information

Article

A Text Mining Approach in the Classification of Free-Text Cancer Pathology Reports from the South African National Health Laboratory Services

Okechinyere J. Achilonu [1,*], Victor Olago [2], Elvira Singh [1,2], René M. J. C. Eijkemans [3], Gideon Nimako [1,4] and Eustasius Musenge [1]

[1] Division of Epidemiology and Biostatistics, Faculty of Health Sciences, School of Public Health, University of the Witwatersrand, Parktown, Johannesburg 2000, South Africa; ElviraS@nicd.ac.za (E.S.); gideonnimako@gmail.com (G.N.); Eustasius.Musenge@wits.ac.za (E.M.)

[2] National Cancer Registry, National Health Laboratory Service, 1 Modderfontein Road, Sandringham, Johannesburg 2131, South Africa; VictorO@nicd.ac.za

[3] Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht University, 3584 Utrecht, The Netherlands; m.j.c.eijkemans@umcutrecht.nl

[4] Industrialization, Science, Technology and Innovation Hub, African Union Development Agency (AUDA-NEPAD), Johannesburg 1685, South Africa

* Correspondence: achilonu.okechinyere@gmail.com; Tel.: +27-73-412-3085; Fax: +27-86-765-3889

Citation: Achilonu, O.J.; Olago, V.; Singh, E.; Eijkemans, R.M.J.C.; Nimako, G.; Musenge, E. A Text Mining Approach in the Classification of Free-Text Cancer Pathology Reports from the South African National Health Laboratory Services. *Information* **2021**, *12*, 451. https://doi.org/10.3390/info12110451

Academic Editors: Mario Ciampi and Mario Sicuranza

Received: 22 May 2021
Accepted: 27 June 2021
Published: 30 October 2021

Abstract: A cancer pathology report is a valuable medical document that provides information for clinical management of the patient and evaluation of health care. However, there are variations in the quality of reporting in free-text style formats, ranging from comprehensive to incomplete reporting. Moreover, the increasing incidence of cancer has generated a high throughput of pathology reports. Hence, manual extraction and classification of information from these reports can be intrinsically complex and resource-intensive. This study aimed to (i) evaluate the quality of over 80,000 breast, colorectal, and prostate cancer free-text pathology reports and (ii) assess the effectiveness of random forest (RF) and variants of support vector machine (SVM) in the classification of reports into benign and malignant classes. The study approach comprises data preprocessing, visualisation, feature selections, text classification, and evaluation of performance metrics. The performance of the classifiers was evaluated across various feature sizes, which were jointly selected by four filter feature selection methods. The feature selection methods identified established clinical terms, which are synonymous with each of the three cancers. Uni-gram tokenisation using the classifiers showed that the predictive power of RF model was consistent across various feature sizes, with overall F-scores of 95.2%, 94.0%, and 95.3% for breast, colorectal, and prostate cancer classification, respectively. The radial SVM achieved better classification performance compared with its linear variant for most of the feature sizes. The classifiers also achieved high precision, recall, and accuracy. This study supports a nationally agreed standard in pathology reporting and the use of text mining for encoding, classifying, and production of high-quality information abstractions for cancer prognosis and research.

Keywords: pathology reports; breast; colorectal; prostate; text mining; machine learning; support vector machine and random forest

1. Introduction

Cancer is one of the most frequent causes of death in South Africa [1], and breast, prostate, and colorectal are among the top five cancers occurring in the South African adult population. According to Stefan [2], adequate attention to cancer diagnoses is needed to improve the overall health of South Africans. Accurate diagnosis is a major concern in the health care system for optimal prognosis and treatment of cancer [3]. A cancer pathology report is a valuable medical document that provides information for clinical management of the patient and evaluation of health care [4]. Lack of accurate and/or readily available data that can offer feedback to physicians negatively impacts health care quality. The health care system needs to learn from past cancer pathology reports to improve cancer prognosis and the overall health of cancer patients. Several studies in the literature have audited the information content of pathology reports regarding (i) veracity in reporting, (ii) compliance to a set standard, and (iii) comprehensiveness [5,6]. The overall aim of these studies was to evaluate the quality of pathology reports to improve the conformity to a set international or national standard.

The Corporate Data Warehouse (CDW), a division of the South African National Health Laboratory Service (NHLS), collects and archives cancer pathology reports, which covers approximately 80% of the national population [7]. Besides the direct use of these pathology reports in the health care system, these reports can also be used in research, quality and safety evaluations, and cancer incidence reporting [8]. Generally, pathology reports comprise massive valuable data relating to the clinical condition of the patient. However, these reports are usually in semi-structured or unstructured free-text format. In this case, data coders manually read the pathology reports, extract valuable information, and interpret the information based on clinical rules [9]. The final results are recorded in the database, with a minimum computer-assisted process [10].

The high throughput generation of pathology reports due to increasing cancer incidence has made a manual translation of unstructured free-text pathology reports to structured data overwhelming and time-consuming [10]. Therefore, an automated process is needed to abstract information in pathology reports, which may result in significant cost-savings and consistent data structure, especially for research and incidence reporting. In this regard, text mining (TM) has emerged as a powerful computational technique to extract meaningful information and accurately transform pathology reports into a usable structured representation [10–12]. TM techniques have been widely applied to unstructured data such as medical records, social media content, business documents, survey responses, academic publications, and web pages. These unstructured data are effectively transformed into a machine-readable structured representation suitable for text classification using a variety of machine learning (ML) algorithms.

In the medical field, ML has gained a wide range of applications and has provided the potential to improve patient outcomes [13]. The advantages of ML have been emphasised in the analysis of medical images, human genetics, and electronic medical record data, with a focus on diagnosis, detection, and prediction [14]. Precisely, ML has been successfully applied in the text classification of medical documents. Its goal is to automatically build a classifier from training samples to assign documents into a set of pre-defined category labels [15]. For instance, the study by Hyland et al. [16] showed successful application of supervised ML in predicting circulatory failure in the intensive care unit. In the automatic prediction of heart disease, Ali et al. [17] proposed integration of deep learning technique and feature fusion approaches using sensor and electronic medical health data. ML has also been extensively applied to other medical data with high accuracy in text categorisation [14,18–21].

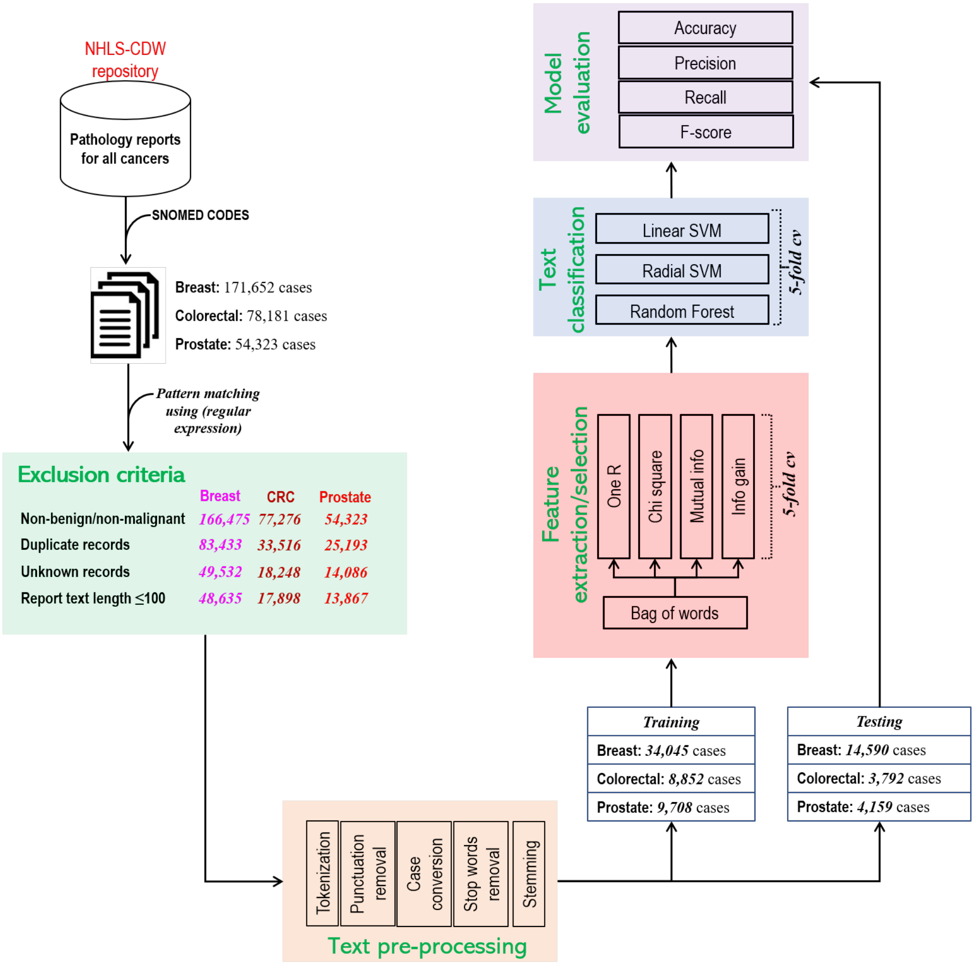
In cancer research, several studies exploited the advantages of using ML for automated text classification [8,11,22–27]. More specifically, Kasthurirathne et al. [22] evaluated the accuracy of cancer case identification within a free-text pathology report using public health data. Logistic regression (LR), naïve Bayes (NB), k-nearest neighbour (KNN), random forest (RF), and J48 decision tree were evaluated. Information gain was used as the feature selection method. The study showed a comparable performance of about 80–90% for the decision models for most of the evaluation metrics used. Ranking of the specific tokens associated with the presence and absence of cancer was shown. Kalra et al. [25] investigated the performance of SVM, XGBoost, and LR based on the features selected using the term frequency-inverse document frequency (TF-IDF) method in classifying pathology report into different diagnosis categories. The XGBoost classifier achieved the highest classification accuracy of 92%. Another study by Jouhet et al. [8] trained SVM and NB to categorise pathology reports. Based on the ICD-03 attribution, the SVM model achieved 72% and 85% F-measure for topography and morphology. In a more recent study by our group [27], all 2016 cancer pathology reports from the Western Cape province, curated by the NHLS-CDW South Africa, were analysed. Various ML algorithms, including RF, LR, NB, SVM, and KNN, were explored to classify pathology reports into malignant, non-malignant, and no diagnosis. Among other classifiers, SVM achieved the highest F-measure of 97%.

Generally, supervised ML algorithms have been used in the classification of cancer pathology. Among these algorithms, SVM and RF are commonly used and have shown good performance in the classification of cancer pathology reports [8,27]. This study considered these supervised ML techniques for categorising cancer pathology reports into malignant or non-malignant classes. Although previous studies used different approaches in processing unstructured report, substantial variations and noise that exist in pathology reports have been shown to impact the generalisability of these studies [11,22]. Hence, there seems to be no optimal and evidence-based practice published to extract information from the pathology report. Further, studies done in pathology report classification have not considered assessing consistency in reporting over the years and using a combination of filter feature selection techniques. In this study, we allowed the filter feature methods to automatically reduce the dimensionality of thousands of tokenised features by ranking the importance of each feature in relation to the target cancer. Such a framework will reduce computational time and the chances of modelling noise.

This study aimed to use ML approaches to evaluate the veracity of over 80,000 free-text breast, colorectal and prostate cancer pathology reports archived in the South African NHLS-CDW. Figure 1 is an illustration of the methodological approach used in this study. The two diagnosis classes ("benign" and "malignant") were defined by mapping the attributes of the Systematised Nomenclature of Medicine (SNOMED) codes to the International Classification of Diseases for Oncology (ICD-03) codes. This study is an important and useful development within the context of cancer pathology reporting in South Africa. It may form the foundation for several research studies that will eventually utilise the NHLS-CDW database. Unlike in the study by Olago et al. [27] and previous studies, our contributions are as follows:

- Our effort was based on the premise that almost a decade of cancer pathology reports (used in this study) may yield more valuable insights into the integrity of the curation of these pathology reports and consistency and style in cancer pathology reporting over the years. Therefore, our data coverage is more comprehensive, which also reflects the population of South Africa.
- Various cancers have some unique clinical key terms specific to that cancer (in their reports) and are entirely different from other cancers. For instance, we expect a breast cancer pathology report to include key terms such as "estrogen receptor and progesterone receptor", which should not be present in colorectal or prostate cancer pathology reports. Hence, we analysed cancer-specific reports instead of aggregating reports of all cancers in a single analysis.
- We perform a detailed comparative evaluation of features selected by the different feature selection techniques and assessed if the integration of the selected features could impact the algorithms' efficiency. The selected features were manually reviewed to ascertain the reliability of any term used in the model building.

Therefore, our research questions are formulated as follows: (1) How efficient can TM be used to process cancer pathology reports? (2) How well can filter selection methods identify the most effective terms associated with a cancer diagnosis? (3) How effective can the two classes be discriminated against by the ML classifiers?



	Breast	CRC	Prostate
Non-benign/non-malignant	*166,475*	*77,276*	*54,323*
Duplicate records	*83,433*	*33,516*	*25,193*
Unknown records	*49,532*	*18,248*	*14,086*
Report text length ≤100	*48,635*	*17,898*	*13,867*

Figure 1. An overall workflow for this study, illustrating the inclusion and exclusion criteria used to extract the final study data and also the modelling approach used in data analysis.

2. Materials and Methods

2.1. Data Source

The NHLS is a system of population-based public health laboratory services for South Africa [7]. It is positioned as the largest diagnostic pathology service in South Africa and Sub-Saharan Africa. The NHLS has a national network of laboratories across all the provinces of South Africa and provides public health and laboratory services covering approximately 80% of the total South African population. It offers a reference to both provincial and national health departments to support appropriate response to quality and accessible health care delivery and health decision making. The NHLS-CDW is the central data repository of laboratory results within the public health sector in South Africa. The focus of NHLS-CDW is to collect and archive pathology reports by characteristics of type, time, person, and place. More specifically, the database archives several types of cancer pathology reports, which are used by the National Cancer Registry in monitoring and reporting regional or national cancer incidence [7]. Permission to use the study data was obtained from the NHLS research committee, and ethical clearance was obtained from the Human Research Ethics Committee of the University of the Witwatersrand (M1911131).

2.2. Preprocessing

We extracted 171,652, 78,181, and 54,323 electronic pathology reports corresponding to breast, colorectal, and prostate cancers, respectively, diagnosed between 2011 and 2019, as a CSV file from the NHLS-CDW database (Figure 1). The pathology reports are free-text written in English, with few terms in Afrikaans, routinely collected from all the

NHLS pathology laboratories across South Africa and were annotated by the pathologist. We used the *stringr* function package in R [28] for pattern matching of the values of the SNOMED morphology codes (to create the class label) using the International Classification of Diseases for Oncology (ICD-O-3) code as the underlying terminology for histology and anatomical sites in the pathology reports. Within the ICD-O-3 codes, there is more than one possible generic classification of cancer group, including the following: M-8000/0 (benign), M-8000/1 (uncertain), M-8000/2 (in situ), M-8000/3 (invasive), and M-8000/6 (metastatic). We observed that majority of the morphology values fell under the benign or the malignant groups. We then excluded few cases with indecisive diagnosis. We reduced the number of possible class labels by bucketing all the carcinoma cases to create two response categories, "Benign" and "Malignant", for each studied cancer.

Several preprocessing steps were performed [29]. First, duplicate records were identified and filtered out from the data. We also excluded cases with unknown records and cases with a maximum report text length of 100 (Figure 1). Most reports with a text length of 100 contain one, two, or few words that show no specific description of benign or malignant diagnosis. Second, important terms such as "ER+VE" and "PSA" that were inconsistently written in the pathology reports were standardised. The R package "quanteda" was used to replace non-useful characters, such as punctuation and parentheses, from the reports with empty spaces [30]. This package was also used to convert the text in the reports to lower case, discard non-informative (*stopwords*) words [31] and other words such as patient, email, name, etc.

2.3. Model Development and Validation

Before model building, the stemming function in the *quanteda* package was used to combine similar words that could have appeared in different forms in the report, which may affect the classifier's performance. A typical example of such a word is the "Bloom-Richardson" for breast cancer grading. We observed different ways of referring to this word in the reports using different parentheses and abbreviations. These types of words were aggregated into one using the stemming function. A function called tokens in the *quanteda* package was used to tokenise the report into uni-grams. Tokenisation is the process of breaking text in a document into words, phrases, or whole sentences [31]. As it is not possible to directly build a classifier using a text document, we employed "bag of word", an indexing approach to document representation, to map the reports into a compact representation of its content [32]. In other words, a vector representing the frequency of each token was created. There was an imbalance class distribution (72% vs. 28%) with the colorectal cancer data. To reduce the imbalance in the distribution of the class label, we subsampled the majority class (benign) using the *Ross* package in R software [33]. The final distribution was 60% vs. 40% for the benign vs. malignant class. Using a stratified sampling method in the *caret* package [34], 30% of each cancer case studied was set aside for model validation. The overall workflow of this study approach is illustrated in Figure 1.

2.3.1. Feature Selection

To reduce the number of dimensions in the datasets, we identified the most relevant features useful in supporting the classification process. This is particularly valuable as irrelevant or redundant features negatively impact the performance of the classifiers and may also increase the computing run-time [10,35]. Features are selected based on their rank in discriminating a malignant tumour from benign. In this study, we have employed a two-step feature selection approach. We first excluded features occurring fewer than 20 times in fewer than 500 documents in the training data. These features may add noise to the classification or may not help to differentiate a benign from a malignant case report. In addition, features with high frequency that could equally occur between benign and malignant classes were excluded from the training data because they may contribute to the classification. Next, we used four filter feature selection methods to identify terms or features that strongly differentiate the targeted classes. The list of the unique features selected across the feature selection methods for each cancer type was manually reviewed to identify words not familiar for any of the cancer types. The top-ranked 150 features common across the four selection methods were identified. In this process, we significantly reduced the original features from >20,000 to 150. The feature selection methods used in this study are information gain, mutual information, chi-square, and one rule. Each of these methods are well known and have been detailed in the literature [35–38]. Hence, we briefly describe these methods below. Given a collection of the training samples $S = (s_1, s_2, \ldots, s_n)$, each sample i contains a pathology report expressed as a numeric vector representing the features or terms $X = (x_1, x_2, \ldots, x_m) \in R^n$. We assume the diagnosis class $Y = (y_1, y_2)$ represents malignant and benign.

Information gain: Information gain (IG) determines the relevance of a feature to the outcome by counting its presence, or absence, in a report [35,37]. To describe IG, we start with Shannon's entropy, denoted by

$$H(Y) = -\sum_{y} P(y) \log_2 P(y) \tag{1}$$

Shannon's entropy describes the information content or uncertainty in the outcome class Y as described in Equation (1). A measure of the estimated information gain is calculated by comparing the entropy of the class label and the entropy of each specific feature.

$$IG = H(Y) - H(X) \tag{2}$$

Mutual information: Mutual information (MI) measures the dependency or relationship between the outcome class and the feature [35,38]. The conditional entropy is given by

$$H(Y/X) = -\sum_{x}\sum_{y} P(x,y) \log_2(P(y/x)), \tag{3}$$

denoting that the observation of the feature X, reduces the uncertainty in the outcome Y. The decrease in uncertainty is expressed as

$$I(M) = H(Y) - H(Y/X). \tag{4}$$

MI is equal to zero if X and Y are independent; otherwise, it will be greater than zero.

Chi-square test: The chi-square (χ^2) method calculates the association of feature X with class Y [35,36,38]. It measures the the lack of independence between the features and the class. χ^2 is defined by the following expression

$$\chi^2(X,Y) = \frac{n[P(xy)P(\bar{x}\bar{y}) - P(x,\bar{y})(y,\bar{x})]^2}{P(x)P(y)P(\bar{x})P(\bar{y})}, \tag{5}$$

where n is the number of the training sample.

One Rule: The one rule (OneR) learning algorithm is a decision tree with only one split [36]. OneR learning method infers a rule to predict the class for the values of the features. Similar to IG, these rules are based on each input feature.

2.3.2. Classifiers

Support vector machine: SVM is a well-known pattern recognition technique that has shown good performance in text classification studies [8,10,27]. SVM uses a hyperplane of coefficients w and b (with a maximum margin of separation) to model the discrimination between samples of two classes while minimising the total classifier error. A hyperplane in the sample space can be defined as $w.x + b = 0 \in R^n, b \in R$, where w is the weight vector perpendicular to the hyperplane and b is the bias term. If such a hyperplane exists in the training data, then the data is said to be linearly separable, and the optimal hyperplane is identified by solving the Lagrange multipliers (α_i) [10]. The decision boundary is defined by

$$f(x) = sign(\sum_{i=1}^{p} y_i \alpha_i k(x_i, x_j) + b), \tag{6}$$

where the kernel function $k(x_i, x_j) = x_i.x_j$ for the linear SVM. $xi.x_j$ represents the dot product between the input vector and each support vector. We also considered the radial kernel function, a scenario where the training data are not linearly separable. In this process, the training data is mapped into a higher dimensional space, where a linear hyperplane is constructed to perform separation. For this kernel, the $k(x_i, x_j)$ in Equation (6) is substituted by $\exp(-\gamma\|x_i - x_j\|^2)$. As described above, it may not be straightforward to identify a hyperplane with the maximum margin that can perfectly separate the two classes. A regularisation parameter C is used in linear SVM to assess the extent of our intention to perfectly separate the two-class label in the training data [39]. In other words, C controls the cost of misclassification in the training data during the optimisation process. Large values of C are associated with high variance and low bias. Conversely, low values of C result in high bias, low variance, and an increase in the training time. Using five-fold cross-validation (CV), a range of C values (1:2, by 0.05) were used for SVM linear, and performance at any given feature size was based on the F1-score, with support from recall, precision, and accuracy scores. The radial kernel function contains two hyperparameters C and γ [39]. The γ parameter is used to account for the smoothness of the decision boundary and controls the variance of the model. Smaller γ leads to a smoother decision boundary with low variance and vice versa.

Random forest: Random forest algorithm has been frequently used in text classification with good prediction performance [12,40], and this technique has shown more efficiency than the single decision tree. For predicting a new case, the trees' final class label prediction in the forest is observed. Popularity vote is used to select the class label of the new case with the highest votes. In this study, we have used the ranger method in the R caret package to fit the RF models [41]. The models were tuned over n of number tree (300:1000, by 100); the random forest algorithm is described as follows:

1. Randomly select sample from the training data, S.
2. Construct a classifier with selected samples.
3. Randomly select the number of features that maximise the information gain from the total features.
4. Use the best split among the selected features to calculate the node and the daughter node.
5. Repeat steps 3 and 4 until the required N number of nodes is reached.
6. Recursively repeat the above steps i number of times to create t number of forests.
7. For a new case i, predict Y class using the rule from step 6, calculate votes v, and use the majority votes.

2.4. Evaluation Metric

We have discussed that feature selection methods can be used to efficiently extract essential features from the individual pathology reports, which are used as input features for the classifiers. Therefore, it is worth evaluating how the extracted features represent the critical content within the pathology report based on the performance of the classifiers. The performance of the classifiers was assessed on the test set. The test set for each studied cancer contained 14,590, 3792, and 4159 for breast, colorectal, and prostate cancers, respectively. Preprocessing steps were conducted as was done in the training set. The same features were also selected from the testing set during the prediction process. Based on the prediction result, we computed four indicators (F1-score, precision, recall, and accuracy) that have been successfully applied as evaluation matrices in text classification within the cancer domain [10,27].

- Precision (P) = $\frac{TP}{TP + FP}$

- Recall (R) = $\frac{TP}{TP + FN}$
- $F_1 = 2\frac{PR}{P + R}$
- Acc = $\frac{TP + TN}{TP + TN + FP + FN}$,

where
TP—true positive
TN—true negative
FP—false positive
FN—false negative

Precision estimates the percentage of classified positive cases that are correctly positive, while recall estimates the percentage of correct positive cases in the classification. F1-score is the weighted harmonic mean of precision and recall. Accuracy measures the percentage of correct prediction overall. We computed the performance of the classifiers at a varying feature size to assess the usefulness of feature selection for these classifiers.

3. Result

The distribution of the class labels for the three studied cancers is shown in Table 1. The class distribution for prostate cancer data is approximately balanced compared with others, with 51% malignant cases and 49% benign cases. More than 72% of colorectal cancer cases are benign. However, this was subsampled during the classification stage. A box and histogram plots describing the unprocessed versus processed word and character counts as tabulated in Table 1 are shown in Figures 2 and 3. These figures and the table show reduced word and character counts in the reports after preprocessing. The illustration of the median text length of the pathology reports across the study period is shown in Figure 4. On average, we observed that the text length in malignant breast cancer reports is consistently higher than that of the benign breast cancer reports across the years assessed. Between 2015 and 2019, this figure shows a broader variation in text length between the two classes. We observed an equal or close distribution between the two classes for the colorectal cancer reporting until 2015. Prostate cancer also shows a wider variation in reporting from the year 2015. We have shown an example of a pathology report of a patient with a malignancy diagnosis (Figure 5). This figure shows the transformation from a raw to processed report, ready for analysis. The reports for each cancer type are presented in a word cloud format to illustrate the overall frequent terms in both the unprocessed and processed data (Figure 6). We observed that words synonymous with each cancer type are more pronounced in the processed reports than in the unprocessed reports. Table 2 shows the top-ranked ten features selected by each filter method across the three cancer types.

Table 1. Descriptive statistics of word counts and text length distribution for unprocessed and processed pathology reports across the three cancer type.

Cancer Type	Class	Frequency	Percent	Word Counts		Character Counts	
				Unprocessed Median (Range)	Processed Median (Range)	Unprocessed Median (Range)	Processed Median (Range)
Breast	Benign	29,741	40.0%	88.0 (12.0–916.0)	61.0 (8.0–688.0)	696.5 (106.0–6974.0)	558.0 (78.0–5549.0)
	Malignant	19,791	60.0%	130.0 (7.0–1197.0)	94.0 (7.0–909.0)	1093.0 (102.0–10,483.0)	875.0 (82.0–7122.0)
Colorectal	Benign	13,084	72.0%	83.0 (8.0–968.0)	83.0 (6.0–678.0)	689.0 (102.0–7697.0)	553.0 (88.0–6207.0)
	Malignant	5164	28.0%	86.0 (9.0–619.0)	58.0 (7.0–490.0)	735.0 (109.0–5346.0)	594.0 (81.0–4232.0)
Prostate	Benign	6849	49.0%	80.0 (20.0–510.0)	55.0 (16.0–363.0)	654.0 (155.0–3482.0)	523.0 (127.0–2854.0)
	Malignant	7018	51.0%	92.0 (12.0–475.0)	64.0 (12.0–330.0)	758.0 (131.0–3553.0)	603.0 (114.0–2705.0)

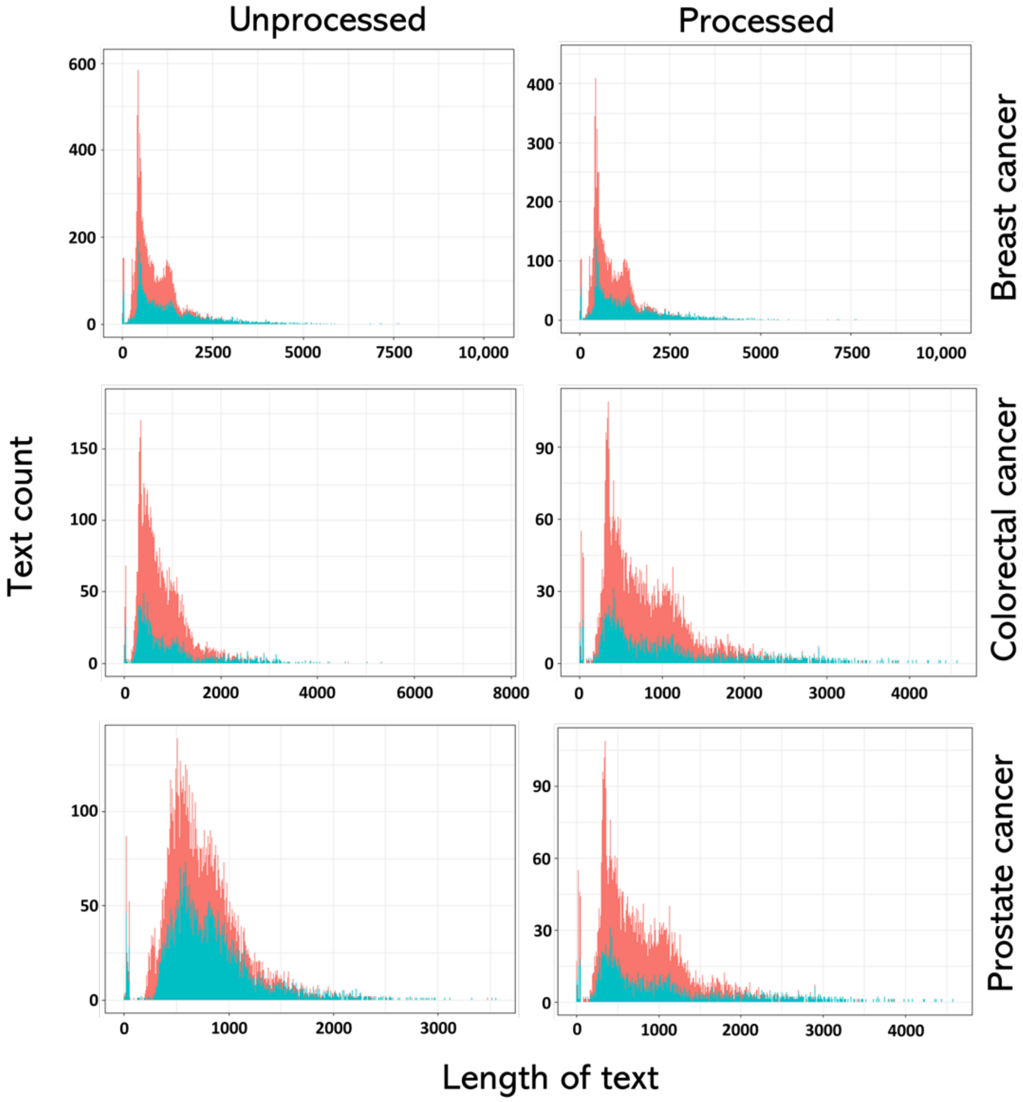


Figure 2. Histogram illustrating frequency counts of pathology report text length for unprocessed and processed benign and malignancy reports for breast (**top**), prostate (**middle**), and colorectal (**bottom**) cancers. The benign class shows higher frequency counts due to large number of cases studied, while the text length of the malignant class is longer than that of the benign class. The text length for each of the cancers was reduced after data preprocessing.

The overall classification results of the RF and SVM (linear and radial) algorithms represented by F1-score measures is shown in Figure 7. A general view of the plots indicates that all the classifiers perform poorly more often than when they have larger feature sizes (Table 3). This shows that an increase in feature size appears to improve classification performance. The classifiers show comparable and higher classification effectiveness in prostate and breast cancers than the colorectal cancer classification. The RF model offers the highest performance across the three cancer types and feature size values, with performance ranging from 90.4–95.2%, 90.0–94.0%, and 93.2–95.3%, in breast, colorectal, and prostate cancer, respectively. The performance of the radial SVM is comparable with RF, especially in prostate cancer classification. On the other hand, linear SVM is the least accurate classifier across all feature sizes and cancer types.

Figure 8 shows the accuracy, recall, and precision (ARP) scores of the classifiers across the three cancer types. For example, in the breast cancer classification, the ARP scores for the RF show a similar trend across all feature sizes than seen with the SVM. Nonetheless, the estimated precision scores are consistently higher than the recall scores across the features sizes. This indicates low misdiagnosis rates than missed diagnosis; that is, a few reports for benign cancer are classified as malignant. A similar pattern is observed with the colorectal cancer prediction, with all the classifiers showing lower recall scores than precision. For both RF and radial SVM in prostate cancer classification, a comparable pattern is seen with the estimated ARP scores, indicating lower missed diagnosis at a large feature size than a misdiagnosis. Overall, the ARP scores also support that RF outperformed SVM across the three cancer datasets.

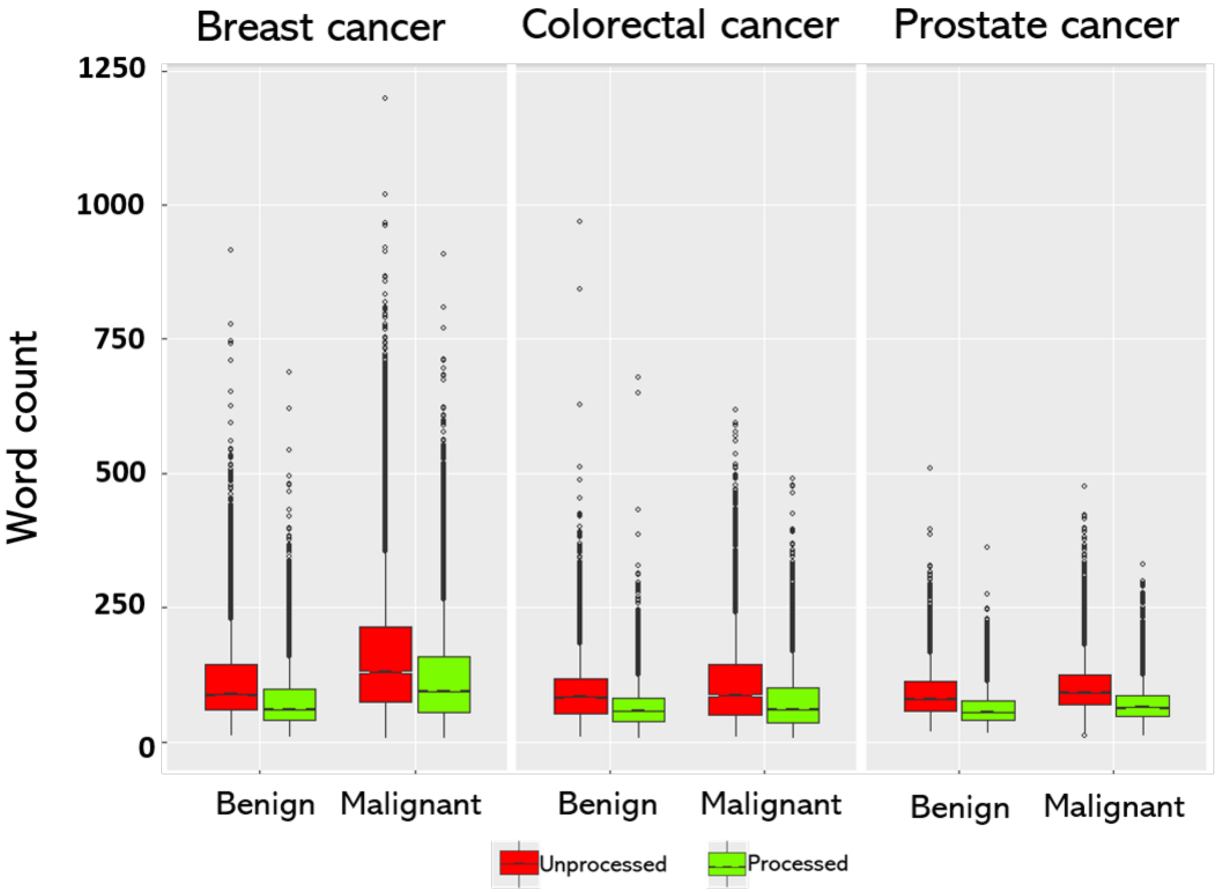


Figure 3. Box plots illustrating the distribution of word counts in the pathology report for the three cancer types. The figure shows reports with higher word counts, which skews the distribution of word counts in the reports. For each cancer target class, the number of word counts reduced after data preprocessing.

As RF outperformed SVM, the classifier was used for further analysis to determine the efficiency of each feature selection method used in this study. Figure 9 shows the F1-scores of the RF model with each feature selection method across feature sizes. For breast cancer, the model achieved a comparable F1-score with the χ^2, IG, and OneR methods than MI across the feature sizes, whereas in colorectal and prostate cancer classification, there is no clear distinction in performance with the four feature selection methods.

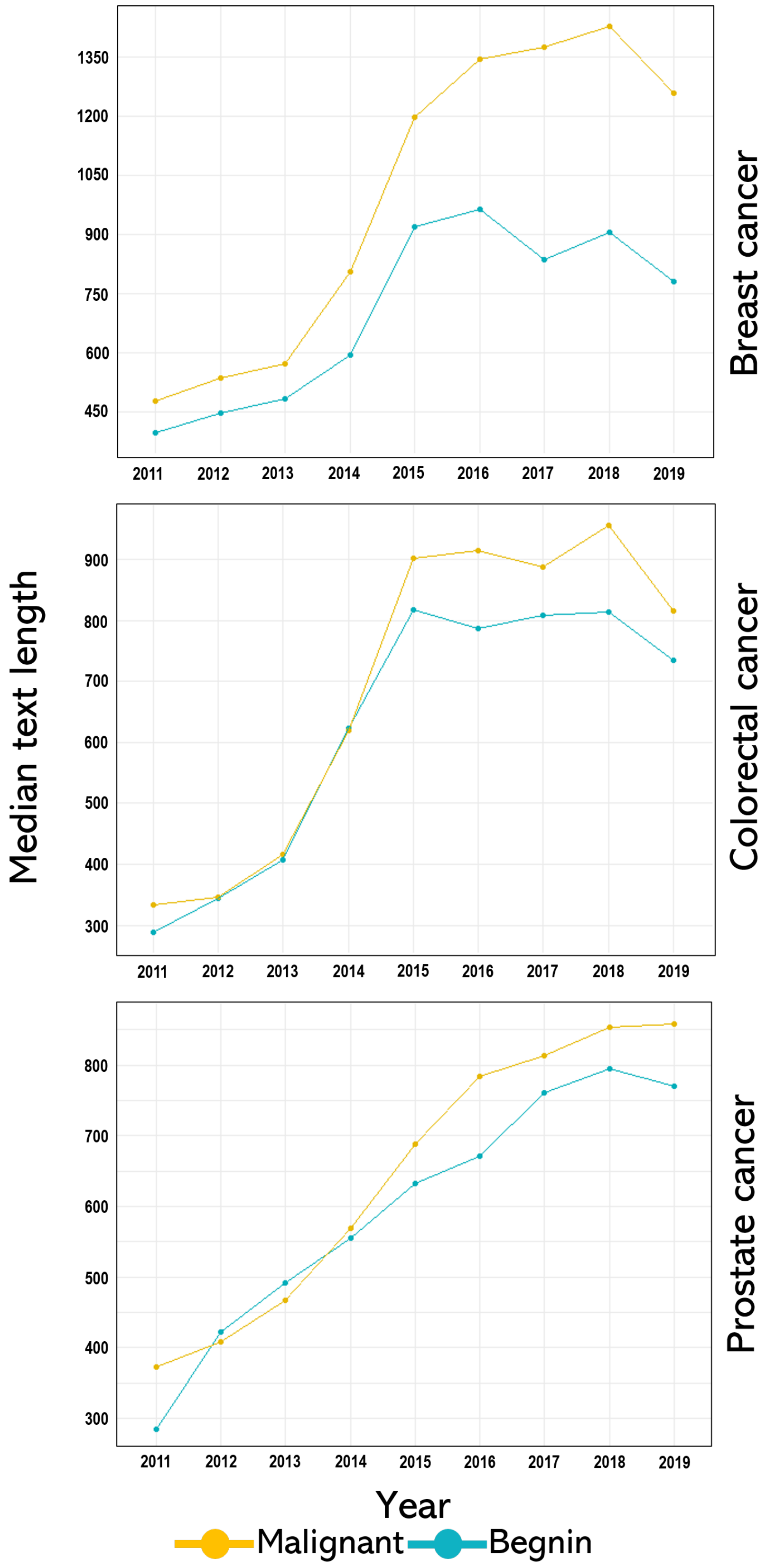
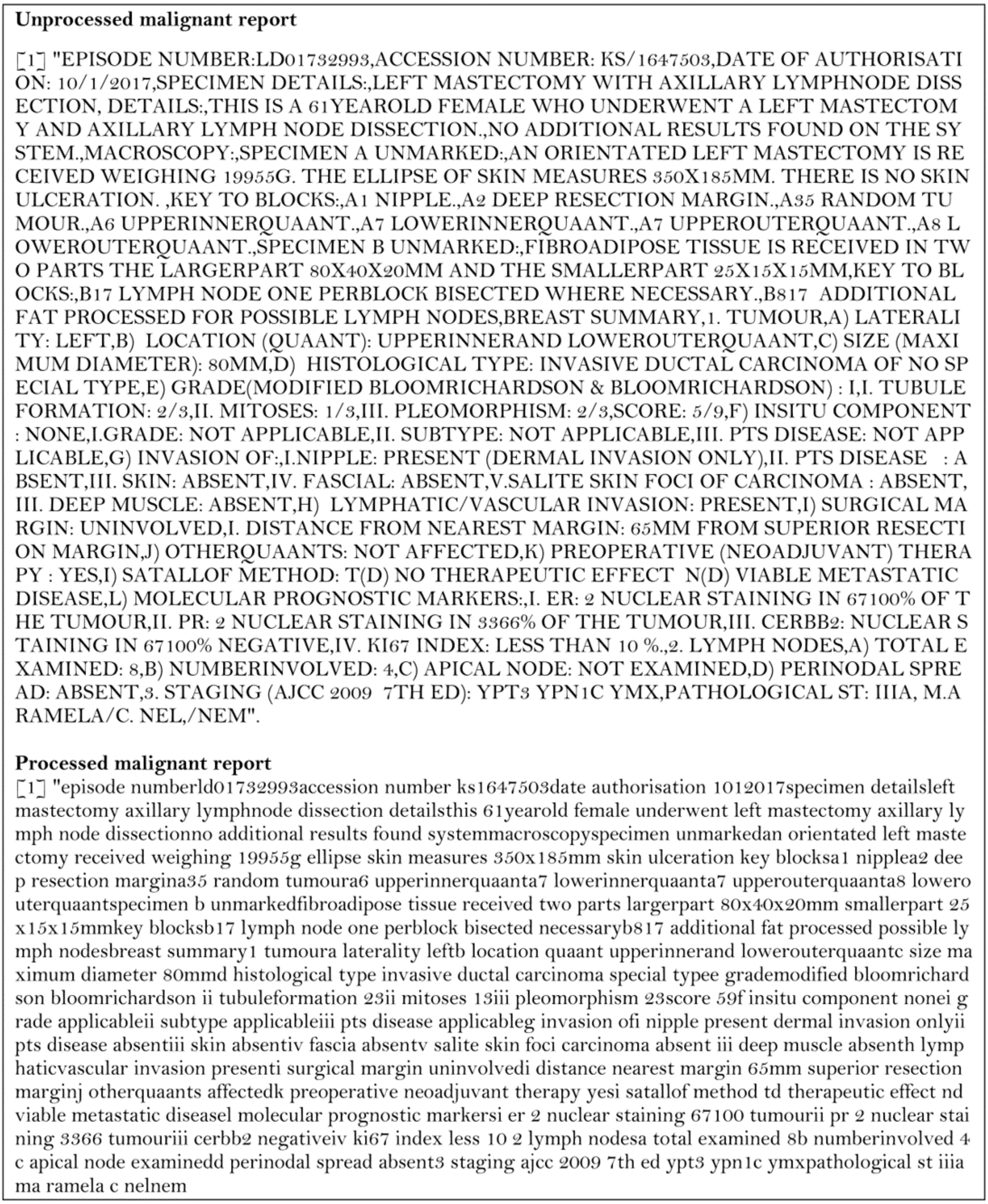


Figure 4. Distribution of median text length across the study period for each cancer diagnosis target class. The figures illustrate the increase in the median pathology reports over the years. Class imbalance is noticeable with the breast cancer reports across the years when compared to prostate and colorectal cancer.



Unprocessed malignant report

[1] "EPISODE NUMBER:LD01732993,ACCESSION NUMBER: KS/1647503,DATE OF AUTHORISATI
ON: 10/1/2017,SPECIMEN DETAILS:,LEFT MASTECTOMY WITH AXILLARY LYMPHNODE DISS
ECTION, DETAILS:,THIS IS A 61YEAROLD FEMALE WHO UNDERWENT A LEFT MASTECTOM
Y AND AXILLARY LYMPH NODE DISSECTION.,NO ADDITIONAL RESULTS FOUND ON THE SY
STEM.,MACROSCOPY:,SPECIMEN A UNMARKED:,AN ORIENTATED LEFT MASTECTOMY IS RE
CEIVED WEIGHING 19955G. THE ELLIPSE OF SKIN MEASURES 350X185MM. THERE IS NO SKIN
ULCERATION. ,KEY TO BLOCKS:,A1 NIPPLE.,A2 DEEP RESECTION MARGIN.,A35 RANDOM TU
MOUR.,A6 UPPERINNERQUAANT.,A7 LOWERINNERQUAANT.,A7 UPPEROUTERQUAANT.,A8 L
OWEROUTERQUAANT.,SPECIMEN B UNMARKED:,FIBROADIPOSE TISSUE IS RECEIVED IN TW
O PARTS THE LARGERPART 80X40X20MM AND THE SMALLERPART 25X15X15MM,KEY TO BL
OCKS:,B17 LYMPH NODE ONE PERBLOCK BISECTED WHERE NECESSARY.,B817 ADDITIONAL
FAT PROCESSED FOR POSSIBLE LYMPH NODES,BREAST SUMMARY,1. TUMOUR,A) LATERALI
TY: LEFT,B) LOCATION (QUAANT): UPPERINNERAND LOWEROUTERQUAANT,C) SIZE (MAXI
MUM DIAMETER): 80MM,D) HISTOLOGICAL TYPE: INVASIVE DUCTAL CARCINOMA OF NO SP
ECIAL TYPE,E) GRADE(MODIFIED BLOOMRICHARDSON & BLOOMRICHARDSON) : I,I. TUBULE
FORMATION: 2/3,II. MITOSES: 1/3,III. PLEOMORPHISM: 2/3,SCORE: 5/9,F) INSITU COMPONENT
: NONE,I.GRADE: NOT APPLICABLE,II. SUBTYPE: NOT APPLICABLE,III. PTS DISEASE: NOT APP
LICABLE,G) INVASION OF:,I.NIPPLE: PRESENT (DERMAL INVASION ONLY),II. PTS DISEASE : A
BSENT,III. SKIN: ABSENT,IV. FASCIAL: ABSENT,V.SALITE SKIN FOCI OF CARCINOMA : ABSENT,
III. DEEP MUSCLE: ABSENT,H) LYMPHATIC/VASCULAR INVASION: PRESENT,I) SURGICAL MA
RGIN: UNINVOLVED,I. DISTANCE FROM NEAREST MARGIN: 65MM FROM SUPERIOR RESECTI
ON MARGIN,J) OTHERQUAANTS: NOT AFFECTED,K) PREOPERATIVE (NEOADJUVANT) THERA
PY : YES,I) SATALLOF METHOD: T(D) NO THERAPEUTIC EFFECT N(D) VIABLE METASTATIC
DISEASE,L) MOLECULAR PROGNOSTIC MARKERS:,I. ER: 2 NUCLEAR STAINING IN 67100% OF T
HE TUMOUR,II. PR: 2 NUCLEAR STAINING IN 3366% OF THE TUMOUR,III. CERBB2: NUCLEAR S
TAINING IN 67100% NEGATIVE,IV. KI67 INDEX: LESS THAN 10 %.,2. LYMPH NODES,A) TOTAL E
XAMINED: 8,B) NUMBERINVOLVED: 4,C) APICAL NODE: NOT EXAMINED,D) PERINODAL SPRE
AD: ABSENT,3. STAGING (AJCC 2009 7TH ED): YPT3 YPN1C YMX,PATHOLOGICAL ST: IIIA, M.A
RAMELA/C. NEL,/NEM".

Processed malignant report

[1] "episode numberld01732993accession number ks1647503date authorisation 1012017specimen detailsleft
mastectomy axillary lymphnode dissection detailsthis 61yearold female underwent left mastectomy axillary ly
mph node dissectionno additional results found systemmacroscopyspecimen unmarkedan orientated left maste
ctomy received weighing 19955g ellipse skin measures 350x185mm skin ulceration key blocksa1 nipplea2 dee
p resection margina35 random tumoura6 upperinnerquaanta7 lowerinnerquaanta7 upperouterquaanta8 lowero
uterquaantspecimen b unmarkedfibroadipose tissue received two parts largerpart 80x40x20mm smallerpart 25
x15x15mmkey blocksb17 lymph node one perblock bisected necessaryb817 additional fat processed possible ly
mph nodesbreast summary1 tumoura laterality leftb location quaant upperinnerand lowerouterquaantc size ma
ximum diameter 80mmd histological type invasive ductal carcinoma special typee grademodified bloomrichard
son bloomrichardson ii tubuleformation 23ii mitoses 13iii pleomorphism 23score 59f insitu component nonei g
rade applicableii subtype applicableiii pts disease applicableg invasion ofi nipple present dermal invasion onlyii
pts disease absentiii skin absentiv fascia absentv salite skin foci carcinoma absent iii deep muscle absenth lymp
haticvascular invasion presenti surgical margin uninvolvedi distance nearest margin 65mm superior resection
marginj otherquaants affectedk preoperative neoadjuvant therapy yesi satallof method td therapeutic effect nd
viable metastatic diseasel molecular prognostic markersi er 2 nuclear staining 67100 tumourii pr 2 nuclear stai
ning 3366 tumouriii cerbb2 negativeiv ki67 index less 10 2 lymph nodesa total examined 8b numberinvolved 4
c apical node examinedd perinodal spread absent3 staging ajcc 2009 7th ed ypt3 ypn1c ymxpathological st iiia
ma ramela c nelnem

Figure 5. A sample of a pathology report for breast cancer malignant neoplasm, illustrating unprocessed free-text report (**top**) and processed free-text report (**bottom**). Processed report shows the absence of some common stop words, punctuation, and special characters.

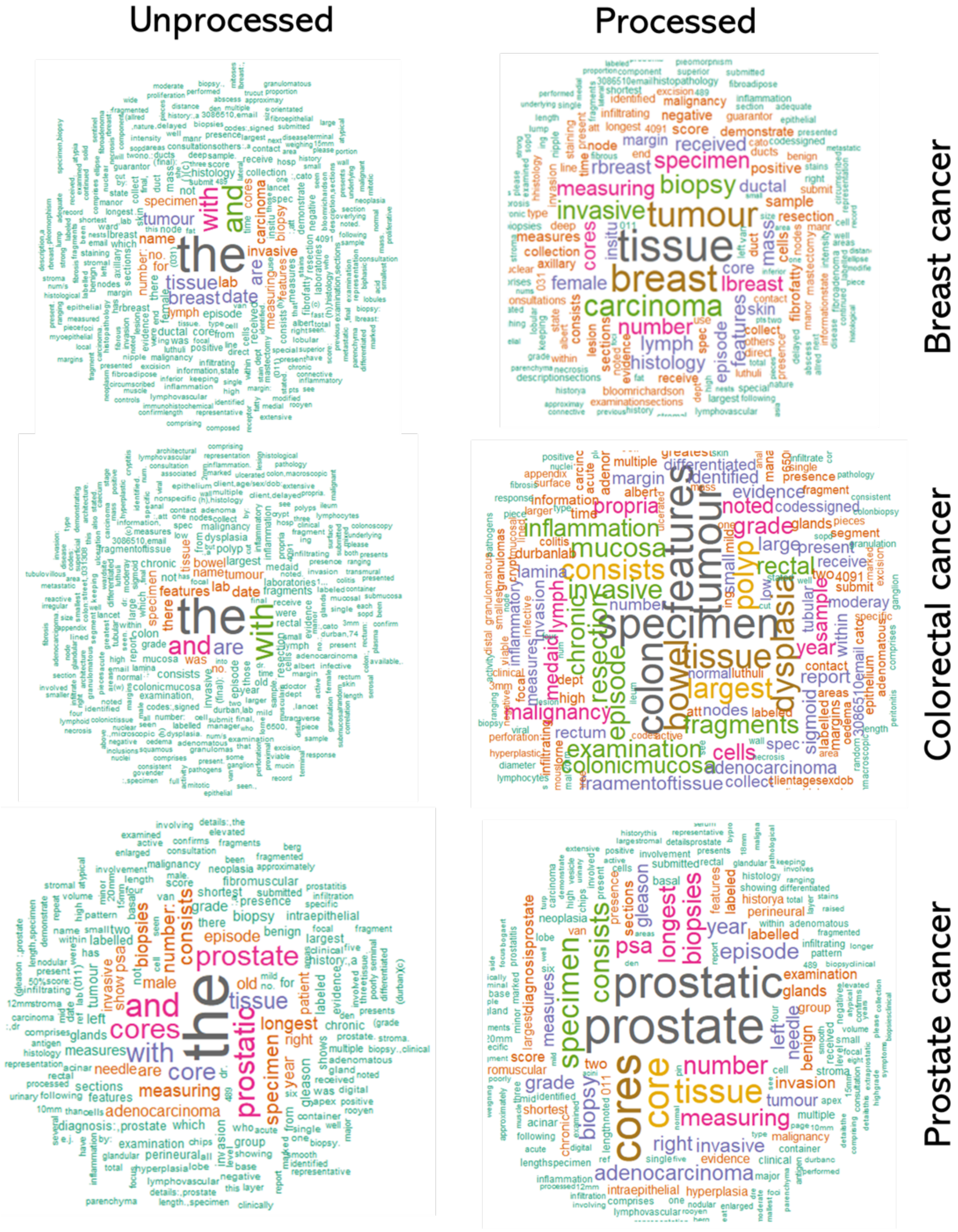


Figure 6. Word cloud of each cancer type for unprocessed and processed pathology reports summarising the most frequently used terms or features in the pathology report. The figures present an overall theme in each cancer type, which is more pronounced in the processed reports.

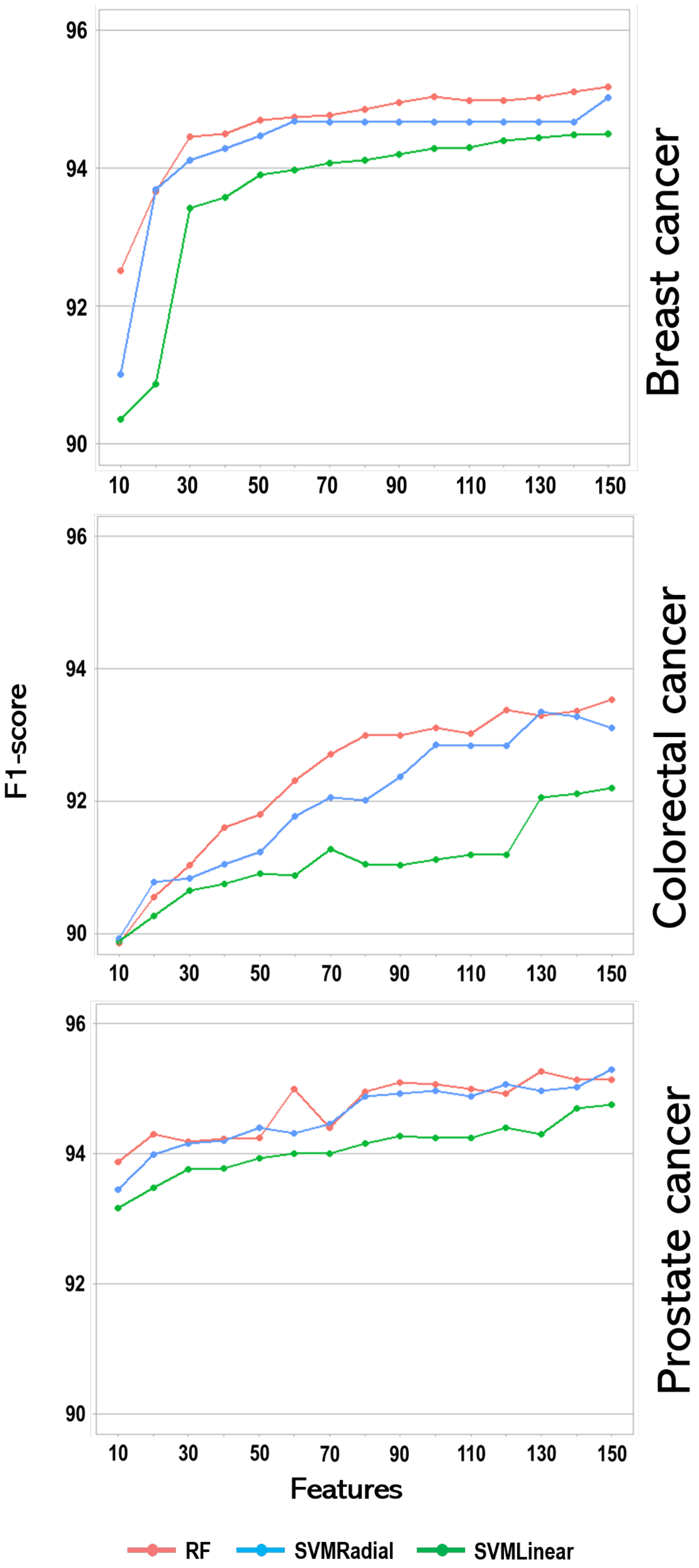


Figure 7. F1-scores of the RF, radial, and linear SVM models across the feature sizes for each cancer type. Different patterns in performance of each model are observed for each cancer type.

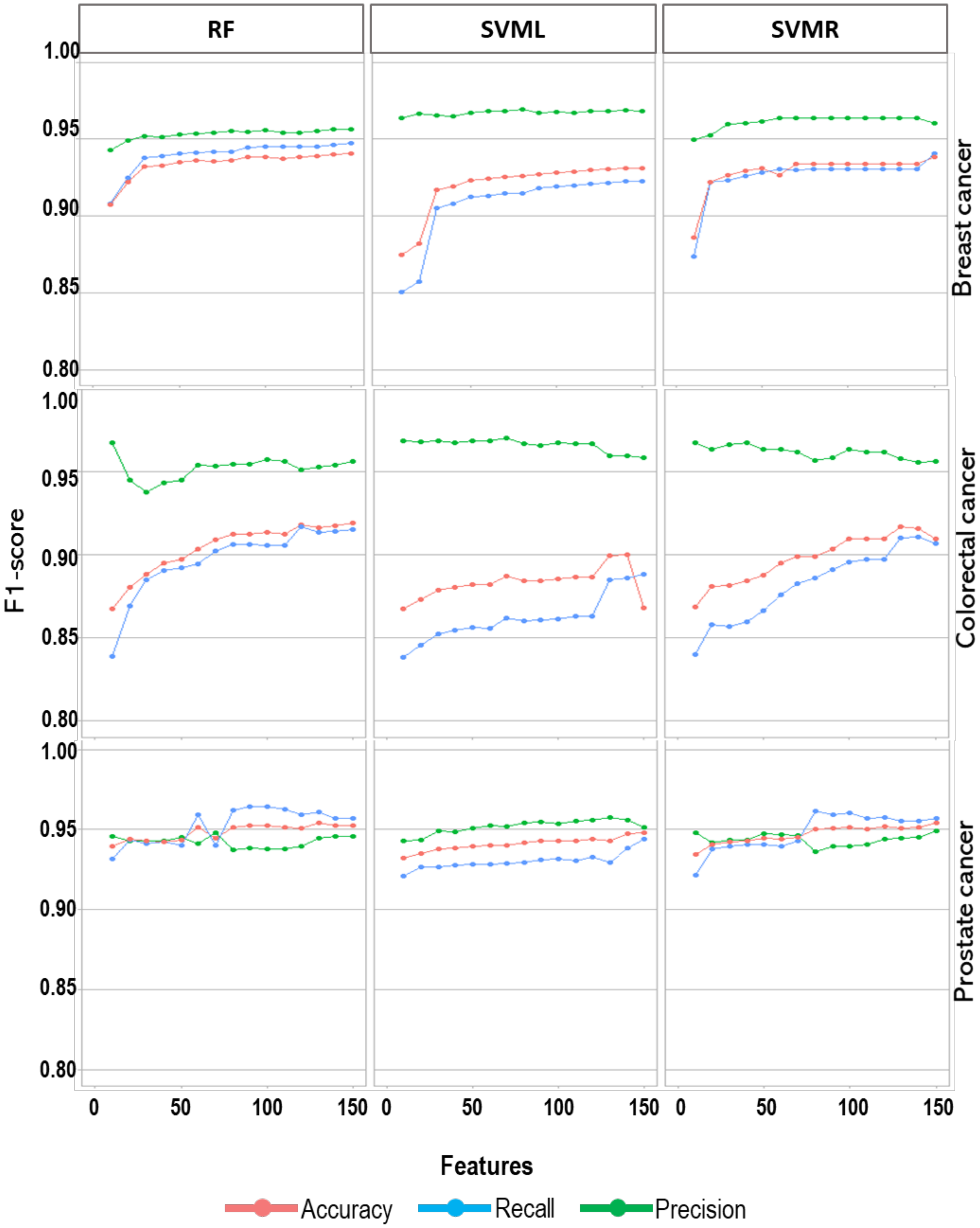


Figure 8. Precision, recall, and accuracy measure of the classifiers over the feature sizes for each cancer type. The figure shows different patterns in each evaluation metric across the feature sizes and for each cancer type.

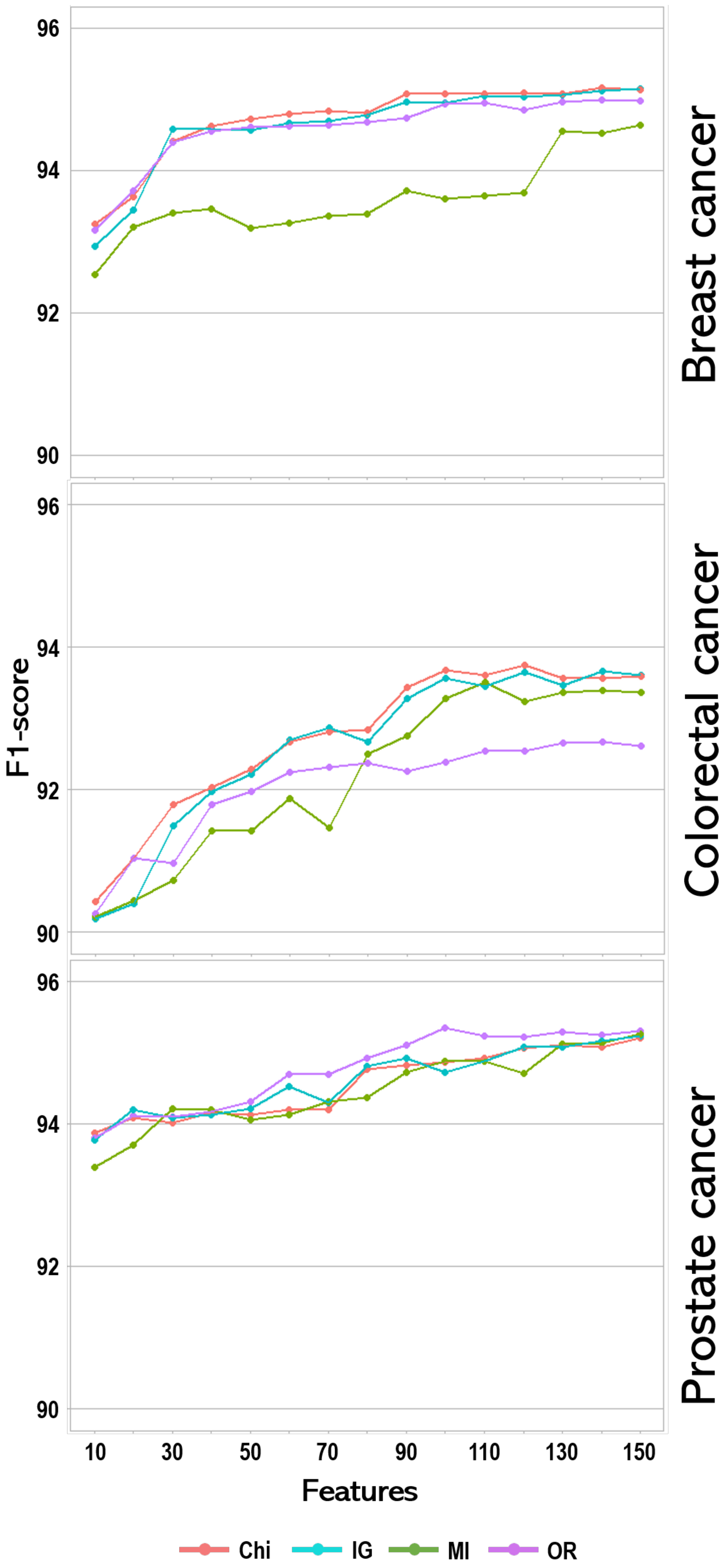


Figure 9. F1-score of RF for the three cancer type, showing patterns in performance with each feature selection method and over feature sizes.

Table 2. Top ten key features or terms identified from each feature selection methods for breast, colorectal, and prostate cancers.

Cancer	Infogain	Chi-Square	OneR	Mutual Info
Breast	carcinoma ki67 tumour score er invasion Age pr infiltrating pleomorphism	carcinoma ki67 tumour score er pr invasion infiltrating invasive pleomorphism	carcinoma ki67 er score tumour pr infiltrating invasion positive her2	ki67 pleomorphism bloomrichardson er invasion lymphovascular nests infiltrating pr score
Colorectal	adenocarcinoma differentiated moderay tumour invasion infiltrating invasive lymphovascular perineural lymph	adenocarcinoma differentiated moderay tumour invasion infiltrating invasive lymphovascular perineural lymph	adenocarcinoma differentiated tumour moderay invasion invasive infiltrating nodes lymph lymphovascular	adenocarcinoma differentiated moderay lymphovascular perineural infiltrating distant differentiation distance ajcc
Prostate	adenocarcinoma gleason perineural invasion score group tumour grade major minor	adenocarcinoma gleason perineural invasion score tumour group grade major hyperplasia	adenocarcinoma gleason perineural invasion score grade tumour group hyperplasia cores	gleason adenocarcinoma perineural score group invasion major minor lymphovascular involved

Table 3. Performance measures of the classifiers at 10 and 150 feature sizes for the three studied cancers.

Features	Cancer Type	Model	F1	Recall	Precision	Accuracy
10	Breast	RF	92.51	90.81	94.28	90.72
		SVMLinear	90.35	85.04	96.39	87.50
		SVMRadial	91.01	87.38	94.96	88.61
	Colorectal	RF	89.86	83.88	96.77	86.74
		SVMLinear	89.89	83.84	96.89	86.76
		SVMRadial	89.93	84.01	96.77	86.84
	Prostate	RF	93.88	93.17	94.60	93.94
		SVMLinear	93.17	92.07	94.31	93.19
		SVMRadial	93.45	92.15	94.79	93.45
150	Breast	RF	95.19	94.73	95.65	94.09
		SVMLinear	94.50	92.26	96.85	93.11
		SVMRadial	95.03	94.05	96.03	93.86
	Colorectal	RF	93.53	91.51	95.65	86.76
		SVMLinear	92.20	88.81	95.86	86.79
		SVMRadial	93.11	90.70	95.65	90.96
	Prostate	RF	95.34	95.69	94.60	95.25
		SVMLinear	94.76	94.40	95.12	94.82
		SVMRadial	95.29	95.55	94.88	95.39

4. Discussion

We developed and validated models for breast, colorectal, and prostate cancers to discriminate benign from malignant using free-text pathology reports. The models achieved good performances. The knowledge extraction phases include pattern matching of the SNOMED ICD-O-3 codes, data subsetting, text processing, feature selection, classification, and evaluation (Figure 1). We provided a thorough description of how each of these steps was conducted.

We acknowledge several studies that show that the quality of free-text pathology reports has potential in enhancing knowledge dissemination [42,43]. Several studies have attempted the auto-annotation of free-text pathology reports with varying degrees of successes. According to Liu et al. [44], extraction of clinical information from the pathology reports poses several difficulties. This study pooled data that spanned nine years, which presented a significant challenge in discriminating benign from malignant due to inconsistencies in the reporting across the studied years. For instance, the median text length has not been consistent over the years for benign and malignancy reporting. In other words, this indicates that at the beginning of the study year, cancer reporting was concise compared to the later study years. This is not unusual because, over the years, there may have been novel biomarkers prescribed for cancer diagnostics, which may necessitate variation in vocabularies and text length of cancer pathology reporting. Be that as it may, new or more essential vocabularies (that have a direct link to the classes) added to the reporting format in recent years might distort the extraction of common terms over the years, thus influencing the overall performance of the classifiers in this study.

Even though we envisaged the influence of analysing pooled data on the classifiers, we observed differences in the text length of the reports (for both malignant and benign cases) within each year. Some reports were overly focused on the patient demographic characteristics or experiments conducted for the diagnosis, other than words describing the actual terms directly related to the pathology outcomes. On the other hand, some reports are so concise that the actual terms that could have helped report interpretations are missing or not comprehensive. In this latter example, it is not clear whether the report is a case of benign or malignancy. Although the first scenario might be preferred, these two case scenarios could be tricky for human and machine classification. These show the imprecise nature of natural language, which could impact the performance of the classifiers [12,43].

Other factors that could impact the model's overall performance are parentheses, notation, and inconsistent usage of important clinical terms (singular terms or a combination of terms) within the reports. For instance, we observed that some of the pathologists report "estrogen receptor-positive" in full, while in some other reports, it was abbreviated as "ER-receptor +ve", "ER-receptor-positive ", "ER-+ve" etc. Moreover, in some cases, punctuation marks were used as a conjuncture in these terms, while in other cases, the clinical terms were enclosed in parenthesis. For example, ER (+ve). This makes the reports somewhat inconsistent. An effort was made to rephrase or standardise those meaningful synonyms that are crucial in discriminating benign from malignant classification. Previous studies noted such inconsistencies in both cancer and non-cancer report classification [11,12,27].

The concept of using filter feature selection methods before text classification has been explored in an email classification study [10]. In this study, we thought it more reasonable to explore four well-known statistical and ML filter selection methods to reduce the excessive number of features generated from reports tokenisation. A combination of features shared across the four feature selection methods did not improve the classifiers' performance compared to using only features selected by chi-square or info gain methods. The selected features are well-known clinical terms that are synonymous with each of the studied cancers, which supports the reliability of our study. Some of these features are visible in the word cloud. An increase in feature size seems to improve classification performance, though in a non-linear fashion across the feature sizes. We recommend the use of a relatively small feature size in model training. In particular, the selected features should be manually assessed to exclude any synonym not corresponding with the study context. Although most cancer case identification and classification studies rely on the algorithm's capacity to discriminate the classes, all the features are used as input to the algorithm without manually assessing the meaningfulness of these features as relating to the task at hand and, hence, may lead to feeding noise to the algorithm.

Generally, malignancy reports (in comparison to benign reports) are expected to embody more elaborate vocabularies, enriched with biomarker terminologies or terms describing the pathology outcome. However, some of the reports of benign cancers were constructed to resemble the malignancy reports. The unique terms for each of these classes are used to profile their characteristics. Ideally, a good classifier should use these profiles to discriminate between the two classes with minimum misclassification cost. Previous studies in cancer case classification have used these profiles to distinguish target classes successfully [8,22,27,45,46]. However, our model showed higher or comparable performances with previous studies. In this study, the RF model consistently performed reasonably well across all values of the feature size and showed an overall performance average of 95.5%, which is comparable to the RF performance in the study by Olago et al. [27]. In addition, the computing time for RF is lower than that of the SVM variants. The radial SVM achieved a better classification performance compared with its linear variant across all the feature sizes. Moreover, the result on the radial SVM is consistently comparable to RF across the three studied cancers. Previous studies have noted the stability and reliability of the classifiers used in this study [10,12,27]. Our study showed comparable performance, especially with the study by Olago et al. [27]. Besides the modelling strategy used in the study by Olago et al. [27], the variation between our study and the study by Olago et al. [27] could be attributed to the use of pooled data across several years and cross several provinces in South Africa.

There is not much variation in the effectiveness of the classifiers across the three datasets. However, we observed more stability and comparability of the classifiers for breast and prostate cancer classification than colorectal cancer classification. Exploring further with the SNOMED ICD-O-3 codes shows that colorectal cancer has about 89 classes compared to the 50 and 20 classes for breast and prostate cancers. This may lead to complexity in the diagnosis and prognosis of colorectal cancer, as seen in this study; hence, the relatively more unsatisfactory performance of the classifiers in colorectal cancer data. The complexity of diagnosis and prognosis of colorectal cancer has also been noted in a study by Wagholikar et al. [46].

Conclusions and Future Studies

We have evaluated a framework to audit the quality of breast, colorectal, and prostate cancer pathology reports archived in the NHLS-CDW between 2011 and 2019 and have developed automated ML algorithms to identify case reports belonging to benign or malignant class. Our modelling strategy appears to generalise the three cancers well and could be adopted in other cancers and beyond cancer studies. Our findings indicate that we can identify the inconsistencies associated with free-text narrative reports in this database. In addition, we can predict the labels of the two classes with F-scores above 90% in all the cancers and feature sizes. We observed the necessity of assessing the tokenised terms to avoid fitting noise in the model and, also, some tokens identified are unique to each cancer studied. We also observed that using a subset of thousands of features generated from tokenisation is enough to build a higher predictive model than using all the features. Using a subset of the features avoids fitting of noise and decreases computational time. Finally, the RF algorithm showed good performance across the three cancers and feature sizes. This study did not comprehensively explore the use of the classifiers (RF and SVM) with other tokenisations, such as bigram, tri-gram, and combinations of the two. Nonetheless, we conducted a preliminary analysis with bigrams using selected feature sizes and observed no improvement compared to uni-gram tokenisation. We also explored other classifiers that are popular in natural language classification, such as naïve Bayes (NB), K-nearest neighbours (KNN), and deep learning (DL) in this study (results not shown); however, RF and the two variants of SVM outperformed NB, KNN, and DL algorithms.

Overall, our study shows that the predictive power of the algorithms used in pathology report classification may be influenced by several factors, including the type of algorithm, data quality, and modelling strategies. In addition, this study supports the use of automated systems such as text mining and machine learning techniques to support human classifiers in labelling a large volume of pathology reports and identifying new cancer cases for incidence reporting. Finally, our findings indicate several interesting directions for future studies. First, we will extend this study by comparing algorithm-based extraction versus manual extraction using tumour topography and morphology sites while considering clustering-based feature selection methods. Correspondingly, for all identified key features critical to the prognosis of studied cancer, we plan to develop a text mining approach that will extract the values associated with these features, irrespective of the degree of inconsistencies observed in the free-text pathology reports. Finally, we would auto-annotate pathology reports targeting the staging and grading using these data, notwithstanding the challenges of extracting such features in heterogeneous free-text reports [44,45].

Author Contributions: Conceptualisation, data analysis, and writing—original manuscript, O.J.A.; review and editing, V.O.; supervision and editing, E.S.; supervision and validation, G.N.; supervision and validation, R.M.J.C.E.; supervision, review and editing, and validation, E.M. All authors have read and agreed to the published version of the manuscript.

Funding: The DELTAS Africa Initiative supported this study. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA). The New Partnership supports it for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Welcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programmer] and the UK government. South African Medical Research CouncilWITS MRC CECRC views expressed in this publication are those of the author(s) and not necessarily those of funders or agencies.

Institutional Review Board Statement: Permission to use the study data was obtained from the NHLS research committee, and ethical clearance was obtained from the Human Research Ethics Committee of the University of the Witwatersrand (M1911131).

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors gratefully acknowledge the NHLS for provision of the study data. We also thank Dr Ikechukwu Achilonu of the School of Molecular and Cell Biology, University of the Witwatersrand for providing the high-performance computer used in this study, which was acquired under the auspices of the NRF/DST SARChI grant number 64788.

Conflicts of Interest: The authors have no conflict of interest to declare.

Abbreviations

The following abbreviations are used in this manuscript:

ARP	accuracy recall precision
AUC	Area under the curve
CDW	Corporate Data Warehouse
DL	deep learning
ICD-O-3	International Classification of Diseases for Oncology version O-3
IG	information gain
KNN	K nearest neighbour
ML	machine learning
NB	naïve Bayes
NHLS	National Health Laboratory Service
OneR	one rule
RF	random forest
ROC	receiver operating characteristics
SVM	support vector machine
TM	text mining

References

1. Statistics South Africa. *Mortality and Causes of Death in South Africa: Findings from Death Notification*; Statistics South Africa: Pretoria, South Africa, 1997.
2. Stefan, D.C. Why is cancer not a priority in South Africa? *S. Afr. Med. J.* **2015**, *105*, 103–104. [CrossRef]
3. Adonis, L.; An, R.; Luiz, J.; Mehrotra, A.; Patel, D.; Basu, D.; Sturm, R. Provincial screening rates for chronic diseases of lifestyle, cancers and HIV in a health-insured population. *S. Afr. Med. J.* **2013**, *103*, 309–312. [CrossRef] [PubMed]
4. Connolly, J.L.; Schnitt, S.J.; Wang, H.H.; Longtine, J.A.; Dvorak, A.; Dvorak, H.F. Role of the Surgical Pathologist in the Diagnosis and Management of the Cancer Patient. In *Holland-Frei Cancer Medicine*, 6th ed.; BC Decker: Hamilton, ON, Canada, 2003.
5. Lankshear, S.; Srigley, J.; McGowan, T.; Yurcan, M.; Sawka, C. Standardized synoptic cancer pathology reports—So what and who cares? A population-based satisfaction survey of 970 pathologists, surgeons, and oncologists. *Arch. Pathol. Lab. Med.* **2013**, *137*, 1599–1602. [CrossRef]
6. Toma, A.; O'Neil, D.; Joffe, M.; Ayeni, O.; Nel, C.; van den Berg, E.; Nayler, S.; Cubasch, H.; Phakathi, B.; Buccimazza, I.; et al. Quality of Histopathological Reporting in Breast Cancer: Results From Four South African Breast Units. *JCO Glob. Oncol.* **2021**, *7*, 72–80. [CrossRef] [PubMed]
7. Service, N.H.L. Annual Report 2011–2017. Available online: http://www.nhls.ac.za/?page=annual_report&id=45 (accessed on 7 August 2018).
8. Jouhet, V.; Defossez, G.; Burgun, A.; Le Beux, P.; Levillain, P.; Ingrand, P.; Claveau, V. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods Inf. Med.* **2012**, *51*, 242.
9. Dube, N.; Girdler-Brown, B.; Tint, K.; Kellett, P. Repeatability of manual coding of cancer reports in the South African National Cancer Registry, 2010. *S. Afr. J. Epidemiol. Infect.* **2013**, *28*, 157–165. [CrossRef]
10. Berry, M.W.; Kogan, J. *Text Mining: Applications and Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
11. Coden, A.; Savova, G.; Sominsky, I.; Tanenblatt, M.; Masanz, J.; Schuler, K.; Cooper, J.; Guan, W.; De Groen, P.C. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J. Biomed. Inform.* **2009**, *42*, 937–949. [CrossRef]
12. Goh, Y.M.; Ubeynarayana, C. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef]
13. Sarkar, S.K.; Roy, S.; Alsentzer, E.; McDermott, M.B.; Falck, F.; Bica, I.; Adams, G.; Pfohl, S.; Hyland, S.L. Machine Learning for Health (ML4H) 2020: Advancing Healthcare for All. 2020. Available online: http://proceedings.mlr.press/v136/sarkar20a.html (accessed on 7 June 2021).
14. Toh, C.; Brody, J.P. Applications of Machine Learning in Healthcare. Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things. 2021. Available online: https://www.intechopen.com/books/smart-manufacturing-when-artificial-intelligence-meets-the-internet-of-things/applications-of-machine-learning-in-healthcare (accessed on 29 June 2021).
15. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142.
16. Hyland, S.L.; Faltys, M.; Hüser, M.; Lyu, X.; Gumbsch, T.; Esteban, C.; Bock, C.; Horn, M.; Moor, M.; Rieck, B.; et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat. Med.* **2020**, *26*, 364–373. [CrossRef]
17. Ali, F.; El-Sappagh, S.; Islam, S.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K.S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inf. Fusion* **2020**, *63*, 208–222. [CrossRef]
18. Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2001**, *2*, 45–66.
19. Nguyen, D.H.; Patrick, J.D. Supervised machine learning and active learning in classification of radiology reports. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 893–901. [CrossRef]
20. Lorenzoni, G.; Bressan, S.; Lanera, C.; Azzolina, D.; Da Dalt, L.; Gregori, D. Analysis of unstructured text-based data using machine learning techniques: The case of pediatric emergency department records in Nicaragua. *Med. Care Res. Rev.* **2021**, *78*, 138–145. [CrossRef]
21. Lewin-Epstein, O.; Baruch, S.; Hadany, L.; Stein, G.Y.; Obolski, U. Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records. *Clin. Infect. Dis.* **2021**, *72*, e848–e855. [CrossRef] [PubMed]
22. Kasthurirathne, S.N.; Dixon, B.E.; Gichoya, J.; Xu, H.; Xia, Y.; Mamlin, B.; Grannis, S.J. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. *J. Biomed. Inform.* **2016**, *60*, 145–152. [CrossRef]
23. Kasthurirathne, S.N.; Dixon, B.E.; Gichoya, J.; Xu, H.; Xia, Y.; Mamlin, B.; Grannis, S.J. Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *J. Biomed. Inform.* **2017**, *69*, 160–176. [CrossRef]
24. Radha, P.; MeenaPreethi, B. Machine learning approaches for disease prediction from radiology and pathology reports. *J. Green Eng.* **2019**, *9*, 149–166.
25. Kalra, S.; Li, L.; Tizhoosh, H.R. Automatic classification of pathology reports using TF-IDF Features. *arXiv* **2019**, arXiv:1903.07406.
26. Nguyen, A.; O'Dwyer, J.; Vu, T.; Webb, P.M.; Johnatty, S.E.; Spurdle, A.B. Generating high-quality data abstractions from scanned clinical records: Text-mining-assisted extraction of endometrial carcinoma pathology features as proof of principle. *BMJ Open* **2020**, *10*, e037740. [CrossRef]

27. Olago, V.; Muchengeti, M.; Singh, E.; Chen, W.C. Identification of Malignancies from Free-Text Histopathology Reports Using a Multi-Model Supervised Machine Learning Approach. *Information* **2020**, *11*, 455. [CrossRef]
28. Wickham, H.; Wickham, M.H. Package 'Stringr' 2019. Available online: https://cran.r-project.org/web/packages/stringr/stringr.pdf (accessed on 15 March 2021).
29. Eler, D.M.; Grosa, D.; Pola, I.; Garcia, R.; Correia, R.; Teixeira, J. Analysis of document pre-processing effects in text and opinion mining. *Information* **2018**, *9*, 100. [CrossRef]
30. Benoit, K.; Watanabe, K.; Wang, H.; Nulty, P.; Obeng, A.; Müller, S.; Matsuo, A. quanteda: An R package for the quantitative analysis of textual data. *J. Open Source Softw.* **2018**, *3*, 774. [CrossRef]
31. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Newton, MA, USA, 2009.
32. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR)* **2002**, *34*, 1–47. [CrossRef]
33. Lunardon, N.; Menardi, G.; Torelli, N. ROSE: A Package for Binary Imbalanced Learning. *R J.* **2014**, *6*, 79–89. [CrossRef]
34. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, R.C.; et al. Package 'caret'. 2020. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf (accessed on 29 June 2021).
35. Parimala, R.; Nallaswamy, R. A study of spam e-mail classification using feature selection package. *Glob. J. Comput. Sci. Technol.* **2011**, *11*, 45–54.
36. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
37. Alhaj, T.A.; Siraj, M.M.; Zainal, A.; Elshoush, H.T.; Elhaj, F. Feature selection using information gain for improved structural-based alert correlation. *PLoS ONE* **2016**, *11*, e0166017. [CrossRef] [PubMed]
38. Kou, G.; Yang, P.; Peng, Y.; Xiao, F.; Chen, Y.; Alsaadi, F.E. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Appl. Soft Comput.* **2020**, *86*, 105836. [CrossRef]
39. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
40. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random Forest and KNN models for the text classification. *Augment. Hum. Res.* **2020**, *5*, 1–16. [CrossRef]
41. Wright, M.N.; Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv* **2015**, arXiv:1508.04409.
42. Parwani, A.V.; Mohanty, S.K.; Becich, M.J. Pathology reporting in the 21st century: The impact of synoptic reports and digital imaging. *Lab. Med.* **2008**, *39*, 582–586. [CrossRef]
43. Ellis, D.; Srigley, J. Does standardised structured reporting contribute to quality in diagnostic pathology? The importance of evidence-based datasets. *Virchows Arch.* **2016**, *468*, 51–59. [CrossRef] [PubMed]
44. Liu, K.; Mitchell, K.J.; Chapman, W.W.; Crowley, R.S. Automating tissue bank annotation from pathology reports–comparison to a gold standard expert annotation set. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2005; Volume 2005, p. 460.
45. Martinez, D.; Li, Y. Information extraction from pathology reports in a hospital setting. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 1877–1882.
46. Wagholikar, K.; Sohn, S.; Wu, S.; Kaggal, V.; Buehler, S.; Greenes, R.; Wu, T.T.; Larson, D.; Liu, H.; Chaudhry, R.; et al. Clinical decision support for colonoscopy surveillance using natural language processing. In Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, La Jolla, CA, USA, 27–28 September 2012; pp. 12–21.