National College of Ireland

Diagnosis of Cervical Cancer using Hybrid Machine Learning Models

MSc Research Project
Data Analytics

Harsh Dev Singh
x16145747

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Harsh Dev Singh
Student ID:	x16145747
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	13/08/2018
Project Title:	Diagnosis of Cervical Cancer using Hybrid Machine Learning Models
Word Count:	6704

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	13th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Diagnosis of Cervical Cancer using Hybrid Machine Learning Models

Harsh Dev Singh
x16145747
MSc Research Project in Data Analytics

13th August 2018

Abstract

With the advancement of technology & its collaboration with health care, the world has gained a lot of benefits. Advanced data mining and machine learning techniques are continuously improving existing statistical methods in medical field. These improved techniques will help to deliver an intelligent medical health care in the 21st century. This research focuses on the diagnosis of cervical cancer by using the data mining techniques. Cervical cancer is one of the most fatal cancer, the reason being the delay in diagnosis of the disease. This gives rise to a strong need to expedite the process, which is the motivation for this research. To efficiently predict true cervical cancer patients, a better subset of attributes is required. This research uses the Genetic Algorithm (GA) as feature selection algorithm to generate better subset for predictors. The classification algorithms used in this research are "svmLinear", "RandomForest" and "gbm" with the oversampling technique, "SMOTE". Bayesian optimisation is used for hyper parameter tuning, to boost the true positive accuracy for above models. Comparative analysis of all the models has been done on the basis of sensitivity and specificity, where, GBM has delivered more promising results with sensitivity of 0.778 (77.8%) followed by "svmLinear" with sensitivity of 0.5558 (55.58%), and "RandomForest" with sensitivity = 0.44 (44.4%). These sensitivity results will be helpful for the real-time application to make sure that no cancer patient remains untreated.

1 Introduction

Data mining is the "process of identifying hidden patterns in the data", which is now a prominent area of research in the medical domain to identify various health problems. Knowledge extracted by data mining techniques can be used as a support for treating deadly diseases by finding efficient cures and remedies. Predictive models created using data mining algorithms can also be helpful in expediting the diagnosis process of several diseases like breast cancer, cervical cancer, heart attacks, etc.

Cervical cancer is one of the most fatal disease due to which women across the globe are losing their lives. Developing nations are severely suffering from this problem because of the lack of required medical facilities. In India alone, 77,000 women have lost their lives because of the delay in diagnosis of this deadly disease (Jemal et al.; 2011).

Developed countries like England and Sweden are also facing dramatic increase in cervical cancer patients. In England, cervical cancer patients have increased from 2.7% to 4.6% (Castanon and Sasieni; 2018). Similarly, in Sweden, cervical cancer patients have increased by 17% (Dillner et al.; 2018). Increase in number of patients and fatalities due to delay in diagnosis process leads the motivation for this research.

This research focuses on data mining techniques to develop predictive models for classifying cervical cancer patients by predicting their biopsy results based on data related to patients' habits, history and choices like smoking history, age, use of birth control methods, number of sexually transmitted diseases, etc. Biopsy is a process where pathologist takes a sample from cervix tissue to verify whether a patient is suffering from cervical cancer or not. This diagnosis process takes 5-6 days to confirm the presence of cancer, but time is very crucial for a cancer patient, delayed diagnosis may cost the life of a patient.

The objective of this research is to expedite the diagnosis process of biopsy using data mining algorithms. The result of the implemented data mining model can be used as the first result for cervical cancer test. This can also help medical clinics and pathologist labs to make priority appointments for the positive cancer patients & start the treatment immediately.

Not only to predict the result but it is very important in medical sciences to have a correct result of the patient actually suffering from the disease, i.e. have a better true positive value of the results of the cervical cancer test, to ensure that no patient remains untreated. Thus, this research sheds light on the sensitivity more rather than determining the overall accuracy.

Looking onto the above mentioned problem statements, the research question on which this research is based is: ***"To what extent can data mining algorithms deliver reliable results in prediction of the cervical cancer using data of patient's habits and choices?"***

In the previous researches related to cervical cancer & data mining, the researchers have used the data of the pap smear images of infected cells of the cervix. This is again a time consuming process. Thus, the idea behind the model built in this research, the potential patients can get the initial report for their biopsy results immediately.

Data-set used to conduct this research is published on UCI machine learning repository. This is a public data-set which contains limited data of 858 patients (observations) with 36 attributes. With the strict data governance policies, it is always important to maintain the privacy of patient's information thus restricting the release of such data on public repositories for analysis, limiting this research (Bertino et al.; 2005). Despite this, methods are used to overcome this limitation in the following research.

Further this report discusses about the related work done in the past in Section 2. The methodology used in this research is mentioned in Section 3. The implementation is then described in Section 4 with the evaluation mentioned in Section 5. Section 6 concludes the report mentioning the future work.

2 Related Work

2.1 Cervical Cancer & Risk Factors

Among the most deadly diseases globally, cervical cancer is ranked fifth on the basis of mortality rate (Fidler et al.; 2017). The National Comprehensive Cancer Network

(NCCN) has pointed that delay in diagnosis of cervical cancer is the main reason of increased female fatalities despite the availability of advanced medical facilities (Koh et al.; 2015). Therefore, a lot of medical research is done for the cure and the prevention of cervical cancer by understanding its causes, symptoms and remedies in a better way. Many researchers have identified several factors which are responsible for the occurrence of this deadly disease.

Averbach et al. (2018) in their research have discussed that IUD increases the risk of cervical cancer. In another research, done by Rousset-Jablonski et al. (2018) has claimed that IUD increases the risk pelvic inflammatory disease which may leads to cervix infection and if not treated timely it increases the risk of cervical cancer. In the same way, other responsible factors were also identified like how hormonal contraceptive pills can increase the risk of cervical cancer.

In a research done in Australia by Xu et al. (2018), the author has concluded that hormonal contraceptives and smoking are responsible for developing cancer cells in cervix after analysing the data of 886 patients. Long term use of hormonal contraceptive pills leads to the severe hormonal change in women's body which may leads to deadly cancers like breast cancer and cervical cancer (Shukla et al.; 2017).

Smoking is also one of the major factor which increases the risk of cervical cancer because excessive smoking leads to Human Papillomavirus Infection (HPV) infection and thus causes cervical cancer (Chatzistamatiou et al.; 2018; Eldridge et al.; 2017; Wentzensen and Arbyn; 2017). Just like smoking increases the risk of HPV infection, Sexually Transmitted Diseases (STDs) also leads to the same (Parthenis et al.; 2018). As mentioned by Santelli et al. (1998), patient having multiple sexual partners have the highest risk of acquiring sexually transmitted diseases/infections along with the risk of developing cervical cancer.

Along with STDs, the age of patient also influences the risk of cervical cancer. In a research done by Teame et al. (2018), the author has identified that women above 38 years of age and with the STD history are most likely to suffer from cervical cancer. In Table 1 we can have a brief overview of such features given by the authors from medical background to show their importance in this research.

Symptoms of cervical cancer includes irregular vaginal bleeding, bleeding after menopause, discomfort in abdominal body, regular pelvic pain[1]. However, women are not aware of cervical cancer's symptoms and this is also one of the reason due to which diagnosis process gets delayed and in most of the cases women lose their lives (Okunowo et al.; 2018). Hence, there is need to create awareness among women regarding the symptoms of cervical.

Considering all the above attributes, this research uses them to conduct the analysis for the prediction of these test results.

2.2 Data Mining Methodologies

Researchers are continuously striving to develop better data mining models to analyse medical data in order extract hidden knowledge from data, so that extracted information can be used as a support for decision making for the disease diagnosis processes. To improve existing data mining models, researchers are applying and testing latest machine learning algorithms in order to identify the best model. This research also uses the

[1] (https://www.cancer.ie/cancer-information/cervical-cancer/symptoms-and-diagnosissthash.s3yIPYfP.dpbs

Authors & Research	Attributes
Averbach et al. (2018) Rousset-Jablonski et al. (2018)	Use of IUDs
Xu et al. (2018) Shukla et al. (2017)	Hormonal Contraceptive
Shukla et al. (2017) Chatzistamatiou et al. (2018) Eldridge et al. (2017) Wentzensen and Arbyn (2017)	Smoking
Parthenis et al. (2018) Teame et al. (2018)	History of STDs
Teame et al. (2018)	Age

Table 1: Attributes responsible for Cervical Cancer

machine learning algorithms to predict cervical cancer patients efficiently by developing a reliable machine learning model using several data mining techniques. Also, this research is inspired from previous research done in the same field.

This research is also inspired by the work where hybrid machine learning algorithms have been used to analyze medical data. Kalantari et al. (2017) have introduced a new computational intelligence architecture after addressing the limitations of previous architectures. This new computational architecture uses the SVM to perform classification tasks on the medical data-sets. The author has also concluded that the SVM has outperformed the other classification algorithms by achieving the better overall accuracy and sensitivity when integrated with feature selection algorithm - Genetic Algorithm (GA). However, this research lacks the justification for using GA as a feature selection algorithm.

Abdel-Zaher and Eldeib (2016) in their research have demonstrated that hybrid machine learning models can boost the sensitivity and overall accuracy too. Their hybrid model of Deep belief network and Back propagation Neural Net has classified cancer cells with 100% sensitivity and 99.66% overall accuracy. In the same way Sartakhti et al. (2012), tested several hybrid machine learning models and concluded that hybrid model of SVM and simulated annealing has delivered the prediction accuracy of 96% while classifying hepatitis patients. These past work has inspired the following research to implementing several hybrid combinations of machine learning algorithms like classification algorithms integrated with feature selection algorithms and then doing hyperparameter tuning using error optimisation algorithms. This will ultimately result in generating more reliable models.

Feature selection is one of the key steps that can help to improve the performance of predictive models by removing redundant and unimportant attributes without any major loss of information. Feature selection can also be used for exploratory data analysis in order to identify important attributes. While developing predictive model it is important to remove irrelevant attributes because they not only badly effect the performance of predictive model, but they also misguide the algorithms (Pritom et al.; 2016). Guyon and Elisseeff (2003) has demonstrated in his work how good subset of predictors can helps to develop better predictive models. The author has also concluded that feature selection algorithms reduces the complexity which further reduces the computational cost and makes the model efficient and reliable. Hence, in this research feature selection process is implemented to obtain better subset of predictors to enhance the performance of the

cervical cancer classification model.

Machine learning algorithms often needs hyperparameter tuning. This Hyperparameter tuning is usually done to improve performance of the machine learning model. Therefore, an efficient optimization algorithm should be used to tune the parameters of algorithms. In a research done by Elith et al. (2008), hyperparameters of 'GBM' were tuned to boost the performance of the predictive model. Friedman et al. (2010) have also mentioned in their research about the importance of hyperparameter tuning by tuning the parameter of Lasso Regression, Ridge Regression and ElasticNet Regression, in order to decrease the prediction error by adjusting the penalties. Therefore, in this research hyper tuning of classification algorithms has been done by using the optimisation algorithm to boost the sensitivity so that actual patient of cervical cancer can be identified correctly.

Handling class imbalance is very important before developing the predictive models because models trained with imbalanced data set often delivers the biased results towards majority class because of the insufficient training done on the minority class. Hence, it becomes very important to handle class imbalance by implementing data mining techniques like oversampling of minority class or undersampling of majority class. According to Chawla et al. (2002), very often real-world data-set mainly contains normal observation and smaller number of abnormal observations. Therefore, sometimes misclassifying abnormal observation cost more than classifying normal observations correctly. They have also tested that undersampling the majority class has helped to increases the sensitivity, later introducing "Synthetic Minority Over-Sampling Technique" (SMOTE) which is a combination of undersampling and oversampling. Author has tested this technique with classifiers like C4.5, Ripper and Naive Bayes and claimed that Sensitivity has remarkably increased. Sun et al. (2018) have also claimed that combination of SMOTE and decision trees have remarkably improve the prediction of minority class. Hence, in this research SMOTE has been used to handle the class imbalance for predicting cervical cancer patients.

2.3 Background for this Research

Previous research done on the UCI cervical cancer data-set are the primary inspiration for this research. (Wu and Zhou; 2017) has used PCA as the feature reduction technique along with SVM-classifier to predict cytology and biopsy results using UCI cervical cancer data-set. The author has concluded that best results were delivered by PCA-SVM model, but their research lacks the justification of using PCA as feature selection and also no preliminary tests like "Bartlett's Test of Sphericity (BTS)" and "Kaiser-Meyer-Olkin's Test - Measure of Sampling Adequacy (KMO)" were performed before applying PCA. These two tests are very important as they test whether the factor analysis will be helpful or not. BTS is used to test whether the correlation matrix is significantly different from the Identity matrix. In other words, if the significance value is less than 0.05 or 0.01, then the data may be suitable for PCA. Whereas, KMO tests the sample adequacy, if value of KMO is closer to 1 then it means that factor analysis may be useful and if KMO value is less than 0.5 then it means the factor analysis will not be useful. (Pechenizkiy et al.; 2004) have used KMO in their research to verify whether the data is suitable for PCA or not. Similary, (Canbas et al.; 2005) have used BTS in their research to test whether the data is suitable for PCA or not. Following them, this research has used same techniques.

On the same data-set that is used in this research, (Ceylan and Pekel; 2017) have used classification algorithms like Naive Bayes, J48 Decision Tree, Sequential Minimal

Optimization, and Random Forest to classify the cervical cancer patients. For their research, the evaluation matrix used are overall accuracy, hamming loss, ranking loss which seems irrelevant because no comparative analysis has been done on the basis of sensitivity/true positive accuracy. Similar type of drawback was identified in the research done by (Hasan et al.; 2017), where author has implemented ensemble of decision trees and achieved the accuracy of 96% but failed to mention the sensitivity.

Sensitivity is important because this dataset is having heavy class imbalance as 96% of the observations belong to negative cancer cases, whereas only "4%" of data contain positive cancer cases. Therefore, it is more important to classify "4%" actual cancer patients correctly. This research thus, is focused on boosting true positive accuracy so that actual patients can be classified correctly. Also, the evaluation matrix for this research is the sensitivity or true positive accuracy.

After analyzing the results and architectures discussed in previous researches the best suited techniques were chosen for this research so, that a reliable machine learning model can be developed to predict cervical cancer patients. This research uses the Genetic Algorithm (GA) as feature selection techniques. Genetic Algorithm is prominently used to solve search and optimisation problems and GA can be used to search best subset of predictors because performance of classifier is very sensitive to the choice of predictors used to develop classification model (Yang and Honavar; 1998). Also, GA works efficiently on mixed dataset which contains numerical as well as categorical attributes. In a research done by Diker et al. (2018), GA has delivered the best results with SVM by achieving the sensitivity of 87% while predicting the myocardial infarction (Heart attack). Therefore, GA is chosen as feature selection algorithm because the data set used in this research also contains mixed attributes.

The data set used in this research is having a heavy class imbalance, the reason why SMOTE has been used to oversample minority class while training the model. Also, repeated K-fold cross validation has been used because data set is having limited observations.

As discussed above, hyper tuning of classification algorithm is very important because tuning parameters for machine learning algorithms are often hard-coded in data mining tools by developers. Hence, optimising tunable parameters results in significant improvement in the model's performance. To optimize these parameters, Bayesian optimization is used as it is a powerful tool because of its high efficiency in finding global optimisation (Shahriari et al.; 2016).

In a research done by Nishio et al. (2018), to diagnose lung nodule author has identified that Bayesian optimization for tuning parameter of SVM-Radial and GBM delivered best results with sensitivity of 89% and 82% respectively. Zhao et al. (2018) compared performance of grid optimization and Bayesian optimization with SVM and concluded that Bayesian optimsation found the better global minima and remarkably enhanced the performance of svm while prediction of liver cancer. Hence, Bayesian optimization is used to tune the hyper parameters of classification algorithms used in this research to boost the sensitivity.

The models used in the past that have motivated this research are mentioned in the Table 2 below:

Classification algorithm used in this research are "svm-Linear", "RandomForest" and "GBM". Previous researches done on the same dataset have used SVM and random forest therefore, this research focuses to improve their performance. This research have also used GBM additionally because as mentioned above the performance of GBM have delivered

Authors & Year	Techniques Used
Kalantari et al. (2017)	SVM with Gentic Algorithm
Abdel-Zaher and Eldeib (2016)	Deep belief Network, Back-propagation Neural Network
Sartakhti et al. (2012)	SVM with Simulated Annealing
Wu and Zhou (2017)	SVM with PCA
Nishio et al. (2018)	SVMRadial, Bayesian Optimisation, Gradiant Boosting Machines
Zhao et al. (2018)	Grid Optimisation, SVM
Turgut et al. (2018)	SVM, kNN, MLP, Decision Trees, RF, LR, Adaboost, GBM

Table 2: Motivation Drivers for this Research

the best results when integrated with Bayesian optimisation (Nishio et al.; 2018). Also, Turgut et al. (2018) used SVM, KNN, MLP, Decision Trees, Random Forest, Logistic Regression, Adaboost and GBM on the cancer data-set and observed that SVM and GBM have outperformed the other methods. Performance of these classification models is evaluated on the basis of Sensitivity (True positive rate/Recall) to meet the objective of this research i.e to identify a reliable model which can predict actual patients accurately.

3 Methodology

CRISP-DM (Cross Industry Standard Process for Data Mining) framework was followed to conduct this research. This framework provides iterative and agile steps to implement data mining projects efficiently. CRISP-DM breaks the life cycle of data mining project into six steps which provides the better clarity for implementation process. It is a comprehensive approach and has been used in many sectors to execute real-time data mining projects (Nadali et al.; 2011). It gives a hierarchical process flow with some generic tasks to be performed which helps to achieve the end goal of the research (Wirth and Hipp; 2000). It is one of the best models to be used for the industrial projects with a list of "sequential steps" that comprises small objectives of the big target (Azevedo and Santos; 2008).

The phases of CRISP-DM are shown in Figure 1 and discussed below:

1) Business Understanding:

In this phase a data mining problem was identified to diagnose cervical cancer by using the data mining techniques. Motivated from previous researches done in the past, this research is focused to achieve better Recall (Sensitivity) and specificity so that positive cancer patients can get treatment as soon as possible. Hence, in this phase research question was formulated to conduct this research after identification of data mining problem. To conduct this research on the basis of research question next phase includes the data gathering and exploratory data analysis.

2) Data Understanding:

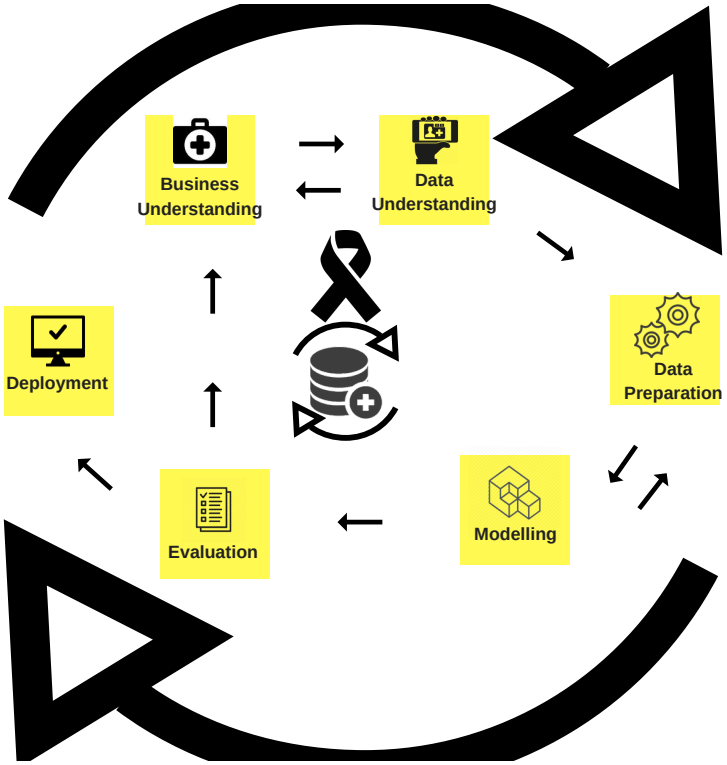

Figure 1: CRISP-DM - Proposed Research Methodology

In this phase data was extracted from UCI Machine learning repository and exploratory data analysis was conducted to understand the nature of data and quality of data. It was important to look for the right data for the analysis specially when it comes to medical data as it is important to deliver right set of information to the world, and an accurate information is always depended on the data on which the analysis is executed. Hence, this UCI data was the best choice to conduct this research as it has been used by researchers in past as well, giving it a validation.

Once data was extracted, correlation matrix was plotted for numerical attributes in order to see if there are any highly correlated variables because highly correlated variables may lead to the redundancy as there are chances that correlated attributes may contain the same information.

The predictors were carefully studied to see if they are relevant to the research or not. Target attribute "Biopsy" was analyzed, and problem of high class imbalance was identified. Boxplots were used to detect the outliers and histograms were used to check the spread of data.

The missing values were identified along with which the incorrect data types for attributes were also identified. Statistical tests like "Bartlett's Test of Sphericity" (BTS) and "Kaiser-Meyer-Olkin's - Measure of Sampling Adequacy" (KMO) were performed to understand if data set is suitable for factor analysis or not. This phase of research is mainly utilized to identify potential problems related to research so that identified issues can be fixed in next phase or take a step back and change the business understanding if required. However, in this research identified issues were fixed in data preparation phase.

The Table 3 below shows the attributes in the dataset

3) Data Preparation:

Attributes	**Data Type**
Age	Int
Number of sexual partners	Int
First sexual intercourse (age)	Int
Num of pregnancies	Int
Smokes	Cat
Smokes (years)	Cat
Smokes (cigarette packs per year)	Cat
Hormonal Contraceptives	Cat
Hormonal Contraceptives (years)	Int
IUD	Cat
IUD (years)	Int
STDs	Cat
STDs (number)	Int
STDs:condylomatosis	Cat
STDs:cervical condylomatosis	Cat
STDs:vaginal condylomatosis	Cat
STDs:vulvo-perineal condylomatosis	Cat
STDs:syphilis	Cat
STDs:pelvic inflammatory disease	Cat
STDs:molluscum contagiosum	Cat
STDs:AIDS	Cat
STDs:HIV	Cat
STDs:Hepatitis B	Cat
STDs:HPV	Cat
STDs: Number of diagnosis	Cat
STDs: Time since first diagnosis	Cat
STDs: Time since last diagnosis	Cat
Dx:Cancer	Cat
Dx:HPV	Cat
Biopsy	Cat

Table 3: Attributes of Dataset in this Research

Data preparation is another important phase of CRISP-DM process. This ensures that the raw data gathered or choosen for the analysis is ready to be used for the models. As we know, the raw data is generally contained with noise, irrelevant information or misclassified datatypes. Hence, it is important to correct them, which is taken care of at this phase.

Most of the issues identified during data understanding were addressed in this phase like categorical and numerical attributes were correctly type-casted as factor and numerical/int respectively. The missing observations were removed along with the removal of the attributes which are not relevant.

Other issues like class imbalance and limited observations were addressed in next phase.

4) Modelling:

Data set received after data preparation was used in this phase for implementation. Few issues were also addressed in this phase like attributes carrying redundant information or irrelevant information were removed using feature selection algorithm (GA) and class imbalance was handled by oversampling minority class using SMOTE.

As the dataset is having limited observations, repeated K- fold cross validation was used to efficiently train classifiers like SVM, RandomForest and GBM to avoid overfitting. The idea of applying repeated K-fold cross validation was influenced from a research done by Yadav and Shukla (2016),where author has suggested to use repeated K-fold cross validation to achieve better results for small datasets. A classification algorithm requires parameter tuning to deliver promising results. Hence, Bayesian optimization is used to tune the hyperparameters of the classifiers.

5) Evaluation:

To evaluate performance of developed machine learning models, the evaluation matrix used are sensitivity, specificity and accuracy. A comparative analysis of performance of machine learning models has been done on the basis of sensitivity in order to identify best model which can classify actual patients (True positive cases) correctly.

6) Deployment:

This the last phase of Data mining project where most reliable data mining model gets deployed in production environment. Most reliable model identified in this research can be deployed on cloud based application or can be used as mobile application where patients can test their biopsy results anytime. If any patients is diagnosed with cancer they can make priority appointments with medical clinics and pathologists labs so that patients can get treated immediately.

4 Implementation

Implementation comprises of the fourth phase according to the CRISP-DM methodology, also known as modelling. This is the part where all the required models are built and executed.

A plan of action was prepared before moving ahead with the further research process. The Figure 2 gives an architecture of the process followed in this research.

To begin with, first the exploratory data analysis was done, to understand the data throughout.

Exploratory Data Analysis:
Exploratory data analysis is very important part of data mining project because it helps the researcher to understand the data set in a better way so that researcher can wisely choose the data mining techniques those are most suitable for the research. In this research, correlation matrix was plotted in order identify highly correlated variables so that information redundancy can be avoided. However, the strongest correlation in the

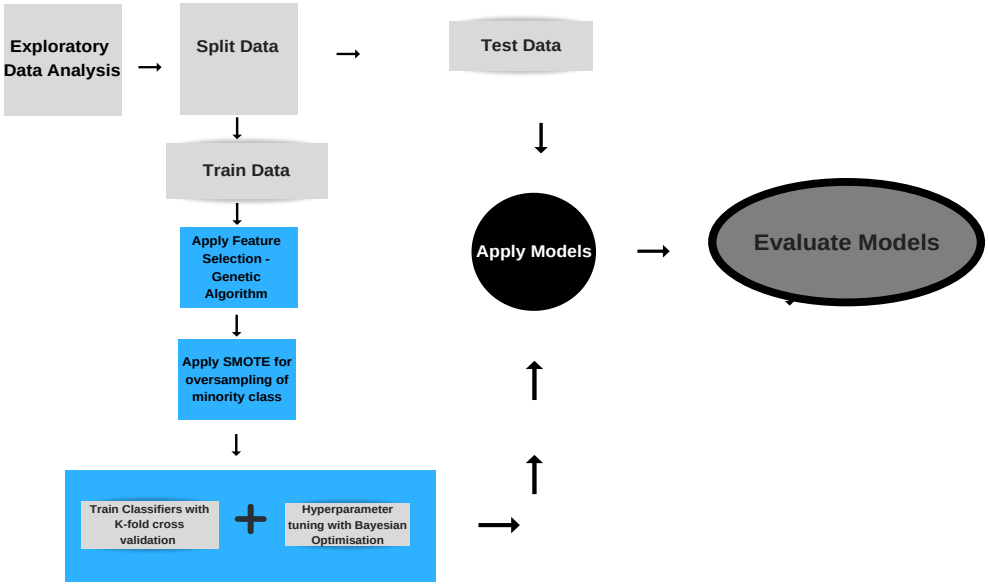

Figure 2: Architecture of the Research

dataset with the value of 0.7 was observed between smoke pack years and smoke year. Here, "smoke pack year" is the number of cigarette packets smoked by the patient and "smokes years" is numerical attribute which contain the information about since how many years the patient is smoking apart from this there is hardly any correlation among other attributes. Below Figure 3 shows the correlation plot created.

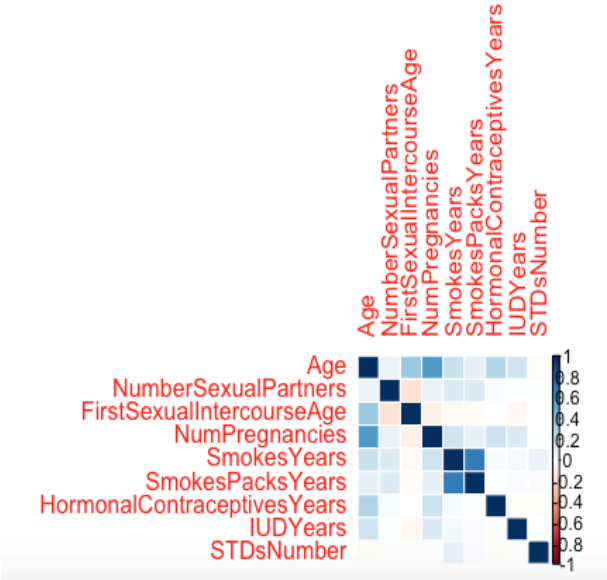

Figure 3: Correlation Matrix

Further, in this research statistical tests like BTS and KMO were performed to see if PCA can be applied of this dataset or not. The objective of this analysis was to understand if PCA will be useful for the analysis on this dataset or not. PCA is a powerful technique of dimensionality reduction and it can be used to plot multidimensional data on 2-D plane, making it easy to explore and visualize so that relationship between attributes

can be studied. However, before applying PCA on a dataset it is important to conduct preliminary tests that can help to understand whether the data set is appropriate for PCA or not.

In this research project, BTS was conducted on the numerical attributes because it tests the hypothesis that correlation matrix is an identity matrix. If the value of p is achieved as less than 0.05, it shows that dataset is suitable for PCA. However, when BTS was conducted on this dataset the p-value obtained is 2.22e-16, which is significantly less than 0.05 and it means that PCA can be applied.

Another important parameter to check for PCA is to make sure that the sampling is adequate on which PCA can deliver useful results. Therefore, KMO test was conducted on the data and achieved the value of 0.48. As the ideal value of KMO test should be any value between 0.6 to 1, thus such low value received in this test tells that PCA will not be useful on the dataset[2]. This is the reason why, PCA was not used.

Splitting Data into Test & Train:

Once dataset was prepared for analysis and exploratory data analysis done, the dataset was then divided into train and test with proportion of 80% and 20% respectively for further steps like feature selection, model building and evaluation.

Feature selection:

As discussed in section 2, feature selection is very important because it provides the subset of best predictors and eliminates the irrelevant predictors. Eliminating irrelevant features increases the performance of predictive models.

In this, research Genetic Algorithm is used for feature selection. GA is powerful technique which follows the stochastic optimisation method and from computational prospective, is very complex. Therefore, this powerful algorithm seems good. This algorithm mimics the procedure of natural evolution of genes in organisms because their genes tends to evolve generation by generation to adapt the changing environment.

GA follows 5 steps to produce better subset of features as shown below:
1) Initialisation: GA follows heuristic optimisation, therefore it randomly creates the population of individuals.
2) Fitness Assignment: Once the population is initialised, fitness assigning takes place for every individual after evaluation of selection error (error received after training the predictive model). Then the individuals with greater fitness will be selected for recombining.
3) Selection: After fitness assignment, the individual with the best fitness value are selected for recombining.
4) Crossover: After the best individuals are selected by selection step, crossover recombines the selected individuals.
5) Mutate: Crossover may generate offspring similar to parents, whereas mutation changes the value of few features in offspring at random to create new generation.
6) Iterate until the stopping criteria met.

Following above process GA shortlisted below mentioned attributes listed in Table 5 as the best subset of predictors. Also, GA has used the repeated cross validation with 5 repeats and 30 iterations while doing feature selection.

[2]https://www.ibm.com/support/knowledgecenter/en/SSLVMB$_2$4.0.0/$spss/tutorials/fac_telco_kmo_01.html$

S.No.	Attributes
1.	Age
2.	NumberSexualPartners
3.	FirstSexualIntercourseAge
4.	NumPregnancies
5.	SmokesPacksYears
6.	HormonalContraceptivesYears
7.	STDsNumber

Table 4: Attributes selected by Genetic Algrithm

Using above mentioned shortlisted features a new dataset was created to develop the machine learning models to predict biopsy results so that performance of machine learning models can be improved.

SMOTE to oversample minority class for training

As discussed in section 3, a heavy class imbalance was identified during exploratory data analysis and in section 2, an oversampling technique "SMOTE" has been discussed which can be very useful here.

The dataset used in this research is having very high class imbalance because 96% of the observations are labelled with negative cancer cases and positive cancer observations are only 4%. Hence, to improve classification rate of positive cancer class, it is important to oversample the positive class during the training of predictive model because according to previous researches this oversampling techniques boosts the classification rate of the minority class.

SMOTE uses the K-nearest neighbours to oversample the minority class. Hence, in this research parameter "k" was optimised by iteratively trying different values of k from 2 to 23, and then using oversampled data to train the SVM based predictive model in order to identify the best value of "k". It was observed that best value of k was 5 followed by k= 7 because when k=5 was used to oversample minority class the training error was least. Hence, k=5 was used to oversample minority class using SMOTE. After applying SMOTE, the class imbalance was fixed and new training data contains the 50% observations for negative cancer cases and 50% for positive cancer cases.

Classifiers trained with repeated K-fold cross validation and hyperparameter tuning using Bayesian optimisation:

While training classifiers, 10-fold repeated cross validation was used with 5 repeats because observations are limited and may leads to overfitting. Also, hyperparameter tuning for classification models was done using Bayesian optimisation in order to get better results.

Model 1: SVMLinear

The classifier "svmLinear" was trained for the further analysis. In svmLinear, tuning parameter is "C" and tuning of C is very important as it is a regularisation parameter responsible to control the tradeoff to achieve low error for training and testing to generalise the classification model for unseen data. In this research, the best value of C, identified by Bayesian optimization is 96.27. Therefore, svmLinear classifier used in this research was tuned with C = 96.27 to predict the biopsy outcomes for cervical cancer

patients. This classifier was tested with test data to check if it's a reliable model or not by evaulting it on the basis of sensitivity and specificity.

Model 2: RandomForest

RandomForest is used for classification in this research to predict the Biopsy outcome. Tuning paramters for randomforest are "mtry", "node size" and "num of trees". In this research, Bayesian optimization is used to tune the above-mentioned parameters in order to boost the the performance of randomforest classifier while predicting the biopsy results.

Below are the optimal values identified:
mtry = 2.896728 node size = 32 number of trees = 326.

Using these tuned parameters, a randomforest based classifier was developed to predict the biopsy results for the test data and further gets evaluated on the basis of sensitivity and specificity.

Model 3: GBM

GBM can be used to implement regression or classification model. In this research, GBM is used to predict biopsy results of cervical cancer patients. GBM is a powerful classification technique, with the tuning parameters as "interaction.depth", "shrinkage", "n.minobsinnode". On tuning these parameters, the performance of the GBM classifier can be improved.

The best values for these parameters were identified as interaction.depth = 2.962253, n.trees = 67.87829, shrinkage = 0.001, n.minobsinnode = 18.06413.

The tuned GBM model was tested with the test data to analyse its performance and reliability.

5 Evaluation

5.1 Model 1: svmLinear

The tuned svmLinear classifier was used to predict the biopsy results for test data. After evaluating the results it was observed that the overall accuracy achieved is 65.47%, with the sensitivity of 55.56% and specificity of 66.15%.

The important variables according to svmlinear can be seen in Figure 4

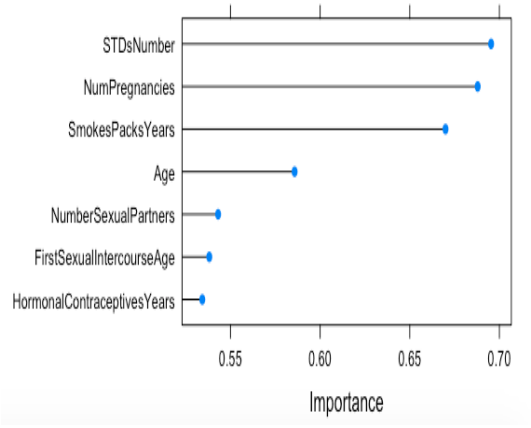

Figure 4: Important Parameters by svmLinear

5.2 Model 2: RandomForest

When tuned randomforest classifier was tested with test data to predict biopsy outcome. The overall accuracy achieved was 70% , sensitivity was 44.4% and specificity was 71.53%.

The important variables according to RandomForest can be seen in Figure 5

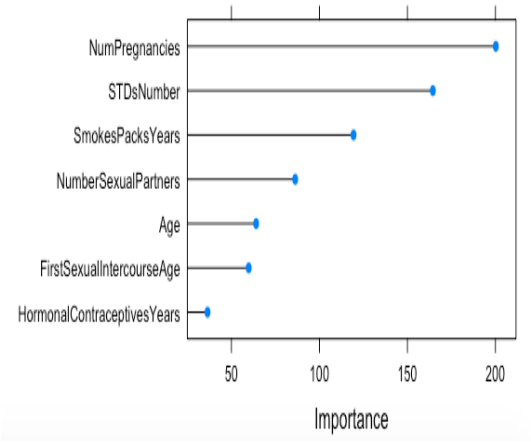

Figure 5: Important Parameters by RandomForest

5.3 Model 3: GBM

When tuned GBM, the classification model was validated against test data. The overall accuracy received was 40.31% , sensitivity was 77.8% and specificity was 41.8%. The important parameters according to GBM can be seen in Figure 6

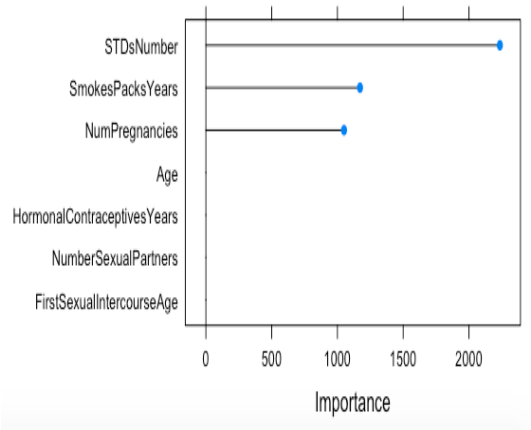

Figure 6: Important Parameters by GBM

5.4 Discussion

As shown in below barchart, the classification model based on Randomforest has outperformed SVM and GBM in terms of overall accuracy. But the results delivered for true

positive cases doesn't seems promising. Whereas, SVM and GBM have perfomed significantly better than RandomForest specially while classifying the True cancer patients. Among SVM and GBM , though GBM has delivered significantly less overall accuracy as compared to svm but sensitivity delivered by gbm is significantly high and therefore makes it more reliable model to classify actual cancer patients.

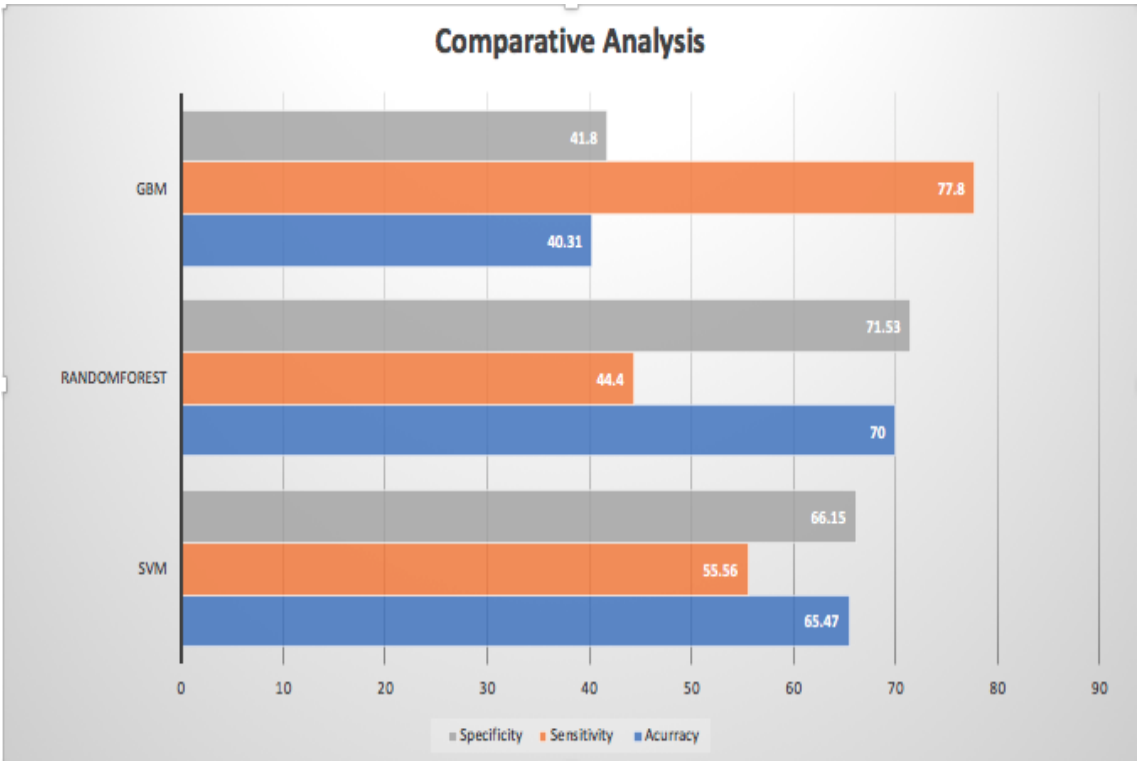

Figure 7: Important Parameters by Comparative Ananlysis

Also, the most important top three factors recognised by all the models are same but according to gbm and svm "number of STDs" by which patient is suffering is most important factor. Whereas, for SVM second most important attribute is "Number of pragnancies" patient had and for GBM second most important attribute is "Number of cigarette packs" smoked by patient in a year. The third most important attribute for SVM is "Number of cigarette packs" and for GBM "number of pragnancies" is the third most important factor. On the other hand, for the random forest, most important attribute is "Number of pragnacies" followed by "number of STDs" and "Number of cigarette packs". On a broader view top three important factors recognised by classification models are "Number of pragnancies", "Number of cigarette packs" and "number of STDs".

	SVM	**RandomForest**	**GBM**
1.	STDNumber	NumPraganancies	STDNumber
2.	NumPraganancies	STDnumber	SmokePacksYears
3.	SmokePacksYears	SmokePacksYears	NumPraganancies

Table 5: Important Factors given by each model

6 Conclusion and Future Work

Cervical Cancer is a deadly disease and also responsible for high mortality rate among women. The reason of high mortality was identified as delay in diagnosis by many researchers. This problem was identified in this research and a research question was formulated around it.

The question asked in this research has been answered by implementing machine learning models to predict the biopsy results of cervical cancer. It has been concluded that hybrid model using GBM and Bayesian optimisation has delivered the reliable predictive model to classify cervical cancer patients by using patient's data, though the overall accuracy is less as compared to SVM and RandomForest.

The important attributes identified are "Number of STDs", "Number of cigarette packs smoked by patient" and "number of pregnancies". These identified attributes can be used for future work to predict the stage of the cervical cancer by including the data of the symptoms by which patient is suffering like "increased vaginal discharge", "longer menstrual bleeding", "persistent pelvic pain", and "Time since symptoms appeared". Including above attributes in research may enhance the performance of predictive model too. The limitation of this research is that it has been conducted on small dataset of 858 observations. Hence, future work is recommended on large dataset so that in depth analysis can be done and better predictive model can be identified for the same problem.

Acknowledgment

I would like to take this opportunity to thank my supervisor Noel Cosgrave for supervising me in my research work with his experience and knowledge. I would appreciate the public data facility available by UCI Machine Learning Repository for providing a data for cancer patients online.

References

Abdel-Zaher, A. M. and Eldeib, A. M. (2016). Breast cancer classification using deep belief networks, *Expert Systems with Applications* **46**: 139–144.

Averbach, S., Silverberg, M. J., Leyden, W., Smith-McCune, K., Raine-Bennett, T. and Sawaya, G. F. (2018). Recent intrauterine device use and the risk of precancerous cervical lesions and cervical cancer, *Contraception* .

Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview, *IADS-DM* .

Bertino, E., Ooi, B. C., Yang, Y. and Deng, R. H. (2005). Privacy and ownership preserving of outsourced medical data, *null*, IEEE, pp. 521–532.

Canbas, S., Cabuk, A. and Kilic, S. B. (2005). Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The turkish case, *European Journal of Operational Research* **166**(2): 528–546.

Castanon, A. and Sasieni, P. (2018). Is the recent increase in cervical cancer in women aged 20–24 years in england a cause for concern?, *Preventive medicine* **107**: 21–28.

Ceylan, Z. and Pekel, E. (2017). Comparison of multi-label classification methods for prediagnosis of cervical cancer, *International Journal of Intelligent Systems and Applications in Engineering* **5**(4): 232–236.

Chatzistamatiou, K., Moysiadis, T., Vryzas, D., Chatzaki, E., Kaufmann, A. M., Koch, I., Soutschek, E., Boecher, O., Tsertanidou, A., Maglaveras, N. et al. (2018). Cigarette smoking promotes infection of cervical cells by high-risk human papillomaviruses, but not subsequent e7 oncoprotein expression, *International journal of molecular sciences* **19**(2): 422.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Diker, A., Cömert, Z., Avci, E. and Velappan, S. (2018). Intelligent system based on genetic algorithm and support vector machine for detection of myocardial infarction from ecg signals, *2018 26th Signal Processing and Communications Applications Conference (SIU)*, IEEE.

Dillner, J., Sparén, P., Andrae, B. and Strander, B. (2018). Cervical cancer has increased in sweden in women who had a normal cell sample, *Lakartidningen* **115**.

Eldridge, R. C., Pawlita, M., Wilson, L., Castle, P. E., Waterboer, T., Gravitt, P. E., Schiffman, M. and Wentzensen, N. (2017). Smoking and subsequent human papillomavirus infection: A mediation analysis, *Annals of epidemiology* **27**(11): 724–730.

Elith, J., Leathwick, J. R. and Hastie, T. (2008). A working guide to boosted regression trees, *Journal of Animal Ecology* **77**(4): 802–813.

Fidler, M. M., Gupta, S., Soerjomataram, I., Ferlay, J., Steliarova-Foucher, E. and Bray, F. (2017). Cancer incidence and mortality among young adults aged 20–39 years worldwide in 2012: a population-based study, *The Lancet Oncology* **18**(12): 1579–1589.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of statistical software* **33**(1): 1.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of machine learning research* **3**(Mar): 1157–1182.

Hasan, M. R., Gholamhosseini, H. and Sarkar, N. I. (2017). A new ensemble classifier for multivariate medical data, *Telecommunication Networks and Applications Conference (ITNAC), 2017 27th International*, IEEE, pp. 1–6.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E. and Forman, D. (2011). Global cancer statistics, *CA: a cancer journal for clinicians* **61**(2): 69–90.

Kalantari, A., Kamsin, A., Shamshirband, S., Gani, A., Alinejad-Rokny, H. and Chronopoulos, A. T. (2017). Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions, *Neurocomputing* .

Koh, W.-J., Greer, B. E., Abu-Rustum, N. R., Apte, S. M., Campos, S. M., Cho, K. R., Chu, C., Cohn, D., Crispens, M. A., Dorigo, O. et al. (2015). Cervical cancer, version 2.2015, *Journal of the National Comprehensive Cancer Network* **13**(4): 395–404.

Nadali, A., Kakhky, E. N. and Nosratabadi, H. E. (2011). Evaluating the success level of data mining projects based on crisp-dm methodology by a fuzzy expert system, *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, Vol. 6, IEEE, pp. 161–165.

Nishio, M., Nishizawa, M., Sugiyama, O., Kojima, R., Yakami, M., Kuroda, T. and Togashi, K. (2018). Computer-aided diagnosis of lung nodule using gradient tree boosting and bayesian optimization, *PloS one* **13**(4): e0195875.

Okunowo, A. A., Daramola, E. S., Soibi-Harry, A. P., Ezenwankwo, F. C., Kuku, J. O., Okunade, K. S. and Anorlu, R. I. (2018). Women's knowledge of cervical cancer and uptake of pap smear testing and the factors influencing it in a nigerian tertiary hospital, *Journal of Cancer Research and Practice* .

Parthenis, C., Panagopoulos, P., Margari, N., Kottaridi, C., Spathis, A., Pouliakis, A., Konstantoudakis, S., Chrelias, G., Chrelias, C., Papantoniou, N. et al. (2018). The association between sexually transmitted infections, human papillomavirus and cervical cytology abnormalities among women in greece, *International Journal of Infectious Diseases* .

Pechenizkiy, M., Tsymbal, A. and Puuronen, S. (2004). Pca-based feature transformation for classification: Issues in medical diagnostics, *Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings. 17th IEEE Symposium on*, IEEE, pp. 535–540.

Pritom, A. I., Munshi, M. A. R., Sabab, S. A. and Shihab, S. (2016). Predicting breast cancer recurrence using effective classification and feature selection technique, *Computer and Information Technology (ICCIT), 2016 19th International Conference on*, IEEE, pp. 310–314.

Rousset-Jablonski, C., Reynaud, Q., Nove-Josserand, R., Durupt, S. and Durieu, I. (2018). Gynecological management and follow-up in women with cystic fibrosis, *Revue des maladies respiratoires* .

Santelli, J. S., Brener, N. D., Lowry, R., Bhatt, A. and Zabin, L. S. (1998). Multiple sexual partners among us adolescents and young adults, *Family planning perspectives* pp. 271–275.

Sartakhti, J. S., Zangooei, M. H. and Mozafari, K. (2012). Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (svm-sa), *Computer methods and programs in biomedicine* **108**(2): 570–579.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. and De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization, *Proceedings of the IEEE* **104**(1): 148–175.

Shukla, A., Jamwal, R. and Bala, K. (2017). Adverse effect of combined oral contraceptive pills, *Asian J Pharm Clin Res* **10**(1): 17–21.

Sun, J., Lang, J., Fujita, H. and Li, H. (2018). Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates, *Information Sciences* **425**: 76–91.

Teame, H., Addissie, A., Ayele, W., Hirpa, S., Gebremariam, A., Gebreheat, G. and Jemal, A. (2018). Factors associated with cervical precancerous lesions among women screened for cervical cancer in addis ababa, ethiopia: A case control study, *PloS one* **13**(1): e0191506.

Turgut, S., Dağtekin, M. and Ensari, T. (2018). Microarray breast cancer data classification using machine learning methods, *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, IEEE, pp. 1–3.

Wentzensen, N. and Arbyn, M. (2017). Hpv-based cervical cancer screening-facts, fiction, and misperceptions, *Preventive medicine* **98**: 33–35.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, pp. 29–39.

Wu, W. and Zhou, H. (2017). Data-driven diagnosis of cervical cancer with support vector machine-based approaches, *IEEE Access* **5**: 25189–25195.

Xu, H., Egger, S., Velentzis, L. S., O'Connell, D. L., Banks, E., Darlington-Brown, J., Canfell, K. and Sitas, F. (2018). Hormonal contraceptive use and smoking as risk factors for high-grade cervical intraepithelial neoplasia in unvaccinated women aged 30–44 years: A case-control study in new south wales, australia, *Cancer epidemiology* **55**: 162–169.

Yadav, S. and Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, IEEE, pp. 78–83.

Yang, J. and Honavar, V. (1998). Feature subset selection using a genetic algorithm, *Feature extraction, construction and selection*, Springer, pp. 117–136.

Zhao, Y., Liu, Y. and Huang, W. (2018). Prediction model of hbv reactivation in primary liver cancer—based on nca feature selection and svm classifier with bayesian and grid optimization, *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, IEEE, pp. 547–551.