A Comparative Analysis of Classification Techniques for Cervical Cancer Utilising At Risk Factors and Screening Test Results

Sean Quinlan, Haithem Afli and Ruairi O'Reilly

Cork Institute of Technology, Ireland
sean.a.quinlan@mycit.ie, haithem.afli@cit.ie and ruairi.oreilly@cit.ie

Abstract. Cervical cancer is a severe concern for women's health. Every year in the Republic of Ireland, approximately 300 women are diagnosed with cervical cancer, 30% for whom the diagnosis will prove fatal. It is the second most common cause of death due to cancer in women aged 25 to 39 years [14]. Recently there has been a series of controversies concerning the mishandling of results from cervical screening tests, delays in processing said tests and the recalling of individuals to retake tests [12]. The serious nature of the prognosis highlights the importance and need for the timely processing and analysis of data related to screenings.

This work presents a comparative analysis of several classification techniques used for the automated analysis of known risk factors and screening tests with the aim of predicting cervical cancer outcomes via a Biopsy result. These techniques encompass methods such as tree-based, cluster-based, liner and ensemble techniques, and where applicable use parameter tuning to determine optimal model parameters.

The dataset utilised for training and validation consists of 858 observations and 36 variables, including the binary target variable "Biopsy". The data itself is heavily imbalanced with 803 negative and 55 positive observations with approximately 11.73% of the data points missing. These issues are addressed during pre-processing by methods such as mean or median imputation, as well as over-sampling, under-sampling and combination techniques which led to the creation of 6 augmented datasets of varying size, consisting of 34 variables including the response Biopsy.

The results show that a SMOTE-Tomek combination resampling method in conjunction with a tuned Random Forest model produced an accuracy score of 99.69% with a recall and precision value of 0.99% for both positive and negative responses.

Keywords— Machine Learning, Classification Techniques, Cervical Cancer

1 Introduction

Cervical cancer is a disease in which healthy cells on the surface of the cervix grow out of control forming a mass of cells called a tumour, which can then spread

to other regions of the body. After breast cancer, it is the second most common cancer among women worldwide [11], and is also one of the most preventable cancers with 90% of cases identifiable and treatable in its early stages [28].

According to the World Health Organisation, comprehensive cervical cancer control includes primary prevention (vaccination against HPV), secondary prevention (screening and treatment of pre-cancerous lesions), tertiary prevention (diagnosis and treatment of invasive cervical cancer) and palliative care [30]. It is at the secondary screening phase that this analysis is to be employed.

Diagnosing cervical cancer requires several physical tests, such as a HPV test, smear test, or colposcopy. This process can take a minimum of 4 weeks for results to return, and during the high demand period results took up to 33 weeks to be returned [13].

The use of classification techniques can provide an informed initial indication of at-risk individuals enabling their tests to be expedited and medical intervention employed at an earlier stage. This is especially useful during periods of high-volume testing such as those seen in Ireland in recent times [12], as delays in diagnosis of cervical cancer are one of the main reasons for increased fatalities despite the availability of advanced medical facilities [17]. Similarly, this method has the potential to be of value in low-resource settings as only an individual's risk factor information is needed to perform an initial screening.

2 Related Work

A woman's risk of developing cervical cancer is affected by several factors, some of which are intrinsic such as genetics and age, others such as smoking habits, methods of contraceptives, and diet are modifiable. An implication of which is that individuals can take actions to reduce the impact of known risk factors. This work aims to analyse these known risk factors, the majority of which are modifiable to determine the outcome of a patient's classification regarding cervical cancer based on biopsy results. The following studies have shown that these risk factors are significant in the development of cervical cancer.

Manderson et al. [19] showed that bearing several children has been found to contribute to increased risk of cervical cancer. In an Australian study, Xu et al. [32] found that hormonal contraceptives and smoking contribute to the development of cervical cancer, while a study by Shukla et al. [26] showed long term use of contraceptive pills might lead to breast and cervical cancer. Averbach et al. [2] highlighted the contribution of IUD contraceptives in the development of cervical cancer, a similar study by Rousset-Jablonski et al. [23] focused on IUD regarding the pelvic inflammatory disease which can further contribute to cervical cancer. Age being an intrinsic feature has been shown by Teame et al. [27] to contribute to the risk of a patient's development of cervical cancer. Eldridge et al. [6] concluded that smoking leads to cervical cancer by increasing the risk of Human Papillomavirus Infection (HPV). Sexually transmitted diseases (STDs) have been shown to also lead to an increased risk of HPV and cervical cancer by Parthensis et al. [21], while a somewhat common sense finding by Santelli et al.

[24] in that patients having multiple sexual partners increase the risk of STDs which in turn leads to a greater risk of developing cervical cancer. Per the Irish Cancer Society 2017 Review [15] HPV has been shown to be a large contributor to the development of cervical cancer, while also highlighting a steep decline (87% down to 50%) over a two-year period prior to the review in the numbers receiving the vaccination due to social media misinformation – this stresses the importance of clear, informed, and available information.

Bosch et al. [4] used linear logistic regression to study the relationship between cervical cancer, HPV, aspects of sexual and reproductive behaviour, oral contraceptives and smoking habits of patients. Finding that HPV was the biggest risk factor in determining occurrences of cervical cancer. The National Cancer Registry Ireland (NCRI) also cites these factors as being leading contributors to the development of cervical cancer [20]. [4] also notes a significant increase in risk for those in low education areas. This increase is also noted by the WHO [30] regarding higher rates of cervical cancer in developing countries.

The advent of big data has seen increased interest in automated solutions for analytical processes. In the context of healthcare, this has resulted in a transition in clinical practice whereby practitioners are encouraged to incorporate technology-based solutions if increased efficiencies, transparency or cost reductions can be achieved by doing so. This transition is materialising itself in the form of advanced artificial intelligence and machine learning-based techniques in areas such as automated decision making, treatment plans and supervision of patients.

3 Methodology

This research utilises classification techniques and patient data consisting of known risk factors such as age, the number of pregnancies, STD's, and smoking habits with the intent of developing predictive models to accurately classify a patient's diagnosis of cervical cancer based on biopsy results. The analysis seeks to assess the dataset via several supervised classification models encompassing areas such as tree, cluster, linear and ensemble technique, and where applicable apply parameter tuning to determine the optimal prediction parameters for each model. Each model is then compared to determine an overall optimal method for predicting the diagnosis of cervical cancer based on the Biopsy classification.

The dataset used in this analysis is the "Cervical Cancer Risk Factors" dataset available from the UCI data repository [16]. This dataset originated from "Hospital Universitario de Caracas' in Caracas, Venezuela and is derived from historical medical records of 858 patients with a Biopsy count of 803 Negative to 55 Positive observation [9]. Similar work has previously been carried out on this dataset, the findings of two such papers are as follows. Alwesabi et al. [1] have previously analysed this dataset regarding classification and feature selection, finding that a decision tree classifier yielded the best results predicting the target "Biopsy" with an accuracy of 97.5%. W. Wu and H. Zhou [31] performed feature selection with PCA and used three methods of Support Vector Ma-

chine to analyse the dataset: Standard SVM, support vector machine recursive feature elimination and support vector machine-principal component analysis. Their standard SVM model produced an accuracy of 94.13 % in predicting the response variable"Biopsy", with 100% sensitivity and 90.21% specificity.

The approach taken in this paper can be differentiated from those mentioned previously in that they have either removed 3 of the 4 response variables ("Hinselmann", "Schiller" and "Cytology") leaving only "Biopsy" as the target or have carried out separate analyses with each of the 4 responses as a target and excluded the other 3.

This analysis proposes to include "Hinselmann", "Schiller" and "Cytology" as features leaving "Biopsy" as the single response. The rationale for this is that each of those variables is the result of a test carried out to determine the presence of abnormal cells [7] [25] [5]. Therefore they can be used as features to contribute to the outcome of a biopsy result and the presence of cervical cancer.

3.1 Implementation

The analysis was carried out using Python, with the loading/summarising of data achieved via NumPy/Pandas, while visualisations were achieved via graphical packages Seaborn and Matplotlib. The pre-processing, model building and evaluation were carried out via the Scikit-learn package, which encompasses a wide range of state-of-the-art machine learning algorithms [22]. To avoid the "Reproducibility Crisis" [3], where applicable, a global integer variable was created and assigned to the random state parameter for each method.

This analysis followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) process [29], which provides a formal standardised framework of 6 cyclical steps for planning and implementing data mining.

1. Business Understanding – Achieved through the related work, introduction and evaluation sections.
2. Data Understanding – The related work showed that the dataset features were suitable for this analysis, and exploratory data analysis gave further insight into the data.
3. Data Preparation – Built on from step 2 and achieved through pre-processing tasks such as missing value imputation, dealing with outliers, class imbalance and train/test splitting.
4. Modelling – Building the models and applying parameter tuning.
5. Evaluation – Comparing the models' results to determine the optimal model.
6. Deployment – Releasing the model to the production environment.

Data preparation involved processing the data with regards to outlier detection, handling missing values via mean/median imputation, and dealing with imbalance using over, under and combination resampling techniques.

The removal of outliers should be considered in the context of the effect their removal would have on analysis. To manipulate the outliers, for instance, replace them with mean/median values or remove observations, could negatively

impact the accuracy of the models either by the reduction in sample size or by the narrowing of values the models could accurately account for. As such, it was decided that potential outliers should be included.

Missing data can occur for several reasons, be it difficulties encountered during an experiment, errors during data collection or entry, or a systemic omission of answers by respondents. The latter occurs here, with respondents choosing not to answer certain questions due to privacy concerns [9]. Missing data rates of less than 1% are generally considered trivial, and those between 1-5% are manageable. However, 5-15% requires imputation techniques to handle, and more than 15% may severely impact any kind of interpretation or conclusions [8].

The dataset has a total possible 30,888 (858 x 36) available data points. Of these, 3,622 or 11.73% data points have missing values, while 27,266 data points are populated. Figure 1 shows the extent of missing data. Note, that only 26 variables are shown as 10 variables had no missing data.

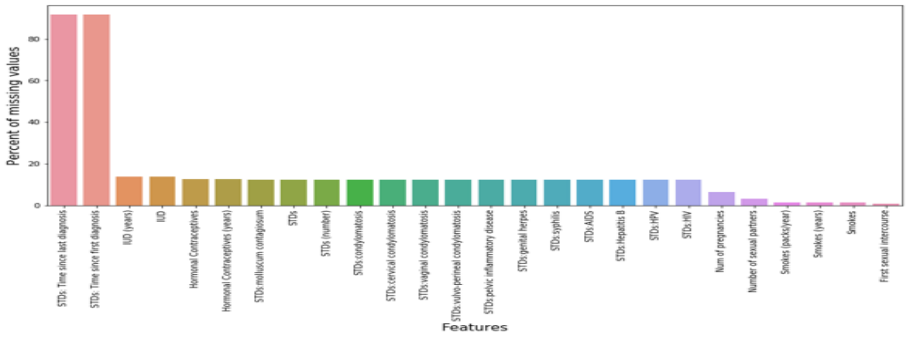


Fig. 1. Barplot shows two features with approx 92% missing data which were removed, the remaining 24 feature's missing data were imputed using the mean or median of the respective feature

Removing observations where missing data occurs will reduce the sample size and in turn, reduce the accuracy of any predictive models, it can also bias the data making any conclusions drawn not truly representative of the population. As such, it is typically preferable to use imputation techniques to estimate the missing values rather than remove observations. Imputation is the process of estimating a missing value based on valid values of other variables and/or subjects/observations in the sample.

A dataset is unbalanced when at least one class is represented by only a small number of training examples while other classes make up the majority. This imbalance gives rise to the class imbalance problem [18], which occurs when the majority class(s) observations greatly outnumber that of the minority class(s) observations in a machine learning problem. Here, the response variable Biopsy has an imbalance of 803 negatives observations to 55 positives observations.

Imbalanced-learn is a python package that offers several resampling techniques that solve this Class Imbalance problem. From this package 6 methods, 2 from each category of over-sampling, under-sampling and combination tech-

niques were used. This led to the creation of 6 augmented datasets of varying size, consisting of 34 features, including the response Biopsy. Table 1 shows the method used, the number of observations and count of the target variable Biopsy in the newly augmented datasets .

Method	**Type**	**Observations**	**Biopsy Response**	
			0	1
Random Oversampling	Oversampling	1606	803	803
Adative Synthetic Sampling	Oversampling	1617	803	814
Random UnderSampling	Undersampling	110	55	55
Neighbourhood Cleaning Rule	Undersampling	725	670	55
SMOTETomek	Combination	1600	800	800
SMOTE Edited Nearest Neighbour	Combination	1429	652	777

Table 1. Balancing datasets: proposed Data frames to address imbalance.

For each augmenting method used, a new dataset was created, each of which along with the original pre-processed dataset were shuffled and split into train and test sets (80/20 split) via the Scikit-learn model_selection module. Following this, 7 lists were created to hold the respective split data from each dataset; this enabled the values to be accessed globally from the function. It should be noted that some augmenting methods produce float values, where bool/int values are required, these were converted/rounded to the desired format.

Following the previously outlined pre-processing steps, the building of the models from the training sets was carried out, and the test sets were then evaluated. This process is associated with steps 4 and 5 of CRISP-DM. Scikit-learn provides several modules and methods to accomplish this. Where applicable the random state for each model was set to 3 for reproducibility, and hyperparameter optimisation techniques to find the optimal values for each model were employed.

Models 1 & 2: Decision Trees are a non-parametric supervised learning technique. For a classification tree, predictions of each observation are made by the most commonly occurring class of training observations in the region to which it belongs. This is achieved through recursive binary splitting – a greedy (better split now rather than later) top-down method that splits the nodes (variable) into two branches moving down at each split towards a leaf decision node which represents the response. Here, the DecisionTreeClassifier method from the Tree module was used. It employes an optimised version of the CART algorithm. With this, two models were created, Model 1 which has it's criterion set to "entropy" and Model 2 where it is set to "gini".

Model 3: Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the assumption that features are independent of one another. The GaussianNB method form the naive_bayes module was used. This method assumes the data follows a normal distribution.

Model 4: Gradient Boosting is a machine learning technique that combines several weak learners, typically decisions trees to form a model. The GradientBoostingClassifier method was implemented via the ensemble module. It has several tuning parameters, n_estimators - the number of boosting stages to per-

form, which was set to 100, learning_rate - shrinks the contribution of each tree, which was set to 1, and max_depth - maximum depth of the individual regression estimators, which was set to 2.

Model 5: K-means clustering is the most widely used unsupervised learning technique. It seeks to partition a dataset into K (specified by the user) distinct, non-overlapping clusters. Implemented via the KMeans method from the cluster module. The n_clusters parameter - the number of clusters and centroids to generate, was set to 2 when tuning this model.

Model 6: K Nearest Neighbours is a non-parametric method used for classification and regression analysis. KNN is sensitive to imbalanced datasets, a point to note in relation to this analysis. If the value for K is too small then it becomes susceptible to noise, if too large it becomes susceptible to bias. Typically when choosing K the square root of the number of samples in the training set is used. The KNeighborsClassifier method from the neighbors module was used to implement KNN. When tuning this model the distance method was set to 2 for euclidean_distance, and the value of K was determined by tuning the n_neighbors parameter as seen in Figure 2 on one of the augmented datasets.

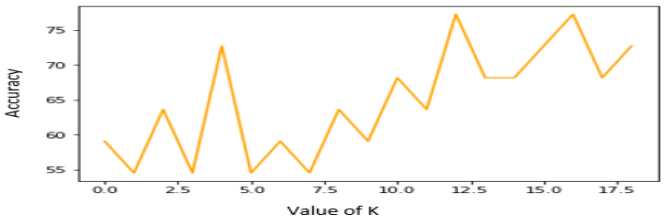


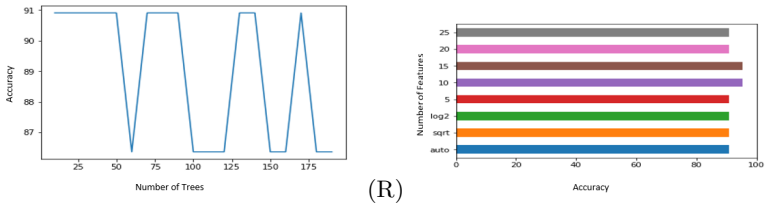
Fig. 2. Accuracy of the KNN model for different values of k when applied to one of the augmented datasets. This was used to tune the n_neighbors parameter when determining the end KNN model.

Model 7: Linear Discriminant Analysis is a classification technique that uses a linear decision boundary, created by fitting class conditional densities to a dataset and using Bayes' rule, it assumes a normal distribution. It is implemented here through the use of the LinearDiscriminantAnalysis method from the discriminant_analysis module. When tuning this model, the solver was set to "svd"- Singular value decomposition.

Model 8: Logistic Regression is a classification algorithm typically used in binary classification problems, such as the case here with negative, 0 and positive, 1 response values. In the logistic model, the log-odds (the logarithm of the odds) for the value "1" is a linear combination of one or more independent features. The LogisticRegression method from the linear_model module was used, with the solver parameter set to "liblinear".

Model 9: Random Forests are an ensemble learning method that construct numerous decision trees during data training, outputting the class that is the mode of the classes for classification of the individual trees. Random Forests correct for a decision trees' habit of overfitting to their training set. The RandomForestClassifier() method from the ensemble module was used for this anal-

yses. Parameters tuned to optimise this model were max_features which is the maximum number of variables RF can test in each node, and the n_estimators parameter, which is the number of trees that are built before the average is taken.



(L) (R)

Fig. 3. (L) Depicts the accuracy of the RF model for a different number of trees (n_estimators). While (R) shows the accuracy of the RF model for the different number of features (max_features) when applied to the dataset.

Model 10: Support Vector Machines (SVM) find a boundary known as a hyperplane in an N-dimensional space that classifies the data points into discrete categories depending on which side of the boundary they lie. Here the svm method was imported through svm module. SVC is a form of SVM for dealing with classification analyses.

4 Results

Many classification algorithms aim to minimise the error rate and obtain a higher accuracy result. They assume that the cost of all misclassification errors is equal. This approach can be problematic, particularly in relation to the area of health.

If a positive result indicates the presence of cancer, and a negative result indicates it's absence, then the consequences of classifying a patient as negative when in fact they are positive - False Negative, is more severe than classifying the patient as positive when they are in fact negative - False Positive [10].

A more accurate metric to use is sensitivity, also known as the True Positive (TP) Rate. This is the proportion of people that tested positive and actually are positive. It can be considered the probability that the test is positive, given that the patient is ill. With higher sensitivity, fewer actual cases of disease go undetected, or in the case of the cancer models, fewer patients that have cancer go undetected. Specificity (TN) is the opposite of this.

The Scikit-learn metric module provides the functionality to produce a classification report which includes values such as Precision, Recall and F1-score, as well a confusion matrix via the accuracy_score, classification_report, and confusion_matrix methods. A description of these metrics can be seen in Table 2.

Table 3 denotes the accuracy, precision, recall, and F1 results of the original cleaned dataset, and the 6 resampled datasets, consisting of 2 over, under, and combination sampled datasets. The legends for the models and databases are denoted on the right hand side of the table.

Term	Formula
Accuracy	(TP+TN)/(TP+FP+TN+FN)
Sensitivity/Recall/ TP Rate	TP/(TP+FN)
Specificity / TN Rate	TN/(TN+FP)
Precision	TP/(TP+FP)
F-Measure	(2*TP*TN)/(TP+TN)

Table 2. Formula for each of the criteria a model is evaluated under.

When taking accuracy as a metric, Table 3 shows that the Naive Bayes model was consistently a poor performer across the 7 datasets, scoring results as low as 9.88% and 12.41% in the original and NCR undersampled datasets respectively. In comparison, both Decision Tree models scored above 90% in all models except for the NCR undersampled dataset. The Random Forest model scored the highest getting above 90% for each dataset.

When viewing the original cleaned dataset (OC) it can be seen that several models failed to predict any of the positive cases correctly. The LDA model had an accuracy of 94.19% and correctly predicted 9 of the 11 positive cases yielding a recall of 82%. The Random Forest model also had an accuracy of 94.19%, however it only had a recall of 55% or predicted 6 of the 11 positive observations.

The Random Over Sampled dataset (ROS) shows that the 3 tree models all produced an accuracy result greater than 98%, with all 3 having a recall of 100% for the positive diagnosis observations.

When viewing the Adaptive Synthetic Sampling Over–Sampled dataset (ASS), it can again be seen that the 3 tree models perform well with an accuracy greater than 98%. They also produce a precision and recall result of 99% for both positive and negative outcomes. The Random Under Sampled dataset (RUS) shows that the Gini Decision Tree model as well as the Linear Discriminant Analysis model perform very well, with an accuracy of 95.45% and both precision and recall for positive and negative observations above 90% in both models.

When viewing the Neighbourhood Cleaning Rule dataset (NCR), it can be seen that 8 of the models produce an accuracy of above 90%, however from these 8 models only 2 (LDA & LR) produce a positive recall value greater than 70%. This again highlights the caution needed when using accuracy as a metric with imbalanced data.

The SMOTE-Tomek combination sampled dataset (S-TOM) produces the model with the most promising results in this analysis. The Random Forest model generates an accuracy of 99.69% with both positive and negative precision and recall values almost being 100%, and an F1 result of 1 for both positive and negative outcomes. Here the KNN model also does well when compared to it's performance in the other datasets.

When viewing the Smote ENN combination sampled dataset (S-ENN), it can be seen that again the three tree methods perform well with high recall and precision results for both positive and negative outcomes. In 5 of the 7 datasets, the Naive Bayes model assigns the majority of observations to the positive category, resulting in its poor overall performance, but high positive recall results.

Accuracy												Models Legend	
		DT-E	**DT-G**	**GNB**	**GB**	**KM**	**KNN**	**LDA**	**LR**	**RF**	**SVC**	**Model**	**Key**
OC		93.02	91.28	9.88	93.6	42.44	93.6	94.19	93.6	94.19	93.6	Decision Tree (Entropy)	DT-E
ROS		98.14	98.45	53.42	90.99	50.62	95.65	91.93	91.93	98.76	81.99	Decision Tree (Gini)	DT-G
ASS		98.77	98.46	50.93	85.19	55.25	93.21	95.06	95.99	99.38	89.51	Gaussian Naive Bayes	GNB
RUS		81.82	95.45	63.64	90.91	45.45	72.73	95.45	86.36	90.91	50	Gradient Boosting	GB
NCR		90.34	90.34	12.41	92.41	60	91.72	93.1	94.48	93.1	92.41	K-Means	KM
S-TOM		97.81	98.75	55.94	87.5	54.69	95.62	69.25	94.69	99.69	89.69	K-Nearest Neighbour	KNN
S-ENN		95.45	96.85	76.57	81.82	46.85	97.2	93.01	92.32	98.6	83.92	Linear Discriminant Analysis	LDA
Precision												Logistic Regression	LR
OC	0	0.96	0.96	1	0.94	0.94	0.94	0.99	0.97	0.97	0.94	Random Forest	RF
	1	0.45	0.33	0.07	0	0.07	0	0.53	0.5	0.55	0	Support Vector Classifier	SVC
ROS	0	1	1	1	0.86	0.49	1	0.89	0.89	1	0.74		
	1	0.97	0.97	0.53	0.97	0.54	0.92	0.95	0.95	0.98	0.95	**Database Legend**	
ASS	0	0.99	0.99	1	0.99	0.55	0.99	0.93	0.95	0.99	0.85	**Dataset**	**Key**
	1	0.99	0.98	0.5	0.77	0.55	0.88	0.98	0.97	0.99	0.95	Original (Cleaned)	OC
RUS	0	0.8	1	0.56	0.83	0.43	0.64	0.91	0.77	0.9	0.48	Random Over Sampled	ROS
	1	0.83	0.92	0.83	1	0.5	0.88	1	1	0.92	1	Adaptive Synthetic Sampling	ASS
NCR	0	0.95	0.96	1	0.92	0.92	0.92	0.98	0.98	0.97	0.92	Random Under Sampled	RUS
	1	0.38	0.4	0.08	0	0.07	0	0.53	0.62	0.54	0	Neighbourhood Cleaning Rule	NCR
S-TOM	0	0.97	0.98	1	0.98	0.51	0.99	0.93	0.92	1	0.84	SMOTETomek	S-TOM
	1	0.98	0.99	0.55	0.82	0.61	0.93	0.99	0.98	0.99	0.95	SMOTE ENN	S-ENN
S-ENN	0	0.98	0.98	0.94	0.97	0.42	1	0.91	0.91	1	0.81		
	1	0.94	0.95	0.7	0.75	0.49	0.95	0.95	0.94	0.97	0.87		
Recall													
OC	0	0.96	0.95	0.04	1	0.41	1	0.95	0.96	0.97	1		
	1	0.45	0.36	1	0	0.64	0	0.82	0.55	0.55	0		
ROS	0	0.96	0.97	0.03	0.97	0.69	0.91	0.95	0.95	0.97	0.96		
	1	1	1	1	0.85	0.34	1	0.89	0.89	1	0.96		
ASS	0	0.99	0.98	0.04	0.72	0.68	0.87	0.98	0.97	0.99	0.96		
	1	0.99	0.99	1	0.99	0.42	0.99	0.92	0.95	0.99	0.83		
RUS	0	0.8	0.9	0.9	1	0.6	0.9	1	1	0.9	1		
	1	0.83	1	0.42	0.83	0.33	0.58	0.92	0.75	0.92	0.08		
NCR	0	0.94	0.93	0.05	1	0.62	0.99	0.95	0.96	0.96	1		
	1	0.45	0.55	1	0	0.36	0	0.73	0.73	0.64	0		
S-TOM	0	0.98	0.99	0.05	0.74	0.66	0.91	0.99	0.97	0.99	0.95		
	1	0.98	0.98	1	0.99	0.45	0.99	0.94	0.92	1	0.85		
S-ENN	0	0.93	0.95	0.52	0.64	0.31	0.94	0.95	0.93	0.97	0.87		
	1	0.98	0.99	0.97	0.98	0.61	1	0.91	0.91	1	0.81		
F1-Score													
OC	0	0.96	0.95	0.07	0.97	0.57	0.97	0.97	0.97	0.97	0.97		
	1	0.45	0.35	0.12	0	0.12	0	0.64	0.52	0.55	0		
ROS	0	0.98	0.98	0.06	0.91	0.57	0.95	0.92	0.92	0.99	0.84		
	1	0.98	0.99	0.69	0.91	0.41	0.96	0.92	0.92	0.99	0.8		
ASS	0	0.99	0.98	0.08	0.83	0.61	0.93	0.95	0.96	0.99	0.9		
	1	0.99	0.98	0.67	0.87	0.48	0.93	0.95	0.96	0.99	0.89		
RUS	0	0.8	0.95	0.69	0.91	0.5	0.75	0.95	0.87	0.9	0.65		
	1	0.93	0.96	0.56	0.91	0.4	0.7	0.96	0.86	0.92	0.15		
NCR	0	0.95	0.95	0.1	0.96	0.74	0.96	0.95	0.97	0.96	0.96		
	1	0.42	0.46	0.15	0	0.12	0	0.62	0.67	0.58	0		
S-TOM	0	0.98	0.99	0.09	0.85	0.57	0.95	0.96	0.94	1	0.9		
	1	0.98	0.99	0.71	0.89	0.52	0.96	0.96	0.95	1	0.9		
S-ENN	0	0.95	0.97	0.69	0.77	0.36	0.97	0.93	0.92	0.99	0.84		
	1	0.96	0.97	0.81	0.85	0.55	0.97	0.93	0.93	0.99	0.84		
		DT-E	**DT-G**	**GNB**	**GB**	**KM**	**KNN**	**LDA**	**LR**	**RF**	**SVC**		

Table 3. Results denoting the accuracy, precision, recall, and F1 of the models tested on the six databases. Model and Database legends are denoted on the upper right hand.

5 Conclusion

This paper shows a comparison of classification techniques used for predicting the outcome of biopsy results based on known risk factors and screening tests. It also highlights the relevance and study of these known risk factors used in this classification process.

Pre-processing techniques were employed to address missing data and imbalance, and where applicable parameter tuning was employed to find optimal values for models. It was shown that imbalanced data can influence the outcome of predictive models, highlighting the need to pre-processing techniques to address said issue. It also showed that accuracy is not an acceptable measure for imbalanced data, and in particular health data.

From the models tested, the Random Forest model was shown to be superior at predicting the biopsy response, yielding high accuracy, precision and recall values, while the Gauissian Naïve Bayes model was the poorest predictor. The combination resampling method SMOTE-Tomek's dataset, in conjunction with a Random Forest model produced the highest result with an accuracy of 99.69%, and a precision and recall of 99% for both negative and positive targets.

References

1. Alwesabi, Y., Choudhury, A.: Classification of cervical cancer dataset (2018)
2. Averbach, S., et al: Recent intrauterine device use and the risk of precancerous cervical lesions and cervical cancer. Contraception 98 (04 2018)
3. Baker, M.: Is there a reproducibility crisis? Journal of Machine Learning Research (2016)
4. Bosch, F.X., et al: Risk factors for cervical cancer in colombia and spain. International journal of cancer 52 5, 750–8 (1992)
5. Dillner, J., et al: Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: joint european cohort study. BMJ 337 (2008), https://www.bmj.com/content/337/bmj.a1754
6. Eldridge, R.C., Pawlita, M., Wilson, L., Castle, P.E., Waterboer, T., Gravitt, P.E., Schiffman, M., Wentzensen, N.: Smoking and subsequent human papillomavirus infection: a mediation analysis. Annals of Epidemiology 27(11), 724–730.e1 (2017)
7. Eraso, Y.: Migrating techniques, multiplying diagnoses: the contribution of Argentina and Brazil to early 'detection policy' in cervical cancer 17 (2010)
8. Farhangfar, A., Kurgan, L., Dy, J.: Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 41(12), 3692–3705 (2008)
9. Fernandes, K., Cardoso, J.S., Fernandes, J.: Transfer learning with partial observability applied to cervical cancer screening. In: IbPRIA (2017)
10. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering (2012)
11. Greenlee, R.T., Murray, T., Bolden, S., Wingo, P.A.: Cancer statistics, 2000. CA: A Cancer Journal for Clinicians 50(1), 7–33 (2000)
12. HSE: Cervicalcheck. http://www.cervicalcheck.ie/news-and-events/information-for-healthcare-professionals-from-cervicalcheck-latest-update.14910.html (2019), [Online; accessed 2019-10-11]

13. HSE: Cervicalcheck. https://www.hse.ie/eng/cervicalcheck (2019), [Online; accessed 2019-10-11]
14. HSE: Cervicalcheck: Screening information. https://www.hse.ie/eng/cervicalcheck/screening-information/why-you-are-offered-a-free-cervical-screening-test/cervical-cancer.html (2019), [Online; accessed 2019-10-11]
15. Irish Cancer Society: Irish cancer society annual report 2017. https://www.cancer.ie/about-us/who-we-are/annual-reports-accounts#sthash.8McZayy5.dpbs (2019), [Online; accessed 2019-10-11]
16. Kelwin Fernandes, J.S.C., Fernandes, J.: Transfer learning with partial observability applied to cervical cancer screening. https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+\%28Risk+Factors\%29, [accessed 2019-10-11]
17. Koh, W.J., et al: Cervical cancer, version 2.2015. Journal of the National Comprehensive Cancer Network : JNCCN 13, 395–404 (04 2015)
18. Lemaître, G., Nogueira, F., Aridas, C.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning 18 (09 2016)
19. Manderson, L., Markovic, M., Quinn, M.: Like roulette: Australian women's explanations of gynecological cancers. Social science & medicine (1982) (2005)
20. NCRI: Cervical cancer trends. https://www.ncri.ie/sites/ncri/files/pubs/CervicalCaTrendsReport_35.pdf (2019), [Online; accessed 2019-10-11]
21. Parthenis, C., Panagopoulos, P., et al: The association between sexually transmitted infections, human papillomavirus and cervical cytology abnormalities among women in greece. International Journal of Infectious Diseases 73 (06 2018)
22. Pedregosa, F., Varoquaux, G., et al: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12 (01 2012)
23. Rousset-Jablonski, C., Reynaud, Q., Nove-Josserand, R., Durupt, S., Durieu, I.: Gynecological management and follow-up in women with cystic fibrosis. Revue des maladies respiratoires 35(6), 592—603 (June 2018)
24. Santelli, J., Brener, N., Lowry, R., Bhatt, A., Zabin, L.: Multiple sexual partners among u.s. adolescents and young adults. Perspectives on Sexual and Reproductive Health 30(6), 271–275 (11 1998)
25. Sesti, F., Ticconi, C., Santis, L.D., Piccione, E.: Clinical value of schiller's test in colposcopic examination of the uterine cervix. Journal of Obstetrics and Gynaecology 10(6), 545–547 (1990)
26. Shukla, A., Jamwal, R.: Adverse effect of combined oral contraceptive pills. Asian Journal of Pharmaceutical and Clinical Research Volume 10, 17–21 (01 2017)
27. Teame, H., et al: Factors associated with cervical precancerous lesions among women screened for cervical cancer in addis ababa, ethiopia (2018)
28. Walsh, J., O'Reilly, M., Treacy, F.: Factors affecting attendance for a cervical smear test: A prospective study. irish cervical screening programme and the national university of ireland, galway
29. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. Journal of Machine Learning Research (2000)
30. World Health Organisation: Hpv and cervical cancer. https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer (2019), [Online; accessed 2019-10-11]
31. Wu, W., Zhou, H.: Data-driven diagnosis of cervical cancer with support vector machine-based approaches. IEEE Access 5, 25189–25195 (2017)
32. Xu, H., et al: "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 (2018)