FA model will be a welcome development for detecting cervical cancer, this system can be relied on 24 hours daily. The system provides Easy access to medical diagnosis services with a non-complicated graphic user interface which can be operated

by almost everybody, low access screening to all, quick and appropriate medical intervention, as such eliminating human error and reducing mortality and morbidity rate of CC patients. For the first time, the capability of the RBF kernels is maximized with FA towards detecting CC. Also major conflict involved in CC as to prominent risk factors on biopsy as well as right kernel trick for CC detection were also addressed with FA, existing work on CC were furthered compared with the developed system, to measure improvement made.

1.7 Project Layout

The remaining chapter therefore gives the various investigation and studies the existing technique to cervical cancer detection process, the various Machine Learning algorithm (MLA) that was employed for realization of the study aim and objectives.

The procedures and methods chosen to develop the project and to ensure that the aims and objectives of the research are met.

The result obtained after evaluating the system using some case photos. Also, the result analysis were discussed.

Summaries of the overall project result in a better simplified form were provided, Contribution to knowledge, Recommendations and Future Improvements were also highlighted as well.

CHAPTER TWO

LITERATIEW REVIEW

2.1 Data Mining in Health Care

Healthcare, also called Medicine, is intrinsic to human existence. It is as old as the human race; humans have always been in need of solutions to various health related issues, such as cure for deadly diseases like cancer (Wiki2, 2019; Encarta, 2008). According to Mashhour *et al*., (2018) The advent of Information and Communication Technology (ICT) through Machine Learning Technique (MLT) in diagnosis and treatment of these disease is gaining ground in all areas of life, and developing nations are taking advantage of this opportunity in various sectors including the development of health care systems (Idowu, Ogunbodede & Idowu, 2016), hence this study.

Data Mining (DM) is one aspect of ICT in which its impact cannot be overemphasized (Ogundele, Popoola, Oyesola & Orija, 2018). Health data requires analytical methodology in identifying vital information that are used for decision making. DM is important for the healthcare sector in identification and detection of diseases, it helps health researchers to make effective healthcare policies, develop recommendation systems and health profiles for patients (Lindner *et al*., 2017), also, Simplifying the difficulties in evaluating large data generated in the healthcare sector, which are used in discovering knowledge and patterns search for decision making. This is because Healthcare data needs to be analyzed accurately in diagnosis, management and treatment

of diseases. DM applications in health have tremendous usefulness and potentials in healthcare industry (Ogundele *et al*., 2018).

2.2 Evolution to Cervical Cancer Detection

In 400 BCE Hippocrates noted that CC was incurable, 1925 Hinselmann invented the colposcope, 1928. Papanicolaou developed the Papanicolaou technique, 1941, Papanicolaou and Traut: Pap test screening began. In 1946, spatula from Aylesbury was established for the cervix scrape and collection of the Pap sample, 1951. First successful in-vitro cell line, HeLa, derived from biopsy of CC of Henrietta Lacks. 1976. Harald zur Hausen and Gisam found HPV DNA in CC and genital warts; Hausen later won the Nobel Prize for his work, 1988, Bethesda System for reporting Pap results The first FDA HPV vaccine was developed in 2006, (2015). HPV Vaccine shown to protect against infection at multiple body sites, 2018, Evidence for single-dose protection with HPV vaccine FDA, (2015).

Epidemiologists working in the early 20th century noted that CC behaved like a sexually transmitted disease. In summary: CC was noted to be common in female sex workers, it was rare in nuns, except for those who had been sexually active before entering the convent. It was more common in the second wives of men whose first wives had died from CC, It was rare in Jewish women.In 1935, The association between RPV (rabbit papillomavirus) and skin cancer has been found by Syverton and Berry in rabbits.(HPV is species-specific and therefore cannot be transmitted to rabbits) (Safaeian, Solomon, &

Castle, 2017). Presently, CC have been reported to be one of the leading gynecological malignancy worldwide causing high mortality and morbidity rate of women round the world (Barker, Berry & Rainwater, 2018).

2.3 General Overview to Cervical Cancer

The cervix causes CC to erupt (NCI, 2015). It is due to the irregular growth of cells that are capable of invading or spreading to other body parts (Wu, & Zhou, 2017). No signs are found at an early stage. During sexual intercourse, late signs include abnormal menstrual bleeding, pelvic discomfort or pain. Although bleeding after sex may not be serious, the presence of CC may also be suggested (Tarney & Han, 2014).

More than 90 % of cases are caused by HPV; most people who have had HPV infections do not develop CC, however. smoking, a poor immune system, birth control pills, beginning sex at an early age, and having multiple sexual partners are other risk factors (Kumar, Abbas, Fausto & Mitchell, 2007). As a consequence of precancerous shifts, cervical cancer usually develops over 10 to 20 years. Squamous cell carcinomas are about 90 percent of CC cases, 10 percent are adenocarcinomas, and others are minor (Tran, Hung, Roden, & Wu 2014). The diagnosis typically consists of cervical and biopsy screening. In order to assess whether the cancer has spread or not, medical imaging is then performed (World Health Organisation, 2014). A woman named Henrietta Lacks (Kumar, Abbas, Fausto & Mitchell, 2007) created the most popular immortalized cell

line, known as HeLa, from CC cells. This research was needed by the urge for a better approach to the CC detection process.

Vaccines against HPV defend this family of viruses from two to seven high risk strains. High risk of cancer still exists, guidelines recommend continuing regular Pap tests. Additional preventive strategies include the use of little to no multiple sexual partners and condoms. CC screening using the Pap test or acetic acid can identify precancerous changes, which when treated, can prevent the development of cancer. Treatment can be a mixture of surgery, chemotherapy and radiation therapy. However, effective treatment are strongly dependent on how early the CC is detected (Charles & Carraher, 2014). Hence SVM-FA for CC detection.

2.3.1 Sign and Symptoms

The early stages of CC may be completely free of symptoms. Vaginal bleeding, contact blood (one of the most common type of bleeding following sex) or vaginal mass (rarely) might imply malignancy.. Also, moderate pain during sexual intercourse and vaginal discharge are symptoms of CC (Gadducci, Barsotti, Cosio, Domenici & Genazzani, 2011). Metastases in the abdomen, lungs, or elsewhere can occur in advanced diseases. Advanced CC symptoms may include: lack of appetite, loss of weight, tiredness, pelvic pain , back pain, leg pain, swollen legs, extreme vaginal bleeding, bone fracture, and (rarely) leakage of urine or feces from the vagina (Snijders, Steenbergen, Heideman & Meijer, 2016), and to compound this, CC can spread to other organs such as abdomen, lungs, kidney (fayz, rizka, Abo & Maghraby, 2017).

2.3.2 Causes of Cervical Cancer

According to fayz, rizka, Abo and Maghraby, (2017) infection with some types of HPV is among the major risk factor for cervical cancer, then smoking. HIV infection is also a risk factor. Not all of the causes of cervical cancer are known, however several other contributing factors are HPV, Cigarette smoking, oral contraceptives, multiple pregnancies, weakened immune system, Birth control pills and Other sexually transmitted diseases (STD) (Jensen, Schmiedel, Frederiksen, Norrild, Iftner & Kjær, 2012; Remschmidt, Kaufmann, Hagemann, Vartazarova, Wichmann and Deleré, 2013; NIHNCI, 2019;).This actually gave rise to using FA to determining the risk factor for CC for improved CC detection and control process.

2.3.3 Diagnosis of Cervical Cancer

Although Screening does not detect cancer but looks for abnormal changes in the cells of the cervix (Lin, Zhou, Dai, Cheng, & Wang, 2017). Without treatment, some abnormal cells can eventually develop into cancer hence gynecologist and health researchers strongly recommend regular screening of CC. some diagnosis treatment are Precancerous lesions, CC smear test HPV DNA Testing, addition test doctor recommend are colposcope**,** Examination under anesthesia (EUA), Biopsy, Cone biopsy LLETZ, CT scan, MRI, and Pelvic ultrasound (Falcetta, Medeiros, Edelweiss, Pohlmann, Stein & Rosa, 2016), this method are normally expensive and are also prone to psychological state and human fatigue hence the FA-RBF-SVM model for early CC detection process.

2.3.4 Cervical Cancer Staging

CC is staged by the International Federation of Gynecology and Obstetrics (FIGO) system. The stage at which a person receives a CC diagnosis can help indicate their chances of survival for at least 5 more years: this stages are Stage 1: In early stage 1, the chance of surviving at least 5 years is 93 percent, and in late stage 1, it is 80 percent. stage 2: In stage 2 it is 63%, but by the end of stage 2 it decreases to 58%. stage 3: Opportunities are lowered from 35% to 32% during this step. stage 4: Those with cervical stage 4 are 15 to 16% likely to survive 5 years later. Stage 5: The survival rates are average and do not extend to everyone. In certain cases, treatment before stage 4 is successful (Curry, Krist, Owens, Barry, Caughey & Davidson, 2018). All these aroused the need for FA-RBF-SVM based system for CC detection.

2.3.5 Cervical Cancer Treatment

CC treatment options include surgery, radiotherapy, chemotherapy, cervical cancer clinical trials or combinations of these (Lin et al., 2017; Zhang, Dai, Zhang & Wang, 2015; Curry et al., 2018; NCI, 2019.). Deciding on the form of treatment depends on many factors, such as the stage of the cancer, as well as age and general health. Treatment for early stage CC has a high success rate because the cancer stays within the cervix. The further a cancer spreads from its original area, the lower the success rate tends to be hence the need for a system to provide timely detection.

2.3.6 Cervical Cancer Prevention

A number of measures can help reduce the chances of developing cervical cancer. This section explain some of the measure for cervical detection process. FA-SVM was used to determine prominent risk factors, output of the FA can be used for creating awareness for both infected and non-infected patients. Hence enhancing the existing prevention techniques. Some of these prevention technique manually employed by gynecologist ar**e** HPV vaccine, Safe Sex with use of condom**,** Cervical screening, having fewer sexual partners, delaying first sexual intercourse and Stopping smoking (Luhn, Walker, Schiffman, Zuna, Dunn, Gold, Smith, Mathews, Allen, Zhang, Wang & Wentzensen, 2016)

2.4 Review of Relevant Data Mining Tools

The most pressing task of bioinformatics has become to analyze and interpret increasing data size and types (Benazir & Nagarajan, 2018). This according to Ogundele *et al*., (2018) calls for an effective and efficient approach to extract vital knowledge that are hidden in the vast data. Thus the process of applying computer based information system (CBIS), including new techniques, for discovering knowledge from this vast data is called data mining (Kurniawati, Permanasari, & fauziati, 2016). By this appropriate MLA must be taken into cognizance for effective and efficient CC detection (Jensen *et al.,* 2012) this inform this project works.

Generally, the data mining models are predictive model and descriptive model, the predictive model often apply supervised learning functions to predict unknown or future values of other variables of interest**.** The descriptive model on the other hand, often apply the unsupervised learning functions in finding patterns (Zolbanin. Delen, & Zadeh, 2015). The model to be employed is determined by the data or task been performed. For instance, clustering, association rules, correlation analysis, are often used for descriptive models. While classification, regression and categorization are used for predictive models (Mamiya, Schwartzman, Verma, Jauvin. Behr, & Buckeridge, 2015). In this study predictive model was developed by classifying the dataset for accurate and timely CC detection.

This section therefore gave a detailed description of some machine learning tools, technique that was used for realization of this study, exploring their strength and weakness to this work

2.4.1 Data Mining Technique (DMT)

Several data mining technique are available, the task to be performed give rise to which of the method to employ (Ogundele et al., 2018). The technique employed for anomaly detection are, standard support vector data description, density induced support vector data description, Gaussian mixture. Then the method widely used for clustering is vector quantization. The methods widely used for classification are statistical, discriminant analysis (DA), Decision Tree (DT), Markov based, Swarm Intelligence (SI), K Nearest

Neighbor (KNN), genetic classifiers, Artificial Neural Network (ANN), SVM and association rule (Abdi & Giveki, 2013).

2.4.1.1 Data Mining Technique in Healthcare

Healthcare covers a detailed processes of the diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans (Kang, Kang, Ko, Cho, Rhee & Yu, 2015). In healthcare DMT are used mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision. The section therefore discuss some data mining techniques used in healthcare (Resig, 2018).

i. Anomaly Detection

Anomaly detection (AD) is used in discovering the most significant changes in the data set. Feng, Fu, Du, Li and Sun, (2016) used three different anomaly detection method, standard support vector data description, density induced support vector data description and Gaussian mixture to evaluate the accuracy of the anomaly detection on uncertain dataset of liver disorder dataset which is obtained from University of California, Irvine (UCI). Nevertheless AD is not part of the technique employed to actualization of this research work.

ii **Clustering**

Clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data. Kang, *et al*., (2015) had used the vector quantization method in clustering approach in predicting the readmissions in intensive

medicine. The algorithms used in the vector quantization method are k-means, k-mediods and x-means. The datasets used in were collected from patient's clinical process and laboratory results. However clustering is not part of the technique employed for actualizing this work.

iii. Classification

Classification is the discovery of a predictive learning function that classifies a data item into one of several predefined classes. There several classification technique these are Statistical, (DA), DT, KNN, Logistic Regression (LR), Bayesian Classifier, SVM and SI algorithm (Jensen *et al*., 2012). However In this study SVM was used as a classifier while a recent SI algorithm called FA, which is more robust than every other SI algorithm (Yang and He, 2018). FA was used as a wrapper approach for selecting the relevant parameter suitable for detecting CC. SVM and FA are discussed in session 2.4.4 and 2.4.4.5

2.4.2 Comparison of Some Classification Technique

According to Divya and Agarwal (2013) classification is one the important technique in data mining several classifier are used but one specific classifier cannot be chosen to be best due to many reason one of which is the factors which affect the usability of such classifier, hence in this session strength and drawback of several classifier were explored, also the capability in the classifier used was appropriately maximized to actualization of this study aim, the significance to this study were also elucidated.

K-Nearest Neighbor (KNN) classifier is one of the simplest classifier that discovers the unidentified data point using the previously known data points (nearest neighbour) and classified data points according to the voting system (Silver, Sakara, Su, Herman, Dolins & O'shea, 2001), It is easy to implement. Training is done in faster manner however it requires large storage Space, Sensitive to noise and slow testing time Hence KNN was not used in realization of the study objectives.

Decision Tree (DT) is similar to the flowchart in which every non-leaf nodes denotes a test on a particular attribute and every branch denotes an outcome of that test and every leaf node have a class label (Goharian and Grossman, 2003). There are no requirements of domain knowledge in the construction of DT.it minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions, it can easily process the data with high dimension, easy to interpret, can handle both numerical and categorical data. Also, it is restricted to one output Attribute, generates categorical Output, an unstable classifier i.e. Performance of classifier depend upon the type of Dataset, complex DT are generated with numeric dataset (Divya & Agarwal, 2013). However DT does not form a basis to realization of this study aim and objectives.

Neural Network (NN) is a classification algorithm that uses the method of gradient descent and is based on the biological nervous system with many interrelated processing elements known as neurons. Functioning in unity to solve specific problem. Rules are extracted from the trained NN help to improve interoperability of the learned network

(Silver, Sakara, Su, Herman, Dolins & O'shea, 2001), NN help to easily identify complex relationships between dependent and independent variables and Able to handle noisy data. But faced with issues like Local minima, Over-fitting, The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks (Divya & Agarwal, 2013). NN however does not form a basis for this study.

Bayesian Belief Network (BBN) is a classifier based on bayes theory is known as Bayesian classification. It is a simple classifier which is achieved by using classification algorithm (Silver, et al., 2001). BBN is widely used by many researchers in healthcare field. Liu et al. develop a decision support system using BBN for analyzing risks that are associated with health (Silver, et al., 2001). Curiac Vasile, Banias, Volosencu and Albu, (2009) analyze the psychiatric patient data using BBN in making significant decision regarding patient health suffering from psychiatric disease and performed experiment on real data obtain from Lugoj Municipal Hospital. Result from several Researcher showed that BBN computations process is easy and also have better speed and accuracy for huge datasets but it does not give accurate results especially cases where there exists dependency among variables. BBN does not form a basis for this study.

2.4.3 Support Vector Machine (SVM)

SVM is a learning machine which was developed by Vladimir Vapnik in aims to construct decision functions in the input space based on the theory of Structural Risk Minimization (Vapnik, 1998). It is widely used binary classifier, steps in form of

Pseudocode used by SVM to solve a giving data mining task is shown in Algorithm 2.1. It separates instances from different classes with hyper-plane that is as far as possible from them and all instances are outside the margin. This can be formally defined with the following expression in equation (2.1).

$$yi(w.xi + b) \geq 1 \; for \; 1 \; for \; 1 \leq i \leq n, w \in Rd; b \in R \tag{2.1}$$

where xi refers to instances, yi ϵ {1, -1}; are labels of instances, an intercept term is b, w is normal vector to the hyperplane, d is the dimension of input vector and n is the number of input data. Instances that lie nearest to the hyperplane define the hyperplane and they are called support vectors. Hyperplane defined by Eq. 1., where all instances from one class must be on the same side of the hyperplane is named hard margin. Problem with this definition is that data from real world usually have a few outliers, instances that are significantly different from other instances from the same class. In that case hyperplane will not be found so classification is impossible (Tuba, Mrkela & Tuba, 2016). In order to solve this problem so-called soft margin was introduced. Soft margin is defined by the following Linear SVM kernel expression in equation (2.2).

$$yi(w.xi + b) \geq 1 - \epsilon i, \epsilon i \geq 0; 1 \leq i \leq n \tag{2.2}$$

Where ϵ_i represent slack variables which allow instances to fall off the margin. Finding optimal soft margin is equivalent to solving quadric programming problem shown in equation (2.3).

$$\text{Min} \;\; 1/2||w||^2 + C \sum_{i=1}^{n} \epsilon i \; ; \qquad \epsilon i \geq 0 \tag{2.3}$$

Where C stands for parameter of soft margin cost function. It can be noticed that for larger value of C model will be similar to hard margin. Parameter C has strong influence on classification Accuracy. Described method works only with linearly separable data, as mentioned in the Introduction, kernel trick was introduced with aim to overcome this problem the idea is to substitute the kernel function for the dot product. Any function that satisfies the condition of Mercer can be chosen for the kernel function.. Linear, polynomial, Gauss (Radial basis function or RBF) and sigmoid functions are the most common used kernel functions to achieve the study aim RBF kernel was employed to this study.

Algorithm 2.1: Pseudo code for Support Vector Machine (SVM) (Wiki 5, 2020)

Step 1: Import the dataset

Step 2: Explore the data to figure out what they look like

Step 3: Pre-process the data

Step 4: Split the data into attributes and labels

Step 5: Divide the data into training and testing sets

Step 6: Train the SVM algorithm

Step 7: Make some predictions

Step 8: Evaluate the results of the algorithm

2.4.3.1 Support Vector Machine Kernel functions

The kernel functions are one of the major kernal tricks of SVM. Those functions are used when the samples are linearly non-separable. The kernel tricks thus expand the class of decision functions to the non-linear case by mapping the samples into a high-dimensional feature R from input space X without ever having to directly compute the mapping, in the hope of obtaining substantial linear structure in R for the samples. In addition, the kernel function can be interpreted as a measure of the similarity between samples xi and xj, allowing even very complex boundaries to be distinguished by SVM classifiers.

There are several possibilities for the choice of this kernel function, including polynomial, sigmoid, Radial Bias Function RBF and linear. In this study RBF were used for CC detection. The SVM kernels will be discussed in this session

i. Linear Kernels (LK)

A LK is often recommended for text classification with SVM because text data has lot of features and is often linearly separable. In text classification, both the number of instances and features are large. In this study the dataset is huge hence result from it will be in-efficient hence the LK was not considered for the study. Equation (2.4) describe the LK

$$k(xi, xj) = (xi) . \quad (xj) \tag{2.4}$$

i. Radial basis function (RBF) Kernels

RBF is usually the first choice for kernel function, IT simplicity, flexibility, ability to obtain high accuracy and efficient that is derived from it when employed for any data

mining task prompted its use for realizing the study aim. RBF is defined by equation (2.5).

$$k(xi,\ xj) = \exp(-\gamma||\text{xi} - \text{xj}||.^2) \quad (2.5)$$

Parameter γ plays a major role in achieving a better classification accuracy. The γ parameter can be described as a parameter that defines how far the influence of a single training example reaches, low values meaning it reaches far and the meaning of high values can be interpreted as close reach. It can also be understood as the inverse of radius of influence of samples that are selected as support vectors. Based on this description of SVM it can be seen that accuracy of the model depends on optimal parameter selected. Grid search is one technique often used for this optimization problem, but better choice for this problem would be stochastic population based algorithm (Tuba, Mrkela & Tuba, 2016). Hence this study.

iii. Polynomial kernel

The Polynomial kernel is a non-stationary kernel. It is well suited for problems where all the training samples are normalized. The vital parameters that must be determined are the slope gamma. The constant term r and the polynomial degree d (hence d=3, r=0) Polynomial kernel is described in equation (2.6) (Rimah, Dorra & Noureddine 2015).

$$k(\text{xi},\ \text{xj}) = \left(1 + \text{sum}(\text{xi} * \text{xj2})^{\text{d}}\right) \quad (2.6)$$

iii. Sigmoid kernel

The kernel must fulfill the theorem of Mercer, and also requires that the kernel be certainly positive. However, the Sigmoid kernel is not a semi-definite positive for some

values of its parameters, despite its wide use. Therefore, otherwise the parameters r must be correctly selected, the results may be dramatically inaccurate, so much so that the SVM performs worse than random, see equation (2.7).

$$k(xi, xj) = \tanh(\sigma\, xi^T xj + r) \tag{2.7}$$

We can see σ as an input sample scaling parameter, and r as a moving parameter that governs the mapping threshold (r=0 therefore). The sigmoid kernel is usually no stronger than the linear kernels and the RBF (Rimah, Dorra & Noureddine 2015).

2.4.3.2 Advantage and Disadvantages of SVM

In order to achieve the aim of the study, SVM was explored. this is due to its simulating behavior to humans, numerous contribution to the field of classification, better accuracy as compared to other classifier, ability to handle complex nonlinear data points using kernel trick and Over fitting problem is not as much as other methods however SVM has certain limitation such as Computationally expensive, The main problem is the selection of right kernel function for every given data mining task, since different kernel function Shows different results, As compare to other methods, training process take more Time, SVM was designed to solve the problem of binary class, It solves the problem of multi class by breaking it into pair of two classes such as one against-one and one-against all, all this setback necessitated this study.

2.4.4 Firefly Algorithm (FA)

Fireflies are winged beetles or insects that produce light and blinking at night. The light has no infrared or ultraviolet frequency, called bioluminescence, and is created

chemically from the lower abdomen. Specifically, they use the flash light to draw mates or prey. The flash light was often used to alert the fireflies about the possible predators as a defensive warning system. Firefly algorithm formulated by Yang (2008) is a met-heuristic algorithm that is inspired by the flashing behavior of fireflies and the phenomenon of bioluminescent communication. Yang (2008) formulated the Firefly Algorithm with the following assumptions:

i. A firefly will be attracted to each other regardless of their sex because they are unisexual.

ii. Attractiveness is proportional to their brightness whereas the less bright firefly will be attracted to the brighter firefly. The attractiveness, however, decreased as the distance between the two fireflies increased.

iii. If the brightness of both fireflies is the same, the fireflies will move randomly. The generations of new solutions are by random walk and attraction of the fireflies (Yang, 2010). The luminosity of the fireflies should be related to the objective position of the related issue.

Their desirability makes them capable of subdividing themselves into smaller groups and swarming around local models for each subgroup. FA is therefore sufficient, as stated by (Yang, 2011), for constrained optimization problems.

2.4.4.1 Pros and Cons of Firefly Algorithm

From the previous literatures, many researchers have stated that FA developed by Yang in 2008 is a very powerful technique to solve constrained optimization problems and NP-

hard problems Johari, Zain, Mustaffa and Udin, (2016) stated that FA has widely been applied to solve continuous mathematical functions but has been rarely reported. For applied mathematics, the algorithm must be just a simple math and logic (Azar, 2009). The behavior of FA is simple and therefore it is suitable to solve the continuous mathematical functions. Apostolopoulos, and Vlachos, (2011), said that FA is very efficient and can outperform other conventional algorithms based on statistical performances measured using standard stochastic test functions. The algorithm operates on the basis of global fireflies contact. It can also be optimal globally and locally at the same time. Yang (2010) said that FA uses random numbers that are primarily actual.

Different fireflies function independently and for parallel implementation it is suitable. Wang, (2013) said that FA is one of the technique that recently used by researchers to solve constrained optimization problems in dynamic environment.

The major drawbacks of FA is that for every iteration, FA is compared with every other firefly in the system, hence increasing the number of computations. As such the number of fireflies in the search space increases, the level of computations also increases to a large extent (Prakasam and Savarimuthu, 2016; Ahmed & Maheswari, 2017). Therefore this study was able to manage this problem by carrying out extensive research on related literature to determine the values that will be assigned to FA parameters, as to achieve optimized features relevant enough for CC detection. This way number of computations were reduced, search space was also reduced.

2.4.4.2 The Attractiveness of the Firefly

The brightness can be term as the attractiveness, I of firefly i on the firefly j is based on the degree of the brightness of the firefly i and the distance rij between the firefly i and the firefly j (Yang, 2010) as in equation (2.8).

$$I(r) = \frac{Is}{r^2} \tag{2.8}$$

Suppose n fireflies exist; and xi corresponds to the firefly I solution. The luminosity of the goal function f (xi) is associated with Firefly i. A firefly's brightness I is selected to disclose its recent location of its fitness value or objective function f(x) as in equation (2.9).

$$I(i) = f(xi) \tag{2.9}$$

The firefly that is less bright (attractive) is drawn and moved to the brighter one; and each firefly has a certain value of attractiveness β. However, depending on the distance between fireflies, the attractiveness value of β is subjective. The firefly's attractiveness function is defined by equation (2.10).

$$\beta(r) = \beta 0 e^{-y r^2} \tag{2.10}$$

Where β0 is the value of firefly attractiveness at r = 0 and γ is the coefficient of media light absorption.

2.4.4.3 The Attractive Movement of Firefly

Equation (2.11) described the movement of a firefly i at position xi moving to a brighter firefly j at position xj, which was also explain by Yang (2010). FA is explain in Algorithm 2.2

$$xi\,(t+1) = xi\,(t) + \beta 0 e^{-yr^2}(xi - xj) + \alpha\epsilon i \qquad (2.11)$$

The attraction of the firefly xj and αε! Is due to $\beta_0 e^{-yr2}$ (xi- xj) a randomization parameter; with $\beta 0 = 0$ is seen as a simple random movement. The algorithm juxtapose the attractiveness of the new firefly position with old one. Higher attractiveness from the new position value, the firefly is shifted to the new position; else the firefly will remain in the current position. Arbitrary predefined number of iterations or predefined fitness value determines the termination criterion of the FA. Equation (2.12) shows the brightest firefly movement in randomly based fashion.

$$xi\,(t+1) = xi\,(t) + \alpha\epsilon i \qquad (2.12)$$

Algorithm 2.2: Pseudo code of the firefly algorithm (FA) (Bahadormanesh, Rabat, and Yarali, 2017)

Begin

i. Objective Function:
ii. Generate an initial population of fireflies
iii. Formulate light intensity I so that it is associated with (for example, for maximization problems, or simply ;)
iv. define absorption coefficient y

While (t< MaxGeneration)

for I = 1 : n (all n fireflies)

for j = 1 : I (n fireflies)

if ( );

vary attractiveness with distance r via;

move firefly I towards j;

Evaluate new solution and update light intensity;

end if

end for j

end for i

Rank foreflies and find the current best;

end while

post-processing the results and visualization;

end

2.4.4.4 Special FA Cases

A close study of FA structure shows that by at looking at Equation (2.13) Closely, α is an important scaling parameter, when light absorption parameter γ is set as γ = 0, which means that there is no exponential decay and thus the visibility is very high. In this case, all the fireflies in the whole domain can see each other it as as expressed in equation (2.13).

$$xi\,(t+1) = xi\,(t) + \beta 0 e^{-y\,r^2}(xi - xj) + \alpha \epsilon i \qquad (2.13)$$

If γ= 0, α= 0 and β0 is fixed, then FA becomes a variant of differential evolution (DE) without crossover (Sundari, Rajaram & Balaraman, 2016). In this special case, if xj is replaced by the best solution in the group g*, this reduced FA is equivalent to a special case of accelerated particle swarm optimization (APSO) (Xiao, Shao, Liang, and Wang, 2016). If β0= 0, FA is equivalent to the basic simulated annealing (SA) with α as the cooling schedule (Yang, 2014). In addition, if ϵi is further replaced by ϵxi, this special case is equivalent to the pitch adjustment of the harmony search (HS) algorithm. It is also clear that the special cases of the regular FA are DE, APSO, SA and HS.. In other words, both in linear and nonlinear method, FA can be referred to be a good combination of

APSO, HS, SA and DE enhanced. It is no surprise that FA can outperform these algorithms for many applications.

FA is controlled by three parameters: the randomization parameter α, the attractiveness β, and the absorption coefficient γ. According to the parameter adjustment, FA distinguishes two asymptotic behaviors. The former appears when γ 0 and the latter when $\gamma \infty$. If γ 0, the attractiveness becomes $\beta = \beta 0$. That is, the attractiveness is does not change anywhere within the search space. This phenomenon is a special case of particle swarm optimization (PSO). If γ ->∞, the second term falls out from Equation (2.12) and the firefly movement becomes a random walk, which is essentially a parallel version of simulated annealing. Infact, each implementation of FA can be between these two asymptotic behaviors. In other to achieve this study aim, proper choice of these parameters were taken into consideration.

2.4.5 Dimensionality Reduction

Dimensionality Reduction (DR) is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into Feature Selection (FS) and Feature Extraction (FE) (Pratap, 2019). The various methods used for DR are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Generalized Discriminant Analysis (GDA). Dimensionality are required in data mining process because It helps in data compression, and hence reduced storage space, reduces computation time also helps in removing redundant features, if

any, the accuracy of the model is improved and also reducing over fitting (Pratap, Abhishek, & Dev, 2019).

2.4.5.1 Feature Selection Methods

FS, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. It usually involves three methods which are Filter, Wrapper Embedded (Pratap, 2019). These three method are extensively described in this session.

i. Filter Method

Filter methods are generally used as a preprocessing step is as shown in Figure 2.1. The selection of features is independent of any MLA. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable (Dong, Hua, & Li, 2007). Examples of this methods are Pearson's Correlation, Linear Discriminant Analysis LDA**,** ANOVA stands for Analysis of variance, Chi-Square, standard deviation and t-test etc.

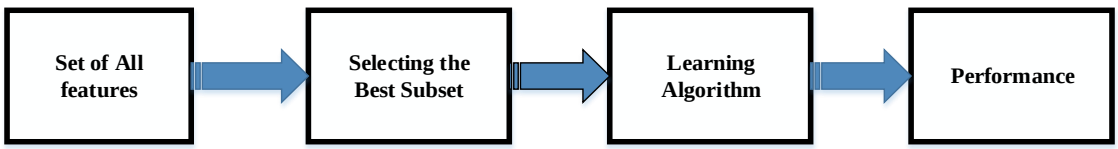


Figure 2.1 Filter Method Wiki1, (2018)

i. Wrapper Methods

Wrapper technique, employs subset of features and train a model using them. Based on the inferences drawn from the previous model, it was decided to add or remove features from your subset as shown in Figure 2.2. The problem is essentially reduced to a search

problem. These methods are usually computationally very expensive but usually provide the best performing feature set for that particular type of model or typical problem (Boutsidis, Zouzias, Mahoney, & Drineas, 2015). Most of the swarm intelligent algorithm uses the wrapper approach. Example are: GA, FA, SA PSO, BA, TS, HS, APSO, EA, MA, DE, ACO etc.

Some common approaches of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

i. **Forward Selection**

Forward selection is an iterative process which begin with no variable in the model. In each iteration, it continue to implement the function that best enhances the model until a new variable is introduced that does not boost the model's performance.

ii. Backward Elimination

This process, start with all features and then begin to remove the least significant feature for backward elimination, iteration which improves the performance of the model at each is repeat until no improvement is observed on removal of features (Wiki1, 2018).

iii. Elimination of Recursive Feature: It is a greedy optimization algorithm that seeks to find the best subset of performing features. It generates models repeatedly and keeps the best or worst performing function at each iteration aside. Until all the features are depleted, it builds the next model with the left features. It then ranks the characteristics based on their elimination order (Wiki1, 2018).

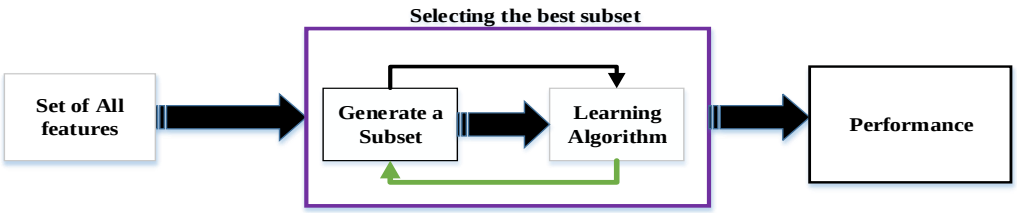


Figure 2.2: Wrapper Method Wiki1, (2018)

iii. Embedded

The characteristics of filter and wrapper techniques are integrated into embedded approaches as shown in Figure 2.3, It is implemented by algorithms which have their own built-in feature selection methods. Some of the more popular examples of these techniques are LASSO and RIDGE regression, which have inbuilt penalization functions to mitigate over-fitting (Wiki1, 2018).

i. Lasso regression performs a regularization of L1, which adds a penalty equal to the absolute value of the coefficient magnitude.

ii. Ridge regression performs regularization of L2, which adds a penalty equal to the square of the coefficients' magnitude.

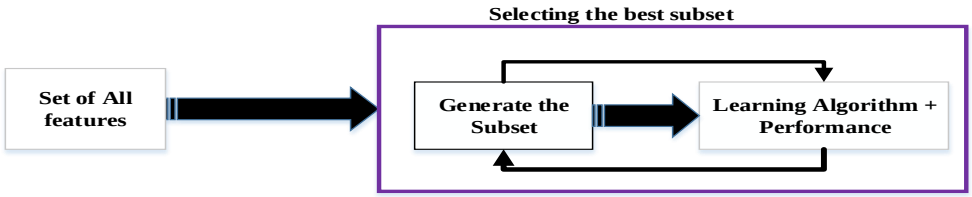


Figure 2.3 Embedded Method Wiki1, (2018)

2.4.5.1a Reasons for Feature selection

The essence of FS to this study is to automatically select those features in the CC dataset that contribute most to CC prediction variable or output in which the study

interest is built on. The benefits of performing feature selected before classification are:

i. Avoid Over fitting: Less redundant data gives performance boost to the model and results in less opportunity to make decisions based on noise
ii. Reduces Training Time: Less data means that algorithms train faster
iii. Most importantly, the risk factors to CC are very necessary as a means to create awareness for non-infected patients.

2.4.5.2 Feature extraction

Extraction of features means reducing the amount of resources needed to define a broad data set. Dependent component analysis, Isomap, Kernel PCA, Latent semantic analysis, Partial least squares, Principal component analysis, Multifactor dimensionality reduction, Nonlinear dimensionality reduction, Multilinear Principal Component Analysis, Multilinear subspace learning, Semi definite embedding and Auto encoder are used in general dimensionality reduction techniques.

2.5 Roles of Swarm Intelligent Algorithm to SVM Optimization

Over-fitting, pair-wise classification and regularization of parameters remain a draw back to the use of SVM. Furthermore classification accuracy of SVM is based on the features and parameters used for training (Oluyinka & Ayobami, 2016), several approaches have been explored to solve this issues but the SI algorithm in recent times have provided better result over other algorithm that have been used (Ramit, Puneet & Ravi, 2018). Hence the study.

According Kisi, Ozgur. Jalal, Sepideh, Shamshirband. Motamedi, Petkovi and Hashim (2015) various optimization algorithms have been used for selection of these parameters such as the grid search algorithm and gradient decent algorithm, but the success rate has been minimal. Computational complexity seems to be the main disadvantage of the grid method, which restricts its applicability to simple cases. The grid search algorithm is also prone to local minima. Multiple local solutions exist for most of the optimization problem meta-heuristic algorithms in recent times have given a competitive result, because they are capable of providing global solution to such problems hence the study.

Nature inspired meta-heuristic optimization algorithms, such as the ant colony optimization (ACO), Genetical Algorithm (GA), Particle Swarm Optimization (PSO) and cuckoo search (CS) and many more have found wide applications in different fields of science for several hybrid algorithms, The basis of these algorithms is the selection of the most appropriate in natural systems. The latest algorithm among the nature-inspired meta-heuristic optimization algorithms is the FA (Styawati & Mustofa, 2019). The FA is believed to be more robust and efficient in finding both global and local optima compared to others. The prediction accuracy of the SVM model highly depends on proper choice of model features and parameters. Even though prearranged approaches for parameter selection are essential, alignment to the model parameter is also important. (Fister, Yang, Fister, Brest & Fester, 2013) all these inspired the choice for FA.

2.5.1 Review of the Various Meta-Heuristic Algorithm

This session gave a brief review to some very prominent meta-heuristic algorithm, the inventors, there strength and weakness, there significance to this study were also explored.

Simulated Annealing (SA) was invented by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953), the notion behind this algorithm is that it involves heating a substance and then gradually cooling it to obtain a high-strength crystalline structure, SA is a neighborhood search, memory-less algorithm with a capability to escape local optima and hence avoid premature convergence. If temperature is low the system can get stuck in local optima at high temperature difficulty in reaching the optimal solution can be experienced, Suited for problem with rough landscape (large number of local optima)

Tabu Search (TS) was invented by Glover and Laguna (1997), it uses the information gathered during the iterations (stored in memory) to make the search Process more efficient, the best feature of TS is its ability to avoid cycles. Appropriate selection of neighborhood operator and search space, it is not efficient enough in moving out of the local minima and requires in-depth knowledge of the problem at hand.

Genetic Algorithm (GA) was invented by Holland (1975), it is inspired by Darwin's theory of evolution. They are characterized by inherent parallelism GA exhibits inherent parallelism and is a good technique for solving complex optimization Problems, GA suffer from premature convergence. The effectiveness of algorithm is based on selection

of objective function, population size, and probability of crossover and mutation. Suitable for large population size which in turns leads to Increase in computational time.

Differential Evolution (DE) was invented by Storn and Price (1995), It resembles evolutionary algorithm (EA) but differs from the traditional ones in the way candidate solutions are generated and the use of greedy selection scheme. DE is simple and efficient but have issues like Increase in number of parents and reduction in scaling factor makes convergence easier but is computationally intensive Crossover constant does have much impact on convergence speeds.

Particle Swarm Optimization (PSO) was invented by Kennedy and Eberhart (1995), it based on the swarm behavior (birds flocking). It maintains a single static population whose members are tweaked in response to the search history, easily Implemented in a wide range of optimization problems. It uses the gbest, pbest and velocity and hence prone to the problem associated with them, also particles fly out of the swarm easily making it difficult to converge before getting to optimal solution

Ant Colony Optimization (ACO) invented by Dorigo et al. (1982) and Dorigo, (1992) based on the food finding (cooperative) behavior of Ants, ACO can be used to find solutions for complex optimization problems if the parameters of the algorithm are carefully chosen. Inherent parallelism with an in-built positive feedback mechanism that helps in reaching the optimum solution faster, ACO have Issues of local optimal hence the need to enhance it, Increase in computation time as well

Memetic Algorithms (MA) was invented by Moscato, (1989) MA integrates EA (such as GA) and a local search procedure (e.g. hill climbing) to generate a hybrid algorithm with features of both its constituents powerful techniques of solving NP-hard optimization problems If the population converges, i.e. no further improvement is likely, For a good MA, solution representation must be carefully selected. Also careful design and use of operators, which are tuned to problem under consideration is highly important to generate better results.

Firefly algorithm invented Yang (2010) based on the flashing patterns and behavior of fireflies. Characterized by flashing light called bioluminescent Can be used for the hardest optimization problem NP-hard problem, suitable for non-linear problem, multi-swarming ability, does not use velocity, gbest and pbest hence not prone to the problems associated with them, scaling control making FA to be highly efficient and flexible, Proper tuning of parameter is required also not the earlier version of firefly not good for discreet variable, discretization, random permutation also modulus function have been made on the other versions of firefly to solve the discreet issues, nevertheless firefly still remain the most robust meta-algorithm for competitive result see session 2.4.4.5.

2.6 Review of Related works on Cervical Cancer

This section gives review of the various study that have been carried out on CC exploring the method used, the result obtained, strength, weakness and relevance to this work.

Nithya1 and Ilango1, (2019) proposed Evaluation of ML based optimized feature selection approaches and classification methods for CC prediction, The research work then intended to attain deeper understanding by applying MLT in R to analyze the risk factors of cervical cancer, this work then build few classifier models using C5.0, RF, KNN and SVM algorithms, C5.0 and RF classifiers have performed reasonably well with comprehensive accuracy for identifying women exhibiting clinical sign of CC, few dataset was used and as such is not enough to determine whether the model is robust enough to predict CC also poor data preprocessing stage, also better approach was also recommended for a more accurate prediction level hence the study.

Thendral and Lakshmi, (2019) carried out Performance Comparison of SVM Classifier Based on Kernel Functions in Colposcopic Image Segmentation for CC, this study proposed a method for automatic CC detection using segmentation and also the various types of kernel function in SVM classifier were also compared, Analysis shows that MLP kernel provides the best performance, Dataset used here was image, parameter selection difficulty affected the system accuracy and poor image preprocessing stage.

Akinrotimi and Olugbebi, (2018) proposed a Neuro-expert system was developed using ANFIS, taking into consideration the combination of eight attributes or factors and one output for the prediction and diagnosis of CC. CC dataset obtained from cancer medical

experts, was used to build the system. An evaluation performance was so as to carry out to determine the level of predictive and explanatory power of the developed system. With an accuracy of 93.54 percent, the resulting test carried out on the systems shows a very good predictive model. The attributes used were restricted so that future work can explore the use of greater sample space.

Fayz, Rizka, Abo and Maghraby, (2018) proposed cervical cancer risk factors to build classification model using RF classification technique with Synthetic Minority Oversampling Technique (SMOTE) and two feature reduction techniques RFE and PCA. To address the question of imbalances, SMOTE was used. In the dataset. Hinselmann, Schiller, Cytology and Biopsy dataset consists of 32 risk factors and 4 target variables:. After comparing the results, random forest classification technique with SMOTE improve the classification performance. There is, however, a strong need to use a different methodology that not only offers a better approach to the handling of this type of dataset, but also gives greater performance to that obtained in this work.

Juliana and Hassan (2018) proposed an ACO-based classification algorithm, Ant-Miner was deployed to analyze cervical cancer data set. The proposed method achieved higher accuracy when comparing with Wu and Zhou (2017) works. However, If only elitist rules instead of pruning each rule constructed by each ant are utilized, more accurate rules will be obtained and number of terms per rule will also be reduce, hence better system

performance, further more recommendation interface, narrowing down predicting target variable as researcher have been silent to this, also reduced feature obtained can further be utilized to develop model that will give better performance hence SVM-FA for CC detection.

Babatunde and Muhammad-Thani, (2018) proposed an ANFIS based Detection of CC, a hybrid intelligent system which combines the fuzzy logic qualitative approach and adaptive neural network capabilities, model to predict either High or low based on the input dataset, Patient with CC where diagnosed, The need for a larger sample space, Need for more enhanced validation, preprocessing of the dataset and also need for a comparative study with other state of art

Sahoo and Chandra, (2018) proposed Improved cervix lesion classification using multi-objective Binary FA (BFA) based feature selection, novel multi-objective BFA is proposed to solve the problem of feature selection. it was applied to a very crucial problem of classifying benign and malignant cervical lesions, Classification error and length of the selected feature subset were taken into consideration; thereby, addressing both the objectives of feature selection, randomized algorithm are needed to solve NP hard problems as to obtain accurate results.

Benazir and Nagarajan, (2018) developed an Expert System for Predicting the CC using DMT, The GA, PSO and ACO was used as the feature selection technique, NN with MLP back propagation classify the patients into two normal and abnormal, GA performing better, patient with cervical cancer were successfully diagnosed, Model need to be implemented on larger sample space, Increase computational time, slow convergence and Need for an enhanced feature selection

Ramit, Puneet, and Ravi, (2018) explores the inclusion of a penalty function to the existing fitness function promoting the BFA to drastically reduce the feature set to an optimal subset, and shows an increase in both classification accuracy as well as feature reduction using a RF classifier for the diagnosis of Breast, Cervical and Hepatocellular Carcinoma-Liver Cancer by the proposed approach in contrast to other contemporary approaches such as those focused on Deep Learning , Knowledge Gain and others. However, by experimenting with other classifiers, deciding the key predicting target variable and capitalizing on the risk factors obtained for developing a better model, this analysis can also be extended by contrasting studies with other approaches to feature selection.

Jiayong, Yanxi, and Tong (2018) presented a novel SVM-based feature screening method and applied it to multispectral PS image classification for CC detection. Using the new feature screening process, comparative studies show substantial improvements in the

accuracy of pixel-level classification. To further validate this research, a much larger Pap smear image collection and an even richer image feature space would aid.

Yong, Liu, Zhang, Zhu. and Zhao (2018) applied various algorithms commonly used in the field of machine learning to CC data classification and comparative research. Based on the extraction of various data features, three types of neural network structures, namely SVM, ANN, and KNN, are used to classify and identify data. The analysis of algorithms are performed in terms of accuracy. The results show that the three ANN introduced in this paper have good performance in the auxiliary diagnosis and classification of medical data and provide a basic way to improve the diagnostic intelligence level.

Wu and Zhou, (2017) study reveals some CC risk factors by using SVM based approaches for classification of CC dataset, he further showed that Both SVM-RFE and SVM-PCA are able to actualize the similar function with less features than SVM. More specially, SVM-RFE and SVM-PCA enjoy the capability to reduce the feature numbers from 30 to 8 to accomplish the classification while SVM based approach suffer high computational cost. Meanwhile, in all the three based approach the classification speed can be further be improved prominently hence SVM-FA based algorithm for CC detection.

Prabukumar and Agilandeeswari, (2016) proposed an Automatic Classification of CC Cell: A Case Study, The suitable features are selected using GA optimization technique. Over this ANN Classifier are used to conclude whether the input image is cancer affected or not, The simulation results indicate that ANN based model using GA optimization technique have the potential for diagnosis of cervical cancer in its early stage, features selection process was time consuming and the difficulty in understanding the ANN process has the neurons increased.

Kurniawati, Permanasari and fauziati, (2016) proposed Comparative Study on DM Classification Methods for CC Prediction using Pap Smear Results, 38 symptoms and 7 classes. Naïve Bayes, SVM, and RFT was used to evaluate the performance of the classifier. The performance matric that used in this study are accuracy, recall, precision, and ROC curve, RFT is the best classifier among other classifiers to classify PS results, degrading performance of SVM due to poor parameter and feature selection also Need for larger sample space

Mukhopadhyay, Kurmi, Dey, Das and Pradhan, (2016) SVM optical diagnosis of colon and CC, the efficacy of SVM-based classification with various kernels has been tested on multifractal parameters such as Hurst exponent, singularity spectrum width to classify cancer tissues, SVM classified cervical cancer patient, raising the need to increase cervical cancer precision, specificity and sensitivity

Mutgi, Murthy, and V, (2015) proposed a NN Based Automated System for the Diagnosis of CC Project, MATLAB and image processing tools was used to extract features from cytology images which was then used to train the neural network, The system was able to classify the images as non-cancerous, low- grade and high-grade cancer cells however Appropriate CC images were not extracted hence accuracy was affected, Difficulty in analyzing countless of CC images and Minimal amount of images for training the model.

Kourou, Exarchos, Exarchos, Karamouzis and Fotiadis, (2015) carried out a review on ML applications in cancer prognosis and prediction, the study extensively discussed the concepts of ML, the work also outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the review of their findings, it is clear that the incorporation of multidimensional heterogeneous data, coupled with the use of various feature selection and classification techniques, can provide promising CC detection inference tools.

CHAPTER THREE
METHODOLODY

3.1 Project Approach

The experimental framework of the study is expressly stated in sequential steps in view of achieving the stated objectives. The experiments are designed to optimize RBF kernels parameters of the SVM with the best features obtained from FA for CC detection. The system was designed to get an optimal subset from the dataset so as to avoid outliers, under-sampling and oversampling in the dataset designed to simulate a data mining sequential process. SVM has been good classification algorithm but to get a more efficient result both feature selection and parameter optimization were performed. The FA was introduced to determine the best features vital enough for obtaining optimal parameters of RBF kernels for effective and efficient CC detection process.

Figure 3.1 show a flowchart that describes the steps involved for actualizing the study aim, this includes, the collection CC dataset from kaggle dataset repository, and then feature selection was performed by optimizing the high dimensional dataset for better classifier performance. This way main factors and the life style responsible for cervical cancer are determined, the best features are then capitalized on for optimizing the Cost C and Gamma Y parameter of the RBF kernels which in turn were used for training the SVM for CC detection. The dataset was partitioned into two fold the training and testing set. Results obtained from the developed model were compared with the existing state of

the art.The training set and testing set of data was divided at a percentage ratio of 75% to25% respectively. The results was evaluated based on ML statistical metrics like the classification accuracy, true positive rate, false negative rate, error rate, and specificity, sensitivity, testing and training time. These process was implemented on a high-performance language for technical computing created by The MathWorks developed in 1984 called MATLAB.

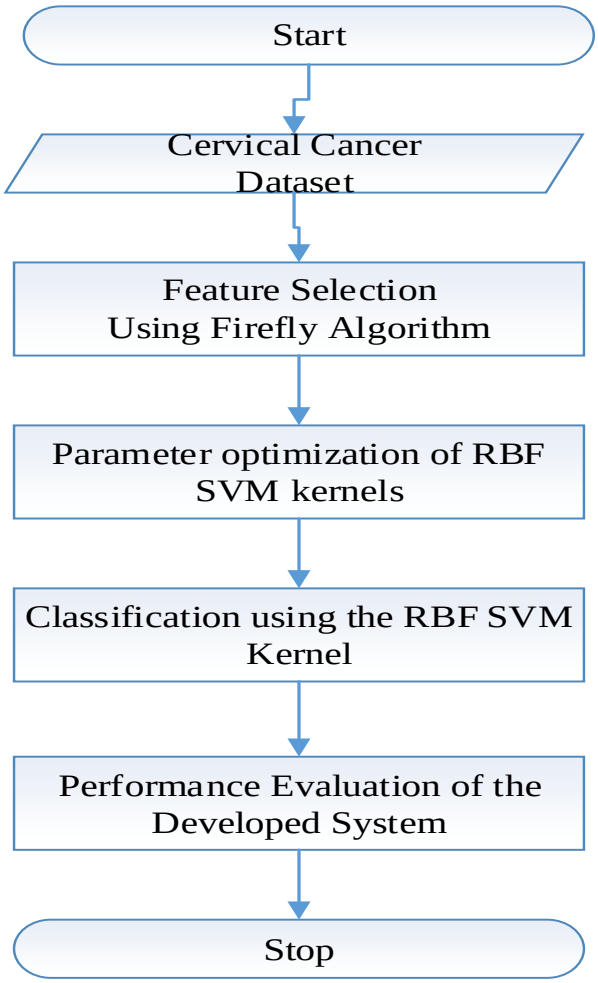


Figure 3.1: Flowchart for the developed System

3.2 Procedural Techniques for the Developed System

The methodology this study adopt to achieving the study aim, was a combined Knowledge Discovery in Databases (KDD) and learning technique using the FA and SVM Classification Algorithm. The data mining techniques adopted are shown in these order, from data acquisition, through feature selection, parameter optimization and then classification.

3.2.1 Dataset acquisition

A dataset was used to formulate a knowledge base for the system so as to build an efficient system. The dataset was obtained from the University of California, Irvine (UCI) repository machine. The Dataset consists of 32 attributes and 4 target variables but Biopsy test was employed based on high test rate from previous literature. This dataset was then passed unto the FA for feature selection stage as shown in Figure 3.1 and 3.2. This is to determine the extent to which the FA can avoid extra preprocessing stage which can further add computational overhead to effective and efficiency of the developed system. Attribute description of the various dataset are shown in Table 3.1.

Table 3.1 Attribute Definition

Feature Index	**Feature Name**
1	(int) Age
2	(int) Number of sexual partners

3	(int) First sexual intercourse (age)
4	(int) Num of pregnancies
5	(bool) Smokes
6	(bool) Smokes (years)
9	(int) Hormonal Contraceptives (years)
10	(bool) IUD
11	(int) IUD (years)
12	(bool) STDs
13	(int) STDs (number)
14	(bool) STDs:condylomatosis
15	(bool) STDs:cervicalcondylomatosis
16	(bool) STDs:vaginalcondylomatosis
17	(bool) STDs:vulvo-perinealcondylomatosis
18	(bool) STDs:syphilis
19	(bool) STDs:pelvic inflammatory disease
20	(bool) STDs:genital herpes
21	(bool) STDs:molluscumcontagiosum
22	(bool) STDs:AIDS
23	(bool) STDs:HIV
24	(bool) STDs:Hepatitis B

25	(bool) STDs:HPV
26	(int) STDs: Number of diagnosis
27	(int) STDs: Time since first diagnosis
28	(int) STDs: Time since last diagnosis
29	(bool) Dx:Cancer
30	(bool) Dx:CIN
31	(bool) Dx:HPV
32	(bool) Dx
33	(bool) Hinselmann: target variable
34	(bool) Schiller: target variable
35	(bool) Cytology: target variable
36	(bool) Biopsy: target variable

3.2.2 Data set pre-processing and Normalization

Removing errors and outliers that may be present in the data are components of the pre-processing task that should be performed to make the information appropriate for modeling. In other to avoid inconsistency, imbalance and missing response normally prone to CC dataset appropriate values were assign to FA parameters hence smooth operation on the dataset were achieved for better optimization processes.

3.2.3 Feature Selection

Feature selection is often referred to as the selection of variable subsets, variable selection, reduction of features, or selection of attributes. Feature Selection (FS) is the tool that can be used in training materials to delete features, so that features that are statistically uncorrelated with class labels are removed and also reduce the set of features to be used in classification hence better performance of the classifier. In this study FA was used to extract the relevant features only pertinent enough to obtaining the optimal parameter of the C and γ parameter. FA at this second stage was used to determine the risk factors responsible for CC, the risk factors were ascertained in a descending format (highest to lowest factors causing CC) risk factors obtained will then be capitalized on for further stages.

3.2.3.1 Objective Function (OF)

Objective Function (OF) employed by Sahoo and Chandra (2018) was used in this study, equation (3.1) describe the OF, NFS is the Number of Features Selected, N is the total number of features, while the last equation is the accuracy of the classifier. *c*1 and *c*2 are constant terms they are weights assigned to both the terms. Their value lies between 0 and 1. Since, classification performance is assumed to be more important than the number of features being selected, *c*1 was set to be smaller than *c*2. The value obtained from OF was used as the initial Light intensity *I* FA.

$$\text{Objective Function (OF)} = c1 * \frac{nfs}{n} + C2 * \frac{FP+FN}{TP+TN+FP+FN} \quad (3.1)$$

3.2.4 Training and Classification

This study aim to predict CC using the FA and SVM. The dataset used was divided into training and testing set at a proportion of 75% to 25% respectively. The training set is passed to the RBF-SVM algorithm so as build an optimized knowledge retention or pattern from the dataset using an optimal FA parameters. The FA-SVM initialization is described in section 3.2.4.1.

3.2.4.1 Firefly Algorithm and Support Vector Machine

The RBF-SVM data classification process begins by initializing the parameters needed for the firefly algorithm search process to determine the number of firefly populations (number of fireflies), the generation (maximum generation), the initial coefficient of attractiveness (β0), the coefficient of light absorption (γ) and the coefficient of random parameters (af). Optimize the parameters C and σ using the optimal FS from FA after initializing the parameters required. There are several steps used in optimization, namely:

i. Determine the objective function f(x)

ii. Initialize the population of firefly noodles (i = 1, 2, N)

iii. Define the light absorption coefficient γ

iv. Calculate the distance

At the location of the coordinates xi and xj, the distance between two fireflies I and j is the Cartesian distance, which is formulated as

$$rij = ||Mi - Mj|| = \sqrt{\sum_{k=1}^{d} (Mi,k - Mi,k)2} \quad (3.2)$$

If an equation uses dimensions (d = 2) then the above equation becomes Equation (3.3)

$$rij = \sqrt{(Mi - Mj)^2 + (ni - nj)^2} \quad (3.3)$$

Calculate Attractiveness of a firefly is proportional to the intensity of light seen by other fireflies. Attractiveness is formulated by Equation (3.4):

$$\beta(r) = \beta 0 e^{-\gamma r^2} \quad (3.4)$$

Information:

$\beta(r)$ = Attractiveness of fireflies at distance r

$\beta 0$ = Attractiveness at distance 0

γ = coefficient of light absorption

r = distance between source fireflies and fireflies

6. Calculate the movements of fireflies

Firefly movements that are attracted to fireflies j (which are brighter or have higher attractiveness) are formulated as shown in Equation (3.4). in general the parameter values used are $\beta 0 = 1$ and $\alpha \in [0, 1]$. The randomization process can be done using a normal distribution of N (0.1) or another distribution. After the values C and σ are obtained, the value is used to train the RBF-SVM. After the SVM model is used to train the data, and accuracy of C and σ determined. Ranking is then performed to evaluate the C and σ parameter values that are most optimal. Then the values of C and σ are used to model the SVM classifier then the model is tested using the test data. Accuracy based on the test data would be acquired from the test results.

3.3 Frame Work for the Developed System

The frame work for the developed system are described in figure 3.2. The model gives a detail description of how the study aim and objective were realized. Cascading processes of the various block are depicted. The first stage shows the input block called the dataset block containing 36 attribute. The output from the input block was fed to the FA block, the 36 attribute was reduced to 15 attributes, and attribute obtained from FA was used to optimize the C and Y of the RBF. the fourth stage showed classification stage were optimized parameter was used for training and classification of the model, to achieved this CC cells were partitioned to 75% for training and 25% for testing, based on the user input the overall output stage was able to detect if a patient was cancerous or not cancerous, user friendliness and easy operation was achieved. Based on the result obtained recommendation were made to patients for treatments.

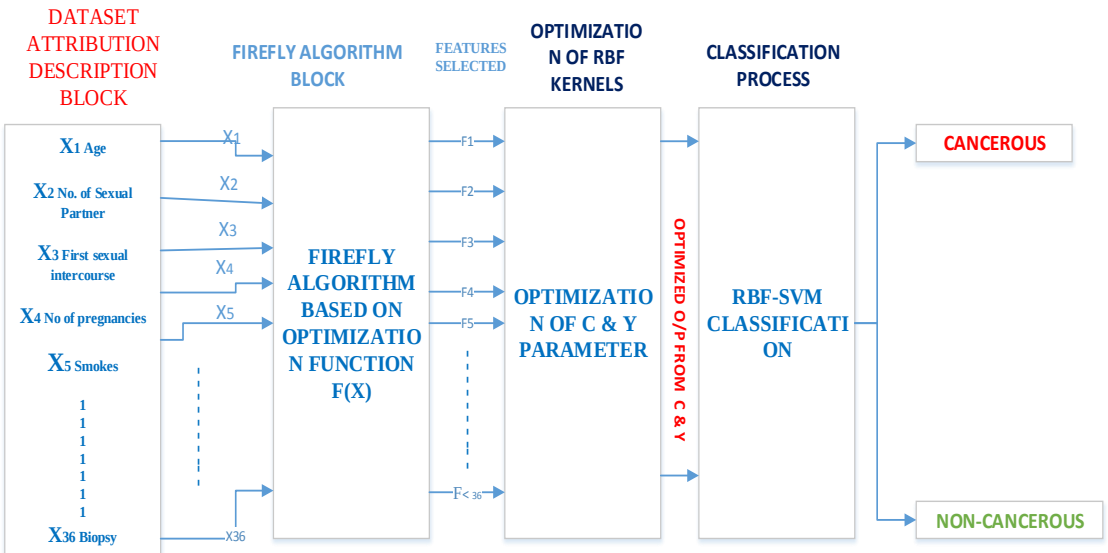


Figure 3.2: Frame Work for the Developed System

3.4 Performance Evaluation Parameters

To determine the effectiveness of the developed model, statistical yardstick was used to measure the effectiveness of the model in terms of the positive predictive rate, negative predictive rate as well as the classification accuracy (Zhu, Zeng & Wang, 2017). Example are the True positive (TP) rate, True Negative (TN), False Positive (FP) rate, False Negative (FN) rate, from this classification computational timing, accuracy, sensitivity (recall) and specificity were derived. If TP Prediction is +ve and patient is cancerous, this is a desirable result, TN Prediction is -ve and patient is non-cancerous, it is also a desirable result with FP Prediction is +ve and patient is non-cancerous, this is a false alarm, and can be regarded as bad and lastly with FN Prediction is -ve and patient is cancerous, this is worst of all and must be avoided (Zhu, Zeng and Wang, 2017).

This study detect CC by giving its output as either cancerous (+ve) or non-cancerous (-ve). To achieve that if it starts with True then the prediction was correct whether cancerous or not, so TP is a cancerous patient correctly predicted and a TN is a healthy patient correctly predicted. Oppositely, if it starts with False then the prediction was incorrect, so FP is a healthy patient incorrectly predicted as cancerous (+) and a FN is a cancerous patient incorrectly predicted as healthy. Positive or negative indicates the output of our program (Baratloo, Hosseini, Ahmed & Ashal, 2015). Correct or incorrect output is judged by true or false . Based on all these the performance Parameters are as described.

i. Classification Accuracy (CA)

The consistency of a test is its ability to accurately distinguish between the patient and stable cases. In order to estimate the accuracy of a test, in all assessed cases, we can determine the proportion of true positive and true negative (Baratloo, Hosseini, Ahmed & Ashal, 2015). See equation (3.5) mathematical illustration of CA:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

ii. Sensitivity (Recall)

The number of True Positives divided by the number of True Positives plus the number of False Negatives is called Recall. Put another way, the number of positive predictions in the test data is separated by the number of positive class values. It is also called Sensitivity or the True Positive Rate (Baratloo, Hosseini, Ahmed & Ashal, 2015). Mathematically, this can be described as shown equation (3.6).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.6}$$

iii. Specificity

A test done to determine the ability of healthy cases correctly is termed specificity. To estimate it, proportion of true negative in healthy cases are calculated. Mathematically, this can be stated as shown in equation (3.7).

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.7}$$

iv. Precision

The sum of True Positives divided by the number of True Positives plus False Positives is precision. In other words, it is the number of positive predictions separated by the total number of predicted positive class values. Good predictive value (PPV) is often referred to. Mathematically, this can be stated as shown equation 22:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.8}$$

3.5 Recommender System

An application interface will be developed with the MATLAB programming language in order to create a user friendly interface that will cater to unfamiliar dataset that will be analyzed on individual basis by the system which will scales on the best FA-SVM optimal classification parameters.

3.6 Choice of Programming Tools

MATLAB is a high-performance language developed by MathWorks in 1984 for technical computing. In an easy-to-use environment, it combines computation, visualization, and programming where problems and solutions are presented in familiar mathematical notation. Typical utilizations include:

i. Creation of Algorithm

ii. Acquiring data

iii. Modelling, prototyping, and simulation

With feedback from many users, MATLAB has evolved over many years. It is the basic teaching method in university environments for introductory and advanced courses in mathematics, engineering, and science. In industry, MATLAB is the instrument of choice for study, growth, and analysis of high productivity.

MATLAB is an environment and programming language for numerical computing. It makes it possible to manipulate the matrix quickly, to plot functions and data, to implement algorithms, to construct user interfaces and to interface programs in other languages. An optional toolbox interface with the Maple symbolic engine, although it specializes in numerical computation, enables it to be part of a complete computer algebra scheme. It can handle differential equations, polynomials, signal processing, and other applications, in addition to working with explicit matrices in linear algebra. Results can be given both numerically and as outstanding graphics. (Products / Mathworks.com).

Within a fraction of the time it would take to write a program in a scalar non-interactive language such as C or FORTRAN, MATLAB solves several technical computing problems, especially those with matrix and vector formulations.

The MATLAB name stands for Laboratory of Matrix. To provide easy access to the matrix software developed by the LINPACK and EISPACK ventures, MATLAB was originally written. Today, the LAPACK and BLAS libraries are built into MATLAB engines, embedding the state of the art in matrix computation applications. A family of add-on application-specific solutions called Toolboxes are available in MATLAB. Toolboxes allow advanced technology to be learned and implemented. Toolboxes are

extensive sets of MATLAB (M-files) functions that broaden the MATLAB environment to solve unique problem classes. Signal processing, control systems, neural networks, fuzzy logic, wavelets, simulation, and many others are the areas in which toolboxes are usable. As of 2004, it was reported that MATLAB was used by more than one million people in industry and academic. In this study MATLAB R2016a (9.0.0.341360) was used.

CHAPTER FOUR

RESULT AND DISCUSSIONS

4.1 Overview to Result Analysis of the Developed System

This section presents the results analysis of the data science approach used in achieving the stated objectives. The experiment was designed to optimize SVM parameters with optimal FA features, as to ensure efficient and effective CC detection process. The results of each stage was designed to obtain the best model which was in turn compared with the state of arts. The developed model was recommended for used in diagnostic centers and further improvement by future researchers.

The implemented system elaborated a stepwise approach of DM technique, as to achieve a high prediction rate for CC process, these were systematically done as follow: CC data collection from kaggle dataset repository, secondly the filtering of data was achieved, which helps to remove outliers and inconsistent data, thirdly the firefly optimization was conducted to determine the risk factors. The adopted objective function which was wrapped with the classification accuracy of the classier was further wrapped with FA obtain the most optimal feature subsets.

Accuracy at the FS stage was competitive with existing studies, the dataset was successfully partitioned into two fold the training and testing set at a percentage ratio of 75% to25% respectively. The results were evaluated based on machine learning statistical

metrics like the classification accuracy, true positive rate, false negative rate, error rate, specificity, sensitivity and training and testing time.

The experimental setup was successfully carried out and developed with the Matlab programming (MATLAB 2016A) with interwoven connected component of the Matlab Guide platform so as to create a friendly user experience. The developed systems made use of various component environment in Matlab for successful presentation of the result from the various DM stages namely data filtering, feature selection, classification and performance evaluation.

4.2 The MATLAB Command Window

The Matlab command window helps to display result of the data mining task in a console screen for readability and easy expression of output see Figure 4.1.

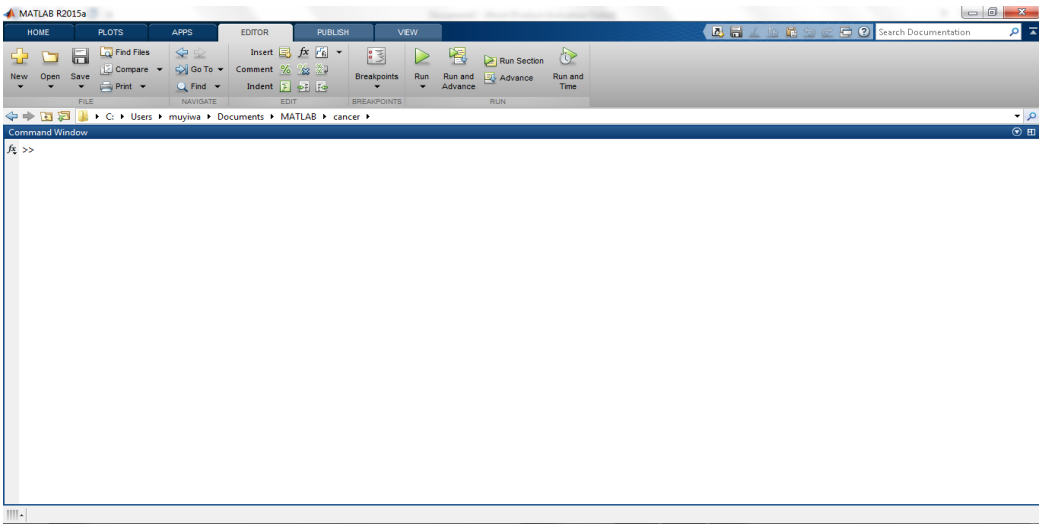


Figure 4.1: The MATLAB Command Window

4.3 Interactive Developmental Stage

The optimized data mining model results was presented based on the following outcomes:

i. Dataset Acquisition
ii. Dataset Filtering
iii. Firefly optimization of the filtered data
iv. Parameters optimization of the RBF-SVM Kernel by using the result from (3)
v. Training of the optimal model obtained in 4
vi. Classification of cancerous cells and the Non-cancerous cells
vii. Cervical Cancer Consultation Interface
viii. Performance Evaluation of the developed model.

Figure 4.2 shows an outlook of the initial startup for the developed model, at run time all modules vital to actualizing the system aim were automated into the working process of developed platform. The interactive graphic user interface automates the loading of data via the load button which loaded the dataset into the platform, the next phase carried out the feature optimization with the FA which resulted to successful selection of optimal feature subset for parameter optimization stage. The reduced features subset when passed into the SVM (RBF kernel) obtained optimal parameter of the RBF kernel at C=2 and Y=0.9. these were successfully used as standard to Train the dataset, validation of the test performance based on the test dataset and as well as to predict new instances.

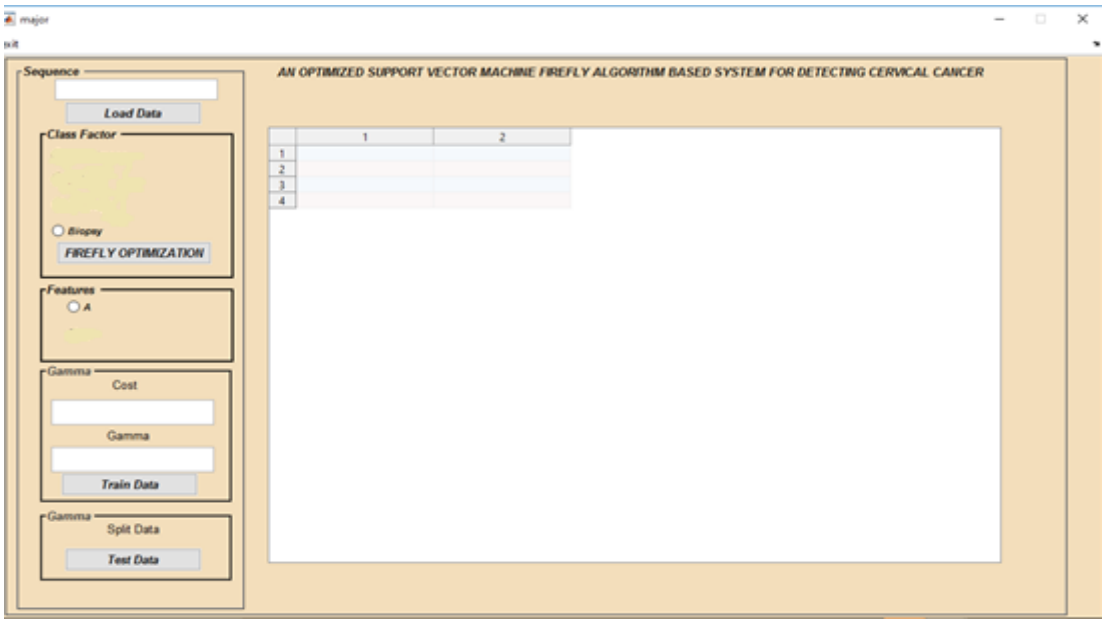


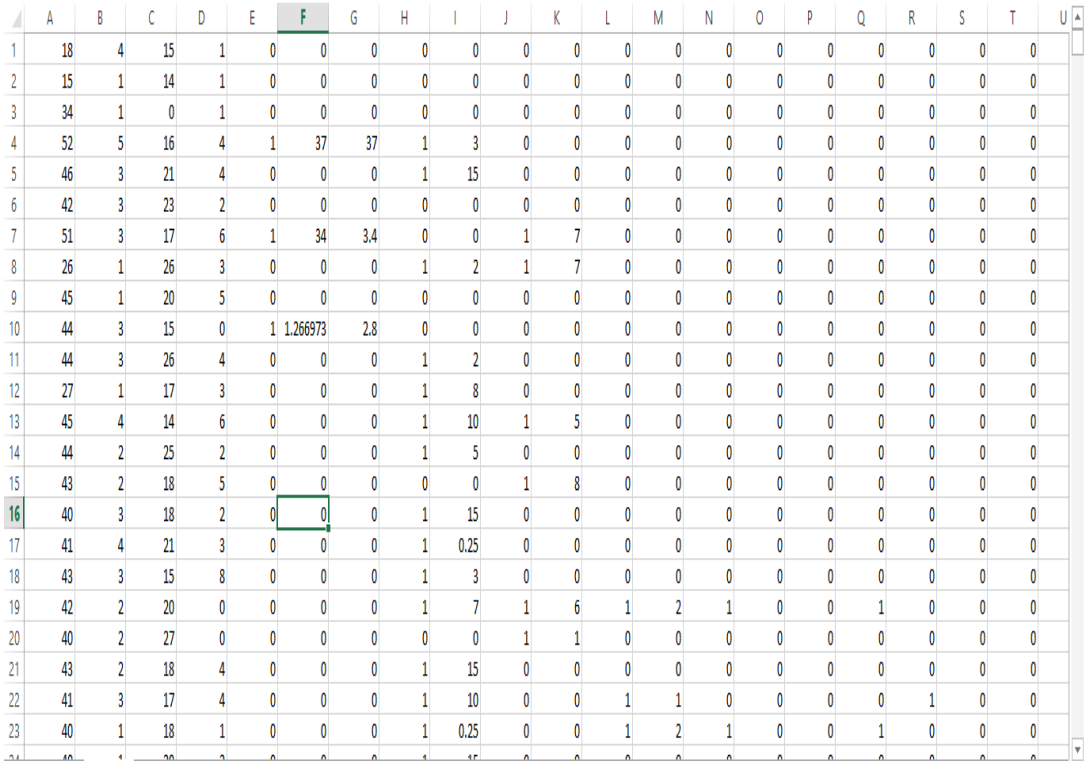
Figure 4.2: Initial start-up of the developed system

4.4 Results Analysis for the Developed System

This section presents the results obtained from each section accordingly.

4.4.1 Dataset Filtering

The filtered data helps to present a well formatted data into the system the data was filtered by converting string variable to numeric variable and removing inconsistent entries. The filtered and normalized data is shown below. An inconsistent factor were also eliminated during this phase attribute 27 and 28 were remove since they were insignificant for the mining process, this was due to missing values present. This was also in line from report from previous researchers. Figure 4.3 therefore gives a detailed look of the filtered data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	34	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	0	
5	46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	
6	42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	51	3	17	6	1	34	3.4	0	0	1	7	0	0	0	0	0	0	0	0	0	
8	26	1	26	3	0	0	0	1	2	1	7	0	0	0	0	0	0	0	0	0	
9	45	1	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	44	3	15	0	1	1.266973	2.8	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	44	3	26	4	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	
12	27	1	17	3	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	
13	45	4	14	6	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0	0	
14	44	2	25	2	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	
15	43	2	18	5	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	
16	40	3	18	2	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	
17	41	4	21	3	0	0	0	1	0.25	0	0	0	0	0	0	0	0	0	0	0	
18	43	3	15	8	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	
19	42	2	20	0	0	0	0	1	7	1	6	1	2	1	0	0	1	0	0	0	
20	40	2	27	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
21	43	2	18	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	
22	41	3	17	4	0	0	0	1	10	0	0	1	1	0	0	0	0	1	0	0	
23	40	1	18	1	0	0	0	1	0.25	0	0	1	2	1	0	0	1	0	0	0	

Figure 4.3: Dataset Filtering

4.4.2 Loading of the Dataset

The figure 4.4 shows the loading of the predicting variables which are also the independent variables that gives precise information about the target which is also the response variable. A total number of 858 instances, 30 attributes with one response/class label was loaded into the system for evaluation.

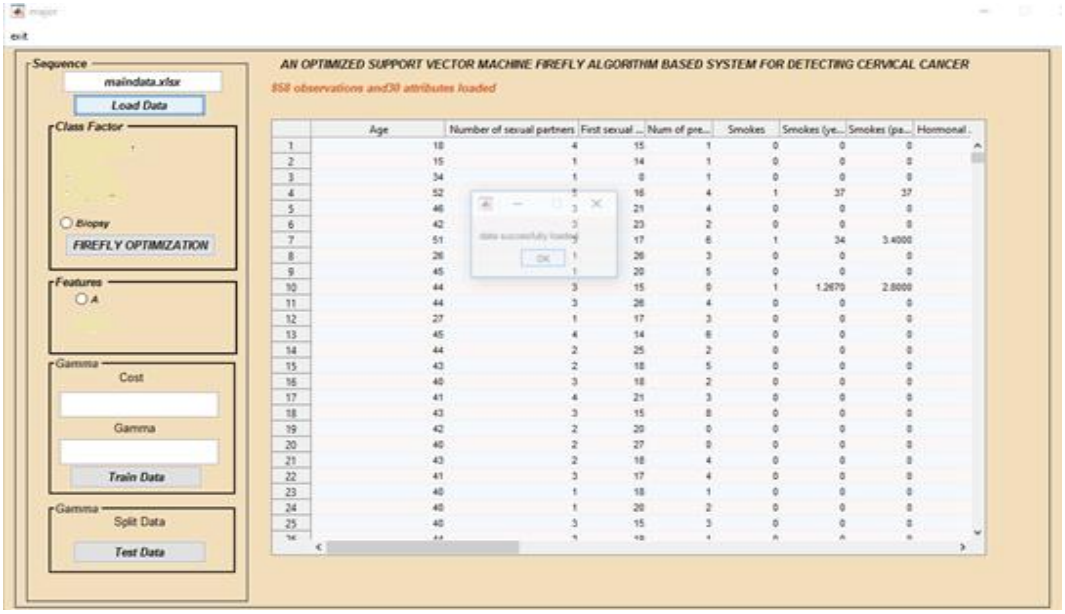


Figure 4.4: Loading of Dataset into the application

4.4.3 Firefly Algorithm and SVM Feature Selection Optimization

After the dataset was presented to the developed model, the features subset vital for the optimization of RBF kernels was achieved, the wrapper technique which collaborated the hybridization of SVM and FA with the adopted Objective function as shown in equation 13 was used, result was obtained when the number of fireflies are 10, the parameters of the fireflies were set at $\alpha = 0.1,\ \beta = 1$ and γ=1. Figure 4.5 illustrate the feature selection stage for Biopsy response variable, Result obtained from the FA stage is described in Table 4.1. The attribute SV represent Selection Value, NF is number of Fireflies, NFS is number of features selected and OPT is Optimization Time. The results given in table 4.1 showed that the accuracy, OPT and SV of biopsy gave a satisfactory result when FA was

used to select vital feature without necessarily performing data preprocessing that may have had overhead to the stage.

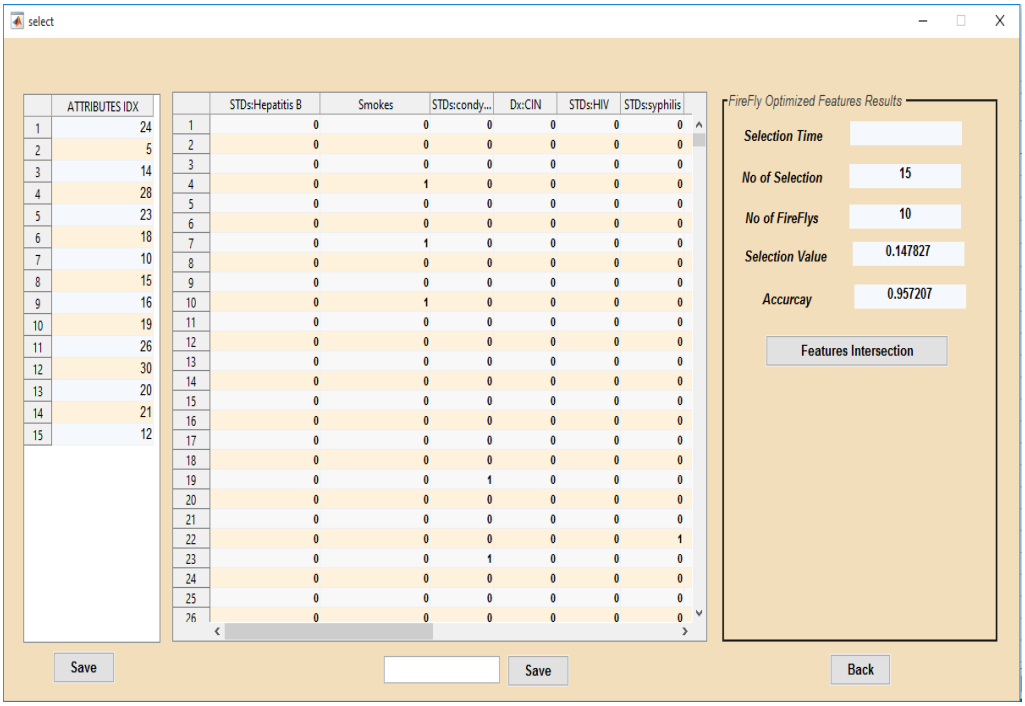


Figure 4.5: Feature Selected with FA with Biopsy as response Variable

Table 4.1: Result obtained After Features were selected using Firefly Algorithm

Target Variable	Accuracy (%)	SV	NF	NFS	Training/OPT
Biopsy	95.721	0.147827	10	15	40.20134

4.4.3.1 The Feature Subset (Risk Factors) Obtained From Using Firefly Algorithm

This section therefore describe the feature obtained after FA was used for FS, Table 4.2 therefore showed that only fifteen (15) attribute were selected after optimization, in order to detect CC, the order of importance of this risk factors were described in descending order (order of importance) as obtained from the FA output hence 30 features were reduced to 15.

Table 4.2: Risk Factors Obtained From Using Firefly Algorithm

No of features	Biopsy
1	'STDs:HIV'
2	'STDs: Number of diagnosis'
3	'Dx'
4	'STDs:HPV'
5	'STDs:cervical condylomatosis'
6	'STDs:vaginal condylomatosis'
7	'Dx:HPV'
8	'STDs'
9	'Smokes'
10	'Dx:CIN'
11	'Dx:Cancer'
12	'STDs:pelvic inflammatory disease'
13	'STDs:syphilis'
14	'STDs:molluscum contagiosum'
15	'STDs:condylomatosis'

4.4.4 Support Vector Machine (RBF Kernel Optimization)

After the feature selection stage the selected features from the FA were optimized with RBF-SVM kernels parameter Cost (C) and gamma (Y).The C values were chosen from the range of 1-3 1.e 1, 2, to 3 while the gamma ranges between 0.1 to 0.9. i.e. 0.1, 0.2, 2.3 to 0.9 result showed that optimal parameter was obtained at C= 2 and Y=0.9, giving an optimization accuracy of 96.4847 see figure 4.6 for the training process. Detailed description of the simulation of the Optimized results for Cost and Gamma Parameter of the RBF, SVM Kernels are illustrated in the table 4.3: optimization accuracy obtained are as shown in Table 4.4

Table 4.3: Optimized results for Cost and Gamma Parameter of the RBF, SVM Kernels

C	Y	Accuracy (%)
1	0.1	93.6335
1	0.2	93.6335
1	0.3	93.6335
1	0.4	93.6335
1	0.5	93.9441
1	0.6	93.9441
1	0.7	93.9441
1	0.8	94.0994
2	0.1	93.6335
2	0.2	93.6335
2	0.3	94.0994
2	0.4	94.0994
2	0.5	94.0994
2	0.6	94.0994
2	0.7	94.0994
2	0.8	94.2547

2	**0.9**	**96.4847**

Table 4.4: Accuracy Obtained at Optimal Value of C and Y

Feature	Optimization Accuracy (%)
15 Risk Factors from the FA	96.4847

4.4.5 Training of the Developed Model

This stage projected the data into training and testing set. The system used 75% of the data for training set which was trained with the most optimal SVM (RBF Kernel) optimized parameter was attained when set as C= 2 and Y= 0.9. See Figure 4.6 for parameter optimization and the training process. The training set helps to build the knowledge retention of the SVM Model while the 25% was used as testing set. That is to validate the performance of the developed model.

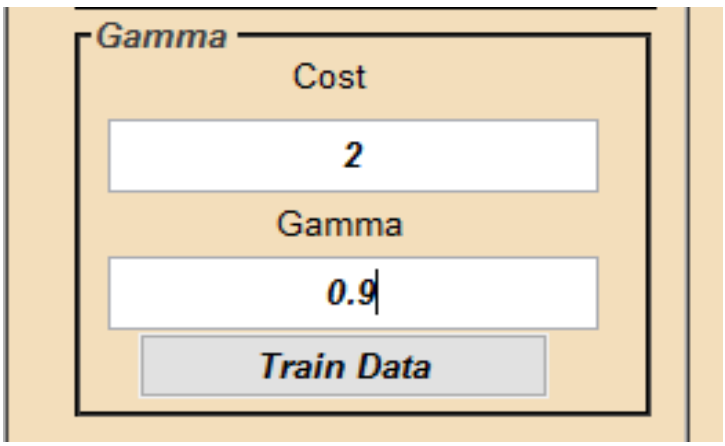


Figure 4.6: Training Process

4.4.6 Testing Phase

This testing phase is triggered by the test data button which outputs the validation results and statistical metrics, the testing interface is as shown in Figure 4.7.

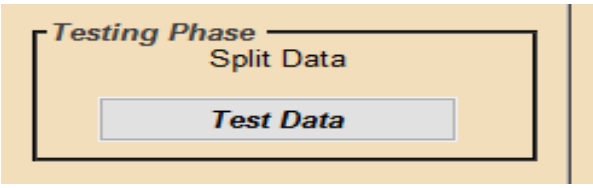


Figure 4.7: Testing Process

4.4.7 Cervical Cancer Consultation Interface

This session provide a consultation platform for CC patients, it generally refers to the results and information that are generated by the system for CC patient; it is the main reason for developing the system and the basis on which CC patient can determine their CC status, providing recommendation for necessary action to hasten early treatment. Thereby preventing CC critical stage. Figure 4.7 show the response from the patient and the system was able to detect if the patient was positive or negative to CC.

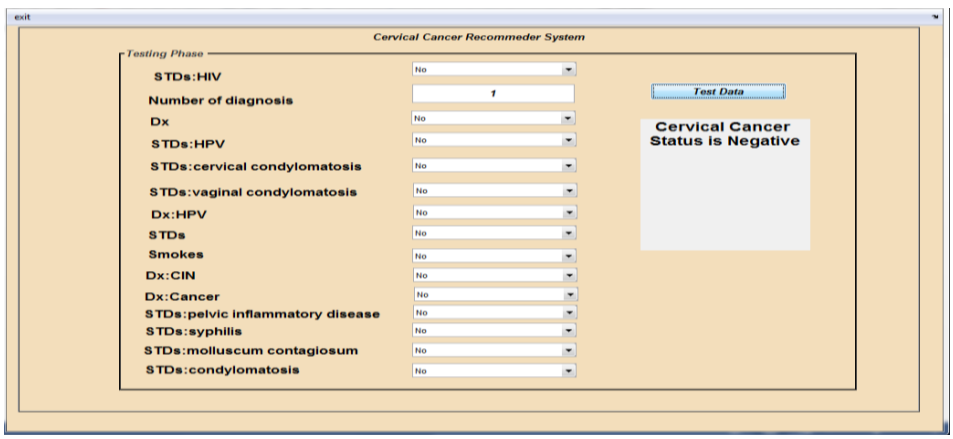


Figure 4.8: Cervical Cancer Consultation Interface

4.4.8 Experimental Results Evaluation

The experimental results are listed based on the classification algorithm. The evaluation parameter shows the result of the RBF-SVM classifier as it was obtained. The testing (probing) evaluation was achieved using the True Positive rate (TP), False Positive (FP), True Negative (TN) and, False Negative (FN), accuracy and error rate as well. The evaluation parameters for classification rate were achieved Classification Accuracy, sensitivity, Specificity and Error Rate. These Statistical Machine learning Results for the developed model obtained are illustrated in the graphical interface shown in Figure 4.9.

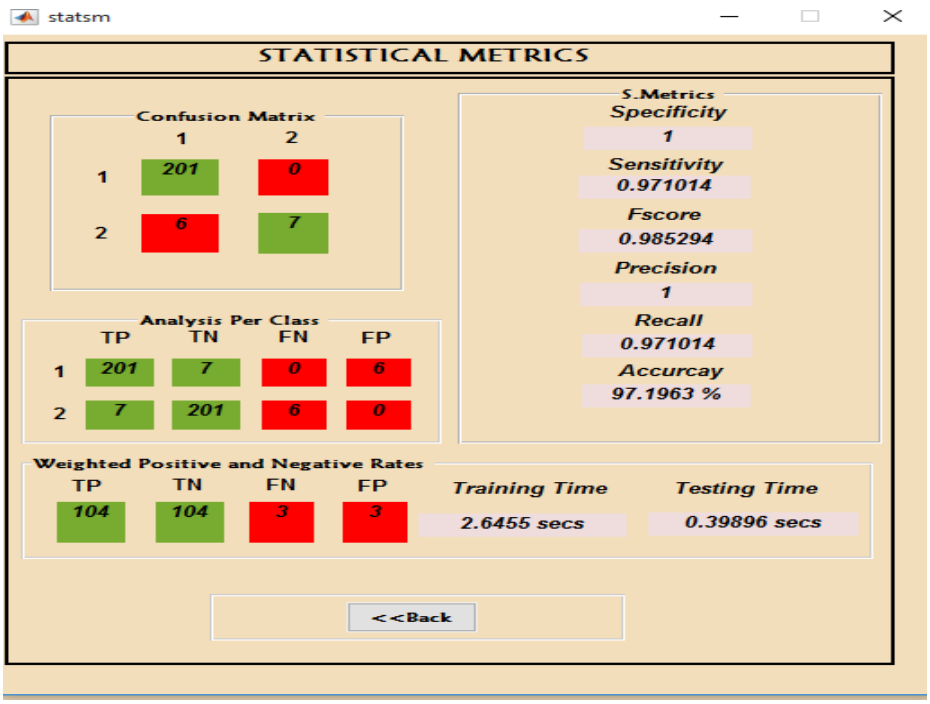


Figure 4.9: Statistical Machine learning Results for the Developed Model

4.4.8.1 SVM (RBF Kernel) Optimal Classification of the Selected Features

Table 4.5 provides an analysis per each class based on the class label from the class represented in the biopsy response variable. The table highlights the true positive value, the true negative value, false positive value and false negative value of each of the class groups.

A. Analysis per class

Table 4.5 Analysis per class

Analysis per class.	True Positive	True Negative	False Positive	False Negative
Class 1 (Non-Cancerous)	201	7	6	0
Class 2 (Cancerous)	7	201	6	0

B. Confusion Matrix

Confusion matrix is a summary of prediction results on a classification problem (see Table 4.6). The number of correct and incorrect predictions are summarized with count values and broken down by each class based on the testing set of data. The class 1 represents the non-cancerous class which gave a total of 201 from the test observation set, a total of 201 was classified correctly and 0 was misclassified, the cancerous class is represented by label 2, which gave a total of 13 from the test observation set, a total of 7 classified correctly and 6 was misclassified.

Table 4.6 Confusion Matrix

	1	2
1	201	0
2	6	7

C. Evaluation Parameters for Classification Phase

The table 4.7 shows the evaluation parameters of the SVM RBF kernel for the selected features based on the most optimal response variable biopsy based on the F-score, specificity, sensitivity, accuracy and error rate.

Table 4.7: Evaluation Parameters for Classification Phase

Technique	**F-score (%)**	**Specificity (%)**	**Sensitivity (%)**	**Accuracy (%)**	**Error rate (%)**
FA-RBF-SVM	98.5294	100	97.1014	97.1963	2.8037

D. Result of System Computational Time (CT)

Table 4.8 gives details of the actual CT for the Developed FA-RBF-SVM Model in terms of the training and testing time, it was measured in terms of the total seconds used time

for executing the training process as well as the time taken for a test or probe instances to be executed The results is shown Table 4.8.

Table 4.8 Computational Time of the Developed Model (Biopsy Response Variable)

Timing Results	**Training Time (Secs)**
Training Time	2.6455
Testing Time	0.39896

4.4.9 Comparative Evaluation of the Proposed Model with Different State of Arts using the Performance Parameters

In order to ascertain the level of improvement in the bid to developing timely, efficient and effective CC detection system, the FA-RBF-SVM was compared with existing state of art. This session therefore gives a detailed illustration of technique employed by several researchers and result obtained so far for CC detection. Table 4.9 showed a description of result from this FA-RBF-SVM, four attribute were used these are NFS, representing Number of Features Selected, CA is Classification Accuracy, SN is Sensitivity and SP is Specificity. Technique are (SVM-RFE for SVM-Recursive Feature Extraction), SVM-PCA (SVM-Principal Component Analysis), SMOTE (Synthetic Minority Oversampling Techniques) –RF (Random Forest) which was applied to RFE and PCA, BFA + Penalty + RF (Binary Firefly Algorithm + Penalty Parameter + RF), Ant-Miner Ant colony optimization based classification algorithm, Ant-Miner. BDT

(Boosted Decision Tree) DF (Decision Forest) and DJ (Decision Jungle). Result showed that the developed FA-RBF-SVM model had the highest CA of 97.20% with a very good specificity and sensitivity rate of 100% and 98.04% respectively. Outperforming existing works in terms of CA and SP with a competitive result as regarding SN. Hence this will be the first time for CC detection that feature selection and parameter optimization of SVM kernels were performed, resulting to better performance of the SVM when compared to existing state of art see table 4.9.

Table 4.9 Comparative Evaluation of the Developed Model with Different State of Art using the Performance Parameters

AUTHORS	Techniques	NFS	CA (%)	SN (%)	SP (%)
Developed Models	**FA-RBF-SVM**	**15**	**97.20**	**97.10**	**100**
(Wu & Zhou, 2017)	SVM	30	94.13	100	90.21
	SVM-RFE	18	94.03	100	90.05
	SVM-PCA	11	94.03	100	90.05
(Fayz *et al*.,2018)	SMOTE-RF	30	96.06	94.55	97.51
	SMOTE-RF-RFE	18	95.87	94.42	97.26
	SMOTE-RF-PCA	11	95.74	94.16	97.76
Ramit *et al*.,2018)	BFA + Penalty + RF	31	97.06	-	-

Juliana & Hassan, 2018)	Ant-Miner	30	94.76	-	-
Alam *et al*.,2019)	BDT	32	93.7	-	-
	DF	32	88.0	-	-
	DJ	32	86.3	-	-

4.4.10 Graphical Analysis

The graphical analysis shows a comparative results of the training time, classification accuracy, specificity, sensitivity, error rate and F-score of the various stages until the developed FA-RBF-SVM model was achieved.

4.4.10.1 Accuracy Obtained at Feature Selection stage using FA + SVM

The CA shows the correct classification rate attained from the FA after the desired feature were selected. The classification accuracy in percentage shows the percentage of instances that were classified correctly. The classification accuracy results shows that with FA 95.7207% was achieved see Figure 4.10. This shows that FA performance was satisfactory with CC dataset. This result obtained provided a solid bedrock for realizing the remaining objective of this project.

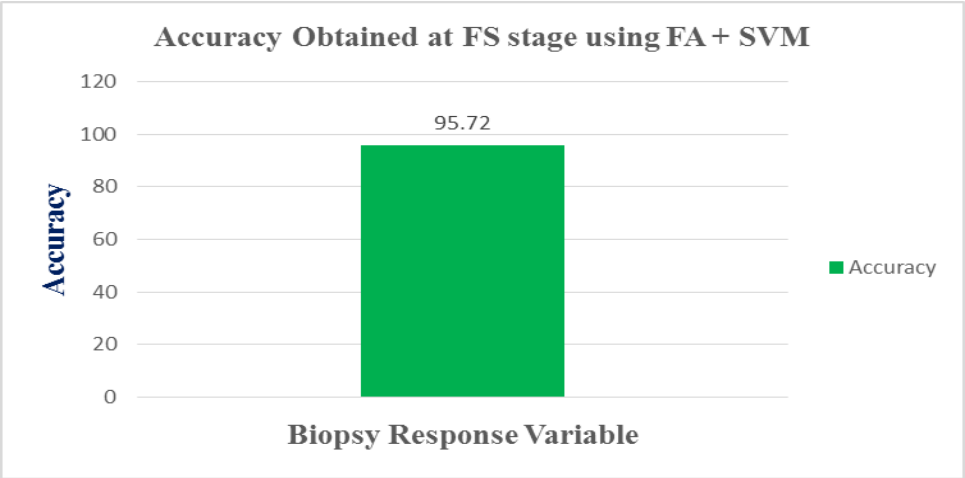


Figure 4.10: Accuracy Obtained at FS stage using FA + SVM

4.4.10.2 Optimization Accuracy for RBF Kernels

During the optimization process of the RBF kernels, optimal accuracy was attain when the value of C=2 and the Y=0.9. Accuracy obtained for the RBF kernels was 96.4847. Figure 4.11 therefore gives a pictorial representation of optimization accuracy for RBF kernels.

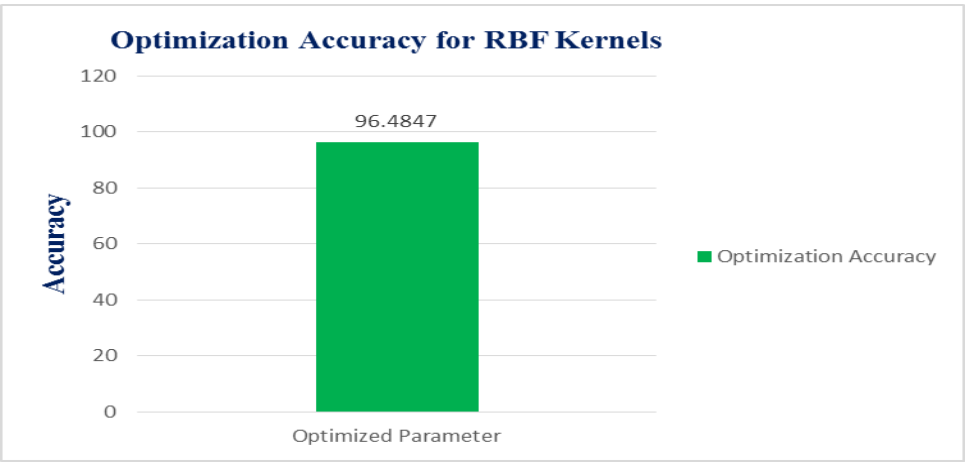


Figure 4.11: Optimization Accuracy for RBF Kernels

4.4.10.3 Result Analysis for Training Time

The training time shows the time taken by the model to create knowledge retention of the data supplied to the RBF-SVM Classifier for the optimized. Figure 4.12 therefore give a pictorial representation for the system training time.

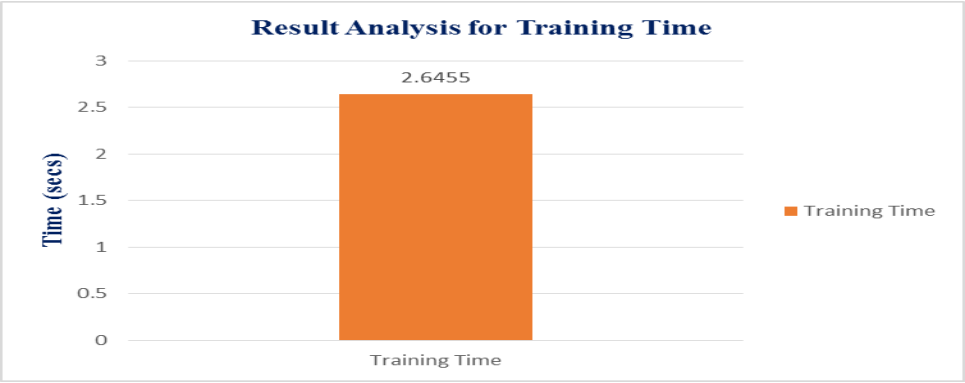


Figure 4.12: Training Time for the Develop Model

4.4.10.4 Result Analysis for Testing Time

The Testing Time (TT) shows the time taken by the model recognize feature supplied to the RBF-SVM Classifier. Figure 4.13 gives a pictorial representation of the testing time for the system.

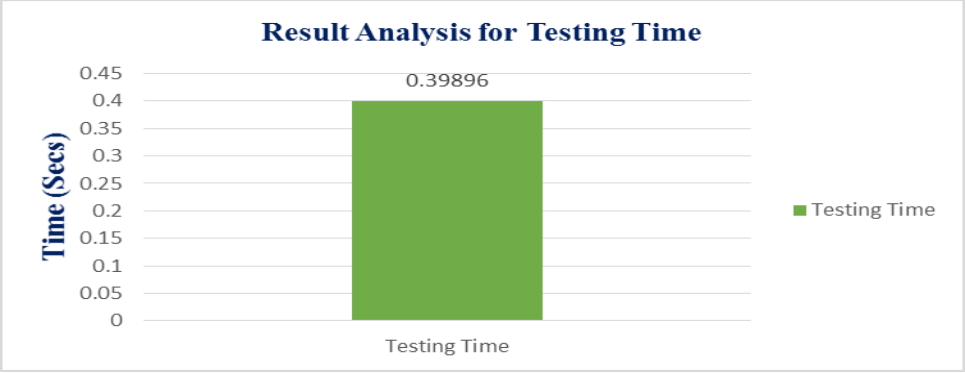


Figure 4.13: Testing Time for the Develop Model

4.4.10.5 Result for Classification Accuracy (CA)

The CA shows the correct classification rate attained by the RBF-SVM Classifier for both cases. The classification accuracy in percentage shows the percentage of instances that were classified correctly. The classification accuracy results obtained for the developed model is 97.20%. See Figure 4.14.

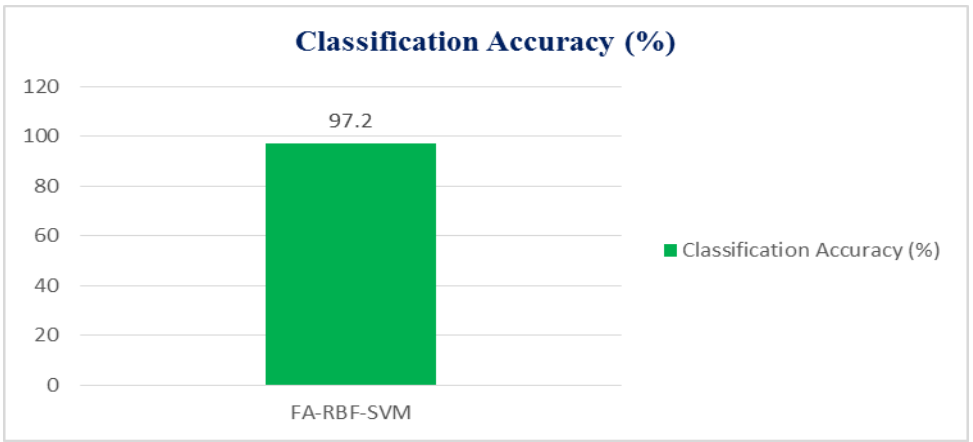


Figure 4:14 Classification Accuracy for the Develop Model

4.4.10.6 Result Analysis for Error rate

The error rate shows the lowest possible error rate for any classifier in a random outcome during the classification. The developed system had an error rate of 2.8037 showing that system performance was satisfactory with very high positive rate for CC detection see Figure 4.15.

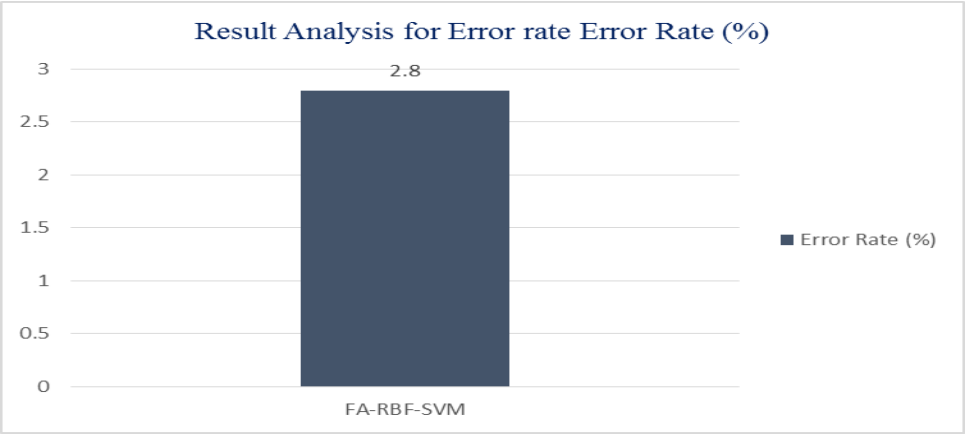


Figure 4.15 Error Rate for the Develop Model

4.4.10.7 Sensitivity and Specificity

The number of positive predictions that are correct divided by the total number of positives is the Sensitivity (SN), Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The best sensitivity and specificity falls at 1 (100%). From the obtained results shows the sensitivity has value close to 1 and the specificity rate has value of 1 indicating a good predictive rate in both case. The developed model had a sensitivity of 97.1% see Figure 4.16, with a Specificity of 100% See Figure 4.17.

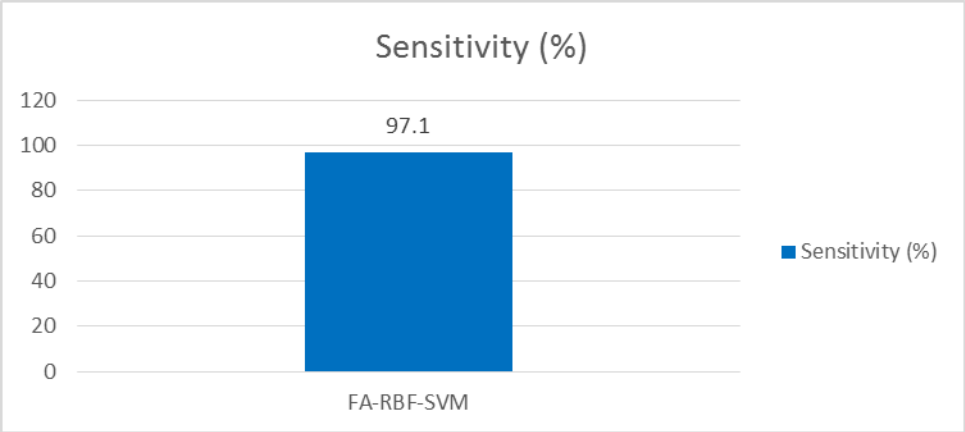


Figure 4.16 Sensitivity for the Develop Model

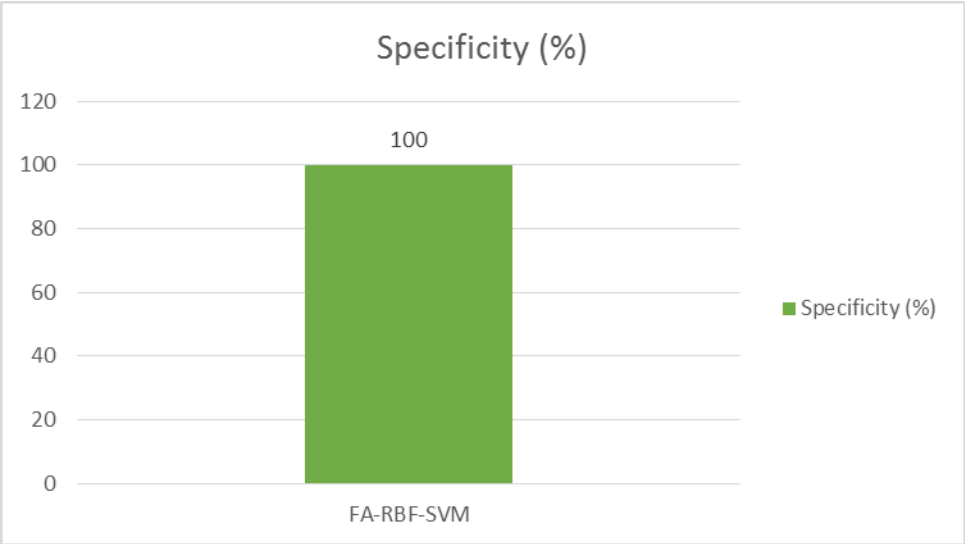


Figure: 4.17 Specificity for the Develop Model

4.4.10.8 Comparative Analysis with Existing State of Arts Based on Classification Accuracy

This section gives a graphical representation of comparative analysis of FA-RBF-SVM with the existing state of arts, from the graphical illustration depicted in Figure 4.18, it showed that FA -RBF-SVM model performed better with a CA of 97.20, showing that

when only the vital features are capitalized on for parameter optimization of SVM kernel, SVM performance was efficient, outperforming previous literature

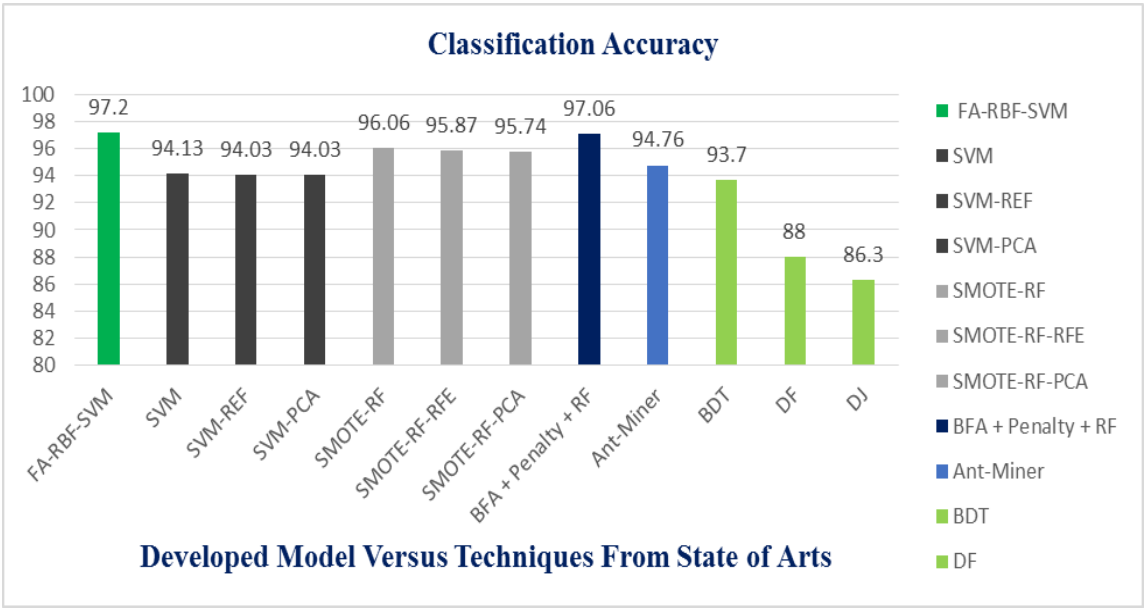


Figure 4.18 Comparative Analysis with Existing State of Art Based on CA

4.4.10.9 Comparative Analysis with Existing State of Arts Based on Sensitivity and Specificity

This section gives a graphical representation of comparative analysis of FA-RBF-SVM with the existing state of art based on the sensitivity and specificity, from the graphical illustration depicted in Figure 4.19, it showed that FA-RBF-SVM model was competitive with SN of 91.10, and outperforming the existing state of art with a SP of 100% see Figure 4.20. Hence when only the vital features are capitalized on for parameter optimization of SVM kernel, SVM performance had a very good predictive rate as compared to some of the previous state of arts.

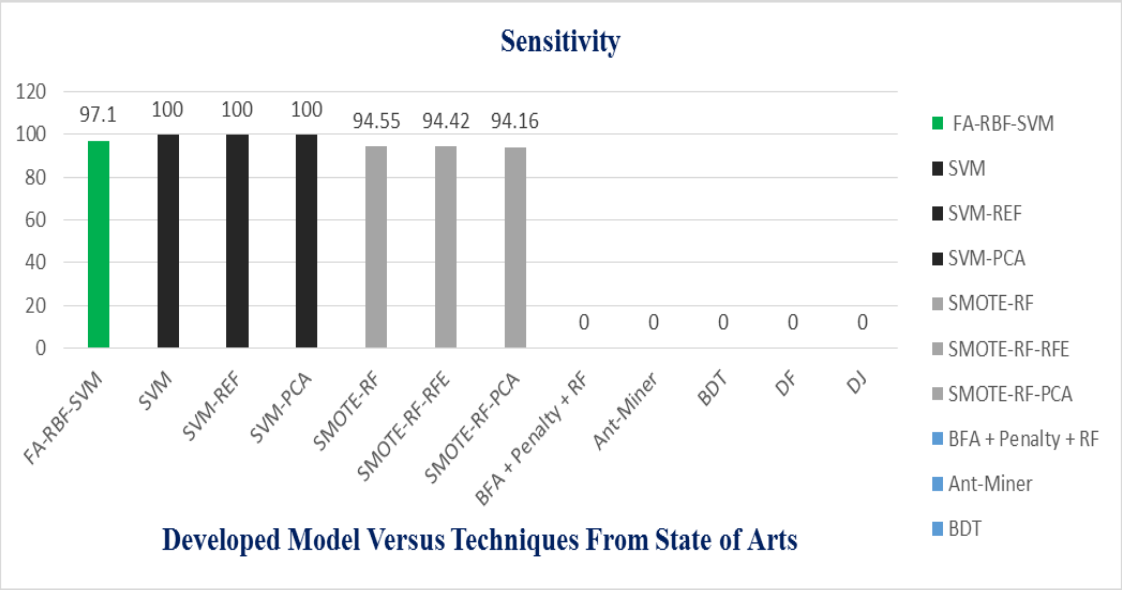


Figure 4.19 Comparative Analysis with Existing State of Art Based on Sensitivity

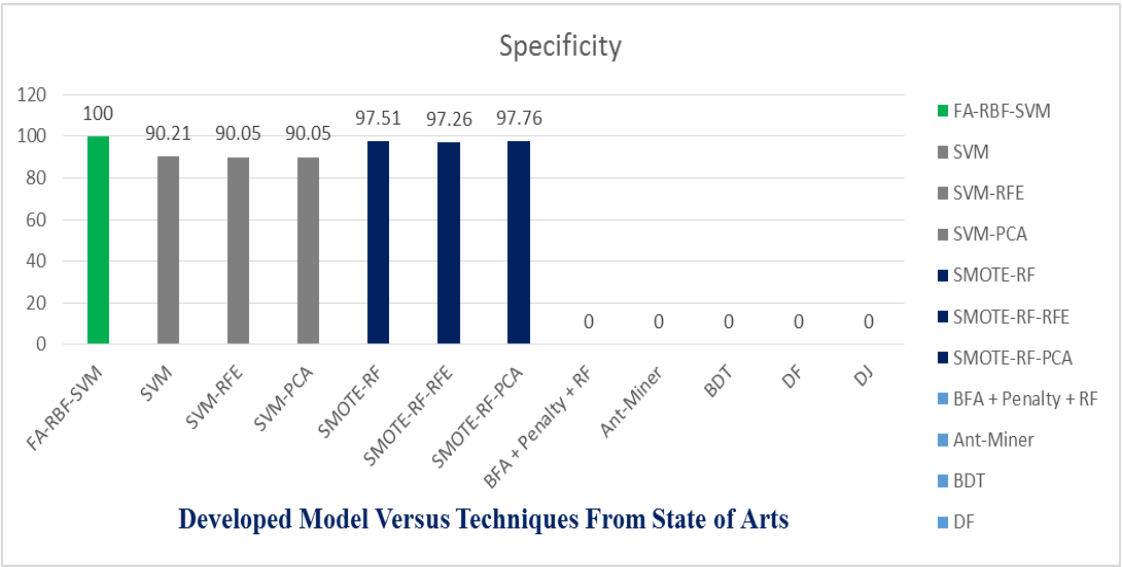


Figure 4.20 Comparative Analysis with Existing State of Art Based on Specificity

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In this Research, a FA-RBF-SVM kernels based approach was used to achieve a high success detection of CC, the model followed a filtering, feature selection, parameter optimization and classification technique of a data mining process. FA was used as wrapper selection technique.15 optimal feature subset (risk factors) necessary for CC detection were selected at an accuracy of 95.7%. selected features attain optimal value of C=2 and Y=0.9 for RBF kernels at an optimal accuracy of 96.5%, When optimal features were employed for training and then for classification, developed model yielded CA of 97.20%, SP of 100% and SN of 97.10%. The developed FA-RBF-SVM model was compared with the existing state of arts in terms of CA, SN and SP matrix. Result from the developed model outperformed existing state of arts for CC detection, and hence FA-RBF-SVM system is strongly recommended for detecting CC patient, because high mortality and morbidity rate of CC patients were reduced by ensuring friendly, timely, effective and efficient CC detection process.

5.2 Major Contributions

Some of the major contribution of this project are

i. Identifying the parameter and values of FA and prominent CC Risk Factors for arriving at the best test to be employed for CC detection process

ii. Determining the value range of RBF kernel Parameter to reduce the computational complexities associated with SVM hence enhanced performance of SVM

5.3 Recommendation

It is recommended that FA be employed for feature selection, and that only the prominent features be obtained from vast features available for prognosis. Pertinent feature obtained will in turn be capitalized on for optimizing parameter of RBF-SVM kernels this is due to an enhanced performance obtained from FA and RBF-SVM kernel for the CC detection process.

5.4 Future Research Directions

The following are promising areas that future researchers can venture into, these areas are highlighted in this section.

i. A comparative analysis of Filter and Wrapper based Approach for CC detection
ii. Ensemble Approach for CC detection using Embedded Approach
iii. SVM kernel Comparison based Approach For CC diagnosis

REFERENCES

Abdi M. J. and Giveki, D. (2013) "Automatic detection of erythemato-squamous diseases using PSO-SVM based on association rules," *Eng. Appl. Artif. Intell.*, vol. 26, no. 1, pp. 603–608, 2013.

Abisoye O.A. and Jimoh R.G. (2015) "A Hybrid Intelligent Forecasting Model to Determine Malaria Transmission" researchgate https://www.researchgate.net/publication/288344295

Ahmed A. Q, Maheswari D. (2017) "Churn prediction on huge telecom data using hybrid firefly basedclassification", Egyptian Informatics J. (2017),http://dx.doi.org/10.1016/j.eij. 2017.02.002

Akinrotimi A. O., and Olugbebi M. A. (2018) "Modelling and Diagnosis of Cervical Cancer Using Adaptive Neuro Fuzzy Inference System" *World Journal of Research and Review (WJRR) ISSN:2455-3956,* Volume-6, Issue-5, May 2018 Pages 01-03

Alam, T. M., Khan, M. M. Afzal, I. M. Atif, W. Abdul M. M. (2019): "Cervical Cancer Prediction through Different Screening Methods using Data Mining", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019

Ali, N. M. Othman, H. Azlishah M. N. and Mohamad H. M. (2014): "A review of firefly algorithm" vol. 9, no. 10, october 2014 issn 1819-6608 ARPN Journal of Engineering and Applied Sciences 2014 Asian Research Publishing Network (ARPN). All rights reserved

Ashok, B. Aruna P. (2016): "Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier", Int. Journal of Engineering Research and Applications, www.ijera.com ISSN: 2248-9622, Vol. 6, Issue 1, (Part - 1) January 2016, pp.94-99

Babatunde R. S., and Muhammad-Thani S. (2018) "Adaptive Neuro Fuzzy Inference System (ANFIS) Based Detection Of Cervical Cancer" International Journal for the Application of Wireless and Mobile Computing (IJFAWMC), Volume 4 October – December, 2018) ISSN: 2141-0720.

Bahadormanesh, N., Rabat, S., Yarali, M., (2016): “Constrained multi objective optimizat ion of radial expanders in organic Rankine cycles by _rey algorithm, Energy Conversion and Management”, 148, 1179{1193 (2017).

Baratloo A., Hosseini, M., Negida, A. and Ashal G. E. (2015): “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity” This open-access article distributed under the terms of the Creative Commons Attribution Non Commercial, 3.0 License (CC BY-NC 3.0). Copyright © 2015 Shahid Beheshti University of Medical Sciences. All rights reserved. Downloaded from: www.jemerg.com Emergency (2015); 3 (2): 48-49.

Barker K., and Berry D., and Rainwater C., eds (2018): “Classification of Cervical Cancer Dataset” Proceedings of the 2018 IISE Annual Conference.

Benazir, B. and Nagarajan A. (2018) “An Expert System for Predicting the Cervical Cancer using Data Mining Techniques**”** International Journal of Pure and Applied Mathematics, Volume 118 No. 20 2018, 1971-1987 ISSN: 1314-3395 (on-line version) url: http://www.ijpam.eu Special Issue

Boutsidis, C., Zouzias, A., Mahoney, M. W., and Drineas, P. (2015): “Randomized dimensionality reduction for-means clustering. Information Theory”, IEEE Transactions on, 61, 1045- 1062.

Charles J. E., and Carraher (2014). “Carraher's polymer chemistry (Ninth ed.). Boca Raton: Taylor & Francis. p. 385. ISBN 9781466552036. Archived from the original on 2015-10- 22.

Cortes, C. and Vapnik, V.N. (1995) "Support-Vector Networks". Machine Learning, springer, NewYork, 20, 1995, 3, 273-297.

Curiac, D. I., Vasile, G., Banias, O., Volosencu C. and Albu, A. (2009) “Bayesian Network Model for Diagnosis of Psychiatric Diseases”, Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, (2009) June 22-25.

Curry S. J, Krist A. H, Owens D. K., Barry M. J, Caughey A. B, Davidson K. W., et al. (August 2018)."Screening for Cervical Cancer: US Preventive Services Task Force Recommendation Statement" .JAMA.**320**(7):674686doi:10.1001/jama. 2018.10897. P M I D 30140884

Devi M., Anousouya R. S., Vaishnavi J. and Punitha S. (2016) “Classification of Cervical Cancer using Artificial Neural Networks” Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016).

Divya T. and Agarwal S. (2013) “A survey on Data Mining approaches for Healthcare” International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241- 266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25ISSN: 2233-7849 IJBSBT Copyright ⓒ 2013 SERSC

Dong, M., Hua, J., and Li, Y. (2007): “A Gaussian Mixture Model to Detect Clusters Embedded in Feature Subspace”. Communications in Information and Systems, 7(4), 337–352.

Eva T., Lazar M. and Milan T. (2016): “Support Vector Machine Parameter Tuning using Firefly Algorithm” 26th Conference Radio elektronika 2016, April 19-20, Koˇsice, Slovak Republic Extraction Algorithm for Cervical Cancer Recognition” Computational and Mathematical Methods in Medicine.

Falcetta F. S, Medeiros L. R, Edelweiss M. I, Pohlmann P. R, Stein A. T and Rosa D. D (November 2016). "Adjuvant platinum-based chemotherapy for early stage cervical cancer". The CochraneDatabaseofSystematicReviews.11:CD005342. :10.1002/14651858.CD005342.pub4. PMC 4164460. PMID 2 7873308

Fatima, M. and Pasha, M. (2017) “Survey of Machine Learning Algorithms for Disease” Diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9, 1-16. https://doi.org/10.4236/jilsa.2017.91001

Fayz S., Rizka, M. A. and Maghraby F., (2018): “Cervical Cancer Diagnosis using Random Forest Classifier with SMOTE and Feature Reduction Techniques”. Volume XX, 2018 10.1109/ACCESS.2018.2874063, IEEE Access. 2169-3536 (c) 2018 IEEE.

FDA (2015) “approves Gardasil 9 for prevention of certain cancers caused by five additional types of HPV". U.S. Food and Drug Administration. 10 December 2014. Archived from the original on 10 January 2015. Retrieved 8 March 2015.

Feng Z., Fu J., Du, D., Li, F. and Sun, S. (2016): “A new approach of anomaly detection in wireless sensor networks using support vector data description” International Journal of Distributed Sensor Networks

Fiste I. Fister I. J., Yang, X., (2014): “A comprehensive review of firefly algorithms” Janez Bresta Swarm and Evolutionary Computatio, 13 (2013) 34–46journal homepage: www.elsevier.com/locate/swevo

Gadducci A., Barsotti C., Cosio S., Domenici L., Riccardo G. A (August 2011). "Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: a review of the literature". Gynecological Endocrinology. **27** (8):597–604. Doi: 10.3109/09513590.2011.558953. PMID 21438669.

Gupta, D. and Gupta M., (2016) “A New Modified Firefly Algorithm” http://dx.doi.org/10.3991/ijes.v4i2.5879

Idowu B, Ogunbodede E., Idowu B. (2016): “Information and Communication Technology in Nigeria” *The Health Sector Experience published in a Journal of Information Technology Impact Obafemi Awolowo University* 3(2), 69-76.

Jensen K., Schmiedel S., Frederiksen K., Norrild B., Iftner T and Kjær S. K. (November 2012). "Risk for cervical intraepithelial neoplasia grade 3 or worse in relation to smoking among women with persistent human papillomavirus infection". Cancer Epidemiology, Biomarkers & Prevention.21(11):1949–55.doi:10.1158/1055-9965.EPI-12-0663. PM C 39 70 16 3. PMID 23019238.

Juliana W. and Hassan F. A. A.: (2018) “Classification of Cervical Cancer Using Ant-Miner for Medical Expertise Knowledge Management”, Knowledge Management International Conference (KMICe) 2018, 25–27 July 2018, Miri Sarawak, Malaysia http://www.kmice.cms.net.my.

Kang, S., Kang, P. Ko T. S., Cho S. R. and Yu K. S. (2015) “An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction,” Expert Syst. Appl., vol. 42, no. 9, pp. 4265–4273, 2015.

Kang, S., Kang, P. Ko T. S., Cho S. R. and Yu K. S. (2015) “An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction,” Expert Syst. Appl., vol. 42, no. 9, pp. 4265–4273, 2015.

Kisi, O. Shiri, J., Sepideh K., Shamshirband. S., Motamedi S., Petkovi, D and Hashim R., (2015) “A survey of water level fluctuation predicting in Urmia Lake using support vector machine with firefly algorithm”journal homepage:www. elsevier.com/locate/amc Vol 4 (7)

Knerr, S., Personnaz, L. and Dreyfus, G. (1990) "Single-layer learning revisited: a stepwise procedure for building and training a neural network". Neurocomputing: Algorithms, Architectures and Applications. Springer-Verlag, F68, pp41-50. 1990

Kourou K., Exarchos T. P., Exarchos K. P. Karamouzis M. V., Fotiadis D. I. (2015) "Machine learning applications in cancer prognosis and prediction" *Research Network of Computationaland Structural* Biotechnology Vol 8-17.https://doi .org/10.1016/j.csbj.2014.11.005

Kumar V., Abbas AK, Fausto N, Mitchell RN (2007). Robbins Basic Pathology (8th ed.). Saunders Elsevier. pp. 718–721. ISBN 978-1-4160-2973-1.

Kurniawati, Y. E., Permanasari, A. E. and Fauziati, S. (2016) "Comparative Study on Data Mining Classification Methods for Cervical Cancer Prediction Using Pap smear Results" 2016 1st International Conference on Biomedical Engineering (IBIOMED), Yogyakarta, Indonesia

Lin Y., Zhou J., Dai L., Cheng Y and Wang J. (September 2017): "Vaginectomy and vaginoplasty for isolated vaginal recurrence 8 years after cervical cancer radical hysterectomy: A case report and literature review". The Journal of Obstetrics andGynaecology Research. 43 (9):1493–1497. doi:10.1111/jog.13375.PMID 2869 1384.

Lindner M., Gramer G., Haege G., Fang-Hoffmann J., Schwab K. O., Tacke U., (2017) "Efficacy and outcome of expanded newborn screening for metabolic diseases" *report of 10 years from South-West Germany. Orphanet J Rare* vol.; 6:44.

Liu K. F. R. and Lu, C. F. (2009) "BBN-Based Decision Support for Health Risk Analysis", Fifth International Joint Conference on INC, IMS and IDC, (2009).

Luhn P., Walker J., Schiffman M., Zuna R. E., Dunn S. T., Gold M. A., Smith K., Mathews C., Allen R. A., Zhang R., Wang S. and Wentzensen N. (2016): "The role of co-factors in the progression from human papillomavirus infection to cervical cancer". GynecologicOncology.128 (2):265270.doi:10.1016/j.ygyno.201 2.11.003.ISSN00908258. PMC 4627848. PMID 23146688.

Mamiya, H. Schwartzman,K. Verma,A. Jauvin,C. Behr, M. and D. Buckeridge, (2015) "Towards probabilistic decision support in public health practice: Predicting recent transmission of tuberculosis from patient attributes," J. Biomed. Inform., vol. 53, pp. 237–242, 2015.

Mashhour, E. E. Houby, M. E. Wassif M. F. Tawfik K. and Salah A. I. (2018): "A Novel Classifier based on Firefly Algorithm" Journal of King Saud University – Computer and Information Sciences journal homepage: www.sciencedirect.com

Mohammad S. Al-Batah, N. Ashidi M. Isa and Mohammed A. A. (2015) **"**Multiple Adaptive Neuro-Fuzzy Inference System with Automatic Features

Mukhopadhyay S., Kurmi, I. D., Rajib D., N. and Pradhan K. S. (2016) "Optical diagnosis of colon and cervical cancer by support vector Machine" Biophotonics: Photonic Solutions for Better Health Care V, edited by Jürgen Popp, Valery V. Tuchin, Dennis L. Matthews, Francesco Saverio Pavone, Proc. of SPIE Vol. 9887, 98870U © 2016 SPIE · CCC code:0277-786X/16/$18·doi: 10.1117/12.2227316

Mutgi, M. A., Murthy, M. R. P., & V, M. T. (2015). "Neural network based automated system for the diagnosis of cervical cancer project" REFERENCE NO.: 38S1497, 3–5.Andries, P.E. (2015). Computational intelligence. John Wiley & Sons, Ltd.

National Cancer Institute (NCI), 2014 Cervical Cancer Treatment ".Archived from the original on 5 July 2014. Retrieved 24 June 2014.

National Cancer Institute (NCI), 2015, "Cervical Cancer Treatment (PDQ)", Archived from the original on 5 July 2014, Retrieved 24 June 2014.

National Cancer Institute SEER Program. (2019) "Cancer Stat Facts: Cervical Cancer".Retrieved 2019-06- 04.

National Institutes of Health, National Cancer Institute (2019): PDQ® "Cervical Cancer Prevention Bethesda, MD: National Cancer Institute. Date last modified 05/01/2010. Accessed 06/04/2019.

Neesha J., N. Abdul R. and Wahidah H., (2015) "Data Mining in Healthcare – A Review" Peer- review under responsibility of organizing committee of Information Systems International Conference (ISICO2015) school of Computer Sciences, Universiti Sains Malaysia, 11800 Minden, Penang Malaysia

Nithya1 B. and Ilango1 V. (2019) "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction" *Springer Nature Switzerland* Vol, 11(7).AG 2019Received: 11 March 2019 / Accepted: 19 May 2019

Nordqvist, C. (2017, August 25). "Cervical Cancer: Causes, Symptoms and Treatments." Medical News Today. Retrieved from http:// www .medicalnewstoday.c om /articles/159821.php.

Ogundele I.O, Popoola O.L, Oyesola O.O, Orija K.T (2018) "A Review on Data Mining in Healthcare" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 7, Issue 9, September 2018, ISSN: 2278 – 1323

Olatomiw L., Mekhilef, S., Shamshirband, S. Mohammadi, K., Petkovic, D.and Sudheer C. h., (2015) "A support vector machine–firefly algorithm-based model for global solar radiation prediction"

Oluyinka A. A. and Ayobami A. A., (2016),"A hybrid firefly and support vector machine classifier for phishing email detection", Kybernetes, Vol. 45 Iss 6 pp.977–994 Permanent link to this document: http://dx.doi.org/10.1108/K-07-2014-0129

Prabukumar M. and Agilandeeswari, L. (2016) "Automatic Classification of Cervical Cancer Cell: A Case Study" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 12, December 2016 ISSN (Print): 2320 – 3765 ISSN (Online): 2278 – 8875

Prakasam A. and Savarimuthu N. (2016) "Metaheuristic algorithms and probabilistic behaviour: a comprehensive analysis of Ant Colony Optimization and its variants". Artific Intell Rev 2016; 45(1):97–130.

Pratap, A. (2019) "Analysis of Big Data Technology and its Challenges", International Research Journal of Engineering and Technology", e-ISSN: 2395-0056, ISSN: 2395-0072, Vol. 6– Issue 03, page 5094-5098, March, 2019.

Pratap, A. Dwivedi A. and Dev Harsh, 2019) "Review of Dimensionality Reduction Techniques in Data Mining from Big Data" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177 Volume 7 Issue V, May 2019- Available at www.ijraset.com ©IJRASET:

Ramit, S., Puneet, M. and Ravi, S., (2018) "A Firefly Algorithm Based Wrapper-Penalty Feature Selection Method for Cancer Diagnosis" Springer

International Publishing AG, partof Springer Nature 2018 O. Gervasi et al. (Eds.): ICCSA 2018, LNCS 10960, pp. 438–449, 2018. https://doi.org/10.1007/978-3-319-95162-1_30

Remschmidt C., Kaufmann A. M., Hagemann I., Vartazarova E., Wichmann O., Deleré Y. (2013). "Risk factors for cervical human papillomavirus infection and high-grade intraepithelial lesion in women aged 20 to 31 years in Germany". InternationalJournal of GynecologicalCancer.23(3):51926.doi:10.1097/IGC.0b013e318285a4b2.PMID23360813.

Resig, J., (2018) "A Framework for Mining Instant Messaging Services" *(PDF). Retrieved 16 March 2018*.

Rimah A., Dorra B., Ayed, N. E. (2015) "Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition" Practical Selection of SVM Supervised Parameters with D

Safaeian M., Solomon D. and Castle P. E., (2017) Cervical Cancer Prevention Cervical Screening: Science in Evolution

Sahoo A., and Chandra S., (2018) "Improved cervix lesion classification using multi-objective binary firefly algorithm-based feature selection**"** Int. J. Bio-Inspired Computation, Vol. 8, No. 6, 2018.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2016): "Pegasos: primal estimatedsub-gradient solver for SVM". Mathematical Programming.127 (1):330. CiteSeerX 10.1.1.161.9629. doi:10.1007/s10107-010-0420-4. ISSN 0025-5610.

Sharma N. and Om H., (2013) "Data mining models for predicting oral cancer survivability," Netw. Model. Anal. Heal. Informatics Bioinforma., vol. 2, no. 4, pp. 285–295, 2013.

Sharma, A., Zaidi, A. Singh, R. Jain, S., and Anita S., (2013) "Optimization of SVM Classifier Using Firefly Algorithm" *Proceedings of the 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*. 978-1-4873-6101

Silver, M. Sakara, T. Su, H. C. Herman, C. Dolins S. B. and O'shea M. J., (2001) "Case study: how to apply data mining techniques in a healthcare data warehouse", Healthc. Inf. Manage, vol. 15, no. 2, (2001), pp. 155-164.

Snijders P. J, Steenbergen R. D., Heideman D. A. and Meijer C. J. (January 2016). "HPV- mediated cervical carcinogenesis: concepts and clinical implications", The Journal of Pathology. 208 (2): 152–64. Doi:10.1002/path.1866. PMID 16362994.

Styawati, and Mustafa K., (2019) "A Support Vector Machine-Firefly Algorithm for Movie Opinion Data Classification" IJCCS (Indonesian Journal of Computing and Cybernetics Systems), Vol.13, No.3, July 2019, pp. 219~230 ISSN (print): 1978-1520, ISSN (online): 2460-7258 DOI: https://doi.org/10.22146/ijccs.41302

Subramaniana, R., Sankaranarayanana, P., Okkuru E. ,Jissa V., Thulaseedharand, R. Swaminathane, S. and Thomasc (2016) "Clinical trial to implementation: Cost and effectiveness considerationsfor scaling up cervical cancer screening in low- and middle-incomecountriesSujha

Sudheer C. h., Sohani S.K., Kumar, D. Malik A., Chahar B.R., Nema A.K. Panigrahi B.K. and Dhiman R.C. (2016) "A Support Vector Machine-Firefly Algorithm based forecasting model to determine malaria transmission"

Sundari, M.G., Rajaram, M. and Balaraman, S., (2016): "Application of improved firefly algorithm for programmed PWM in multilevel inverter with adjustable DC sources", Applied Soft Computing, 41, 169{179 (2016).

Tang, L. Wang, A. Xu, Z. and Li J. (2017) "Online-Purchasing Behavior Forecasting with a Firefly Algorithm-based SVM Model Considering Shopping Cart Use"

Tarney C. M. and Han J. (2014)."Postcoital bleeding: a review on etiology, diagnosis, and management".Obstetrics and GynecologyInternational.2014:192087. doi:10.1 155/2014/19 2087. PMC 4086375. PMID 25045355.

Thendral N. and Lakshmi D. (2019) "Performance Comparison of SVM Classifier Based on Kernel Functions in Colposcopy Image Segmentation for Cervical Cancer" © Springer Nature Switzerland AG 2019

Tran, N. P., Hung, C. F., Roden, R. and Wu, T. C., (2014). "Control of HPV infection and related cancer through vaccination". Recent Results in Cancer Research. 193. pp. 149–71. doi:10.1007/978-3-642- 38965-8_9. ISBN 978-3-642-38964-1. PMID 24008298.

Tuba, E. Mrkela, L. and Tuba M., (2016): "Support Vector Machine Parameter Tuning using Firefly Algorithm" 26th Conference Radioe lektronika 2016, April 19-20, Koˇsice, Slovak Republic

UCI Machine Learning Repository, Cervical cancer (Risk Factors) Data Set. Retrieved February 5, 2019, from https://archi ve.ics.uci. edu/ml/datas ets/Cervi cal+cance r+%28Ris k+Facto rs%29

Vapnik, V.N. (1998) "Statistical Learning Theory". John Wiley and sons Inc., New York, 1998, pp736

Vapnik, Vladimir N. (2014): Invited Speaker. IPMU Information Processing and Management 2014).

WHO (World Health Organization WHO) (February 2014). "Fact sheet No. 297: Cancer". Archived from the original on 2014-02-13. Retrieved 2014-06-24

Wiki1, (2018) https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables

Wiki2 (2019) https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

Wiki4, (2019)https://en.wikipedia.org/wiki/Support-vector_machine#cite_note-HavaSiegelmann-

Wiki 5 (2020) https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the- support-vector-machine-algorithm-explained-with-code-examples/

William, W., Ware, A., Basaza-Ejiri, A., H. and Obungoloch J., (2019): "Cervical Cancer Classification from Pap-smears using an Enhanced Fuzzy CMeans

Wu, W., and Zhou H. (2017) "Data-Driven Diagnosis of Cervical Cancer With SVM-Based Approaches". *special section on data-driven monitoring, fault diagnosis and control of cyberphysical systems*. Vol 5, 2017. Digital Object Identifier 10.11 09/ ACCESS.2017.27 63984 Received September 11, 2017, accepted October 6, 2017, date of publication October 17, 2017, date of current version December 5, 2017.

Xiao, L.Y., Shao, W., Liang, T.L., and Wang, C., (2016) A combined model based on multiple seasonal patterns and modified firefly algorithm for electrical load forecasting, Applied Energy, 167. 135{153 (2016).]

Yang X. (2010). “Nature-Inspired Metaheuristic Algorithms”, 2nd ed. Frome: Luniver Press.

Yang X. and He X. (2018) “Why the Firey Algorithm Works?” Nature-Inspired Algorithms and Applied Optimization (Edited by X.-S. Yang), Springer, pp. 245-259 (2018). https://doi.org/10.1007/978-3-319-67669-2 11

Yang, X. S. (2008). “Nature-Inspired Metaheuristic Algorithms. Luniver ress. ISBN 978-1-905986-10-1.

Yang, X.S., (2014). “Nature-Inspired Optimization Algorithms”, Elsevier Insight, London.

Yong Q., Liu H., Zhang W., Zhu Q. and Zhao Z. (2018). “A classification diagnosis of liver medical data based on Various artificial neural networks”. International Conference on Network, Communication, Computer Engineering (NCCE 2018) *Advances in Intelligent Systems Research*, volume 147.

Zhang X., Dai B., Zhang B., and Wang Z (February 2015). "Vitamin A and risk of cervical cancer: ametaanalysis".GynecologicOncology.124 (2):366 73. doi:10.101 6/j.ygyno.2011.10.012. PMID 22005522.

Zhu W., Zeng, N., and Wang N. (2017) “Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations”

Zolbanin, H. M., Delen, D., and Zadeh A. H., (2015) “Predicting overall survivability in comorbidity of cancers: A data mining approach,” Decis. Support Syst., vol. 74, pp. 150–161, 2015.

APPENDIX

INDEX PAGE

```
% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @major_OpeningFcn, ...
                   'gui_OutputFcn',  @major_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT
  % --- Executes just before major is made visible.
function major_OpeningFcn(hObject, eventdata, handles, varargin)
% This function has no output args, see OutputFcn.
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% varargin   command line arguments to major (see VARARGIN)
 % Choose default command line output for major
handles.output = hObject;
 % Update handles structure
guidata(hObject, handles);

% UIWAIT makes major wait for user response (see UIRESUME)
% uiwait(handles.figure1);


% --- Outputs from this function are returned to the command line.
function varargout = major_OutputFcn(hObject, eventdata, handles)
```

```
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure
varargout{1} = handles.output;

 function edit1_Callback(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Hints: get(hObject,'String') returns contents of edit1 as text
%        str2double(get(hObject,'String')) returns contents of edit1 as a double

% --- Executes during object creation, after setting all properties.
function edit1_CreateFcn(hObject, eventdata, handles)
% hObject    handle to edit1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    empty - handles not created until after all CreateFcns called

% Hint: edit controls usually have a white background on Windows.
%       See ISPC and COMPUTER.
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
   set(hObject,'BackgroundColor','white');
end
```