Classification of noisy free-text prostate cancer pathology reports using natural language processing[⋆]

Anjani Dhrangadhariya[1⋆⋆][0000−0003−1691−1338], Sebastian Otálora[1][0000−0003−2125−8476], Manfredo Atzori[1][0000−0001−5397−2063], and Henning Müller[2][0000−0001−6800−9878]

[1] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
anjani.dhrangadhariya@hevs.ch
[2] University of Geneva (UNIGE), Geneva, Switzerland

Abstract. Free-text reporting has been the main approach in clinical pathology practice for decades. Pathology reports are an essential information source to guide the treatment of cancer patients and for cancer registries, which process high volumes of free-text reports annually. Information coding and extraction are usually performed manually and it is an expensive and time-consuming process, since reports vary widely between institutions, usually contain noise and do not have a standard structure. This paper presents strategies based on natural language processing (NLP) models to classify noisy free-text pathology reports of high and low-grade prostate cancer from the open-source repository TCGA (The Cancer Genome Atlas). We used paragraph vectors to encode the reports and compared them with n-grams and TF-IDF representations. The best representation based on distributed bag of words of paragraph vectors obtained an f_1-score of 0.858 and an AUC of 0.854 using a logistic regression classifier. We investigate the classifier's more relevant words in each case using the LIME interpretability tool, confirming the classifiers' usefulness to select relevant diagnostic words. Our results show the feasibility of using paragraph embeddings to represent and classify pathology reports.

Keywords: Pathology Reports · Natural Language Processing · Paragraph Embeddings

1 Introduction

Pathologists examine tissue via a microscope or in a digital image looking for specific cell and gland morphologies that resemble cancer or healthy tissue. After careful examination, they summarize their findings in a free-text report, as shown

[⋆] Supported by University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland.

[⋆⋆] Corresponding author.

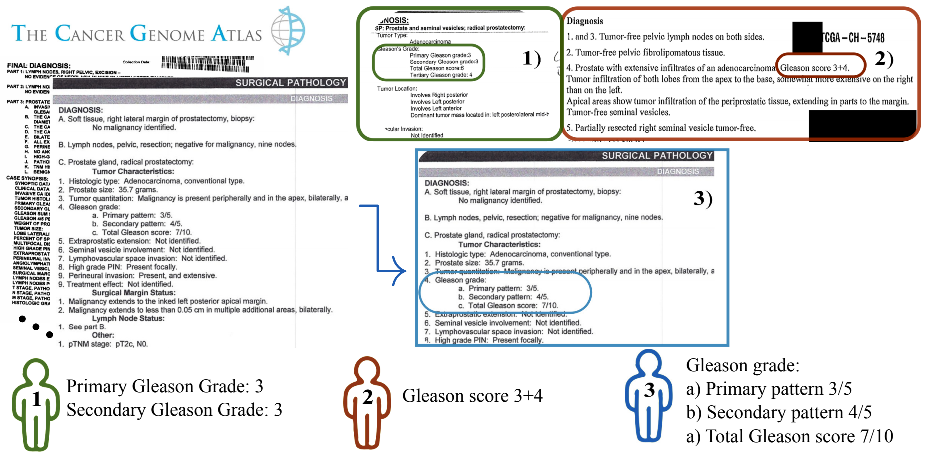
in Figure 1. Pathology reports include a diagnosis or a score in a grading/staging system, despite being an inherently complex and uncertain process [13]. The outcomes are often discussed in tumor boards or given to oncologists and referring clinicians to decide on the best treatment options for the patient.

While free-text reporting has been the main approach in clinical practice for decades (sometimes helped by speech recognition), structured reporting is gaining importance in clinical practice, as it allows to improve quality parameters in diagnostic practice, including timeliness, accuracy, completeness, conformance with current agreed standards, consistency and clarity in communication. In addition, structured approaches can be fundamental (e.g. for cancer registries) for population-level quality monitoring, benchmarking, interventions and benefit analyses in public health management [6].

Automatic analysis and classification of reports can allow to enhance the practice of pathologists. First, it is a possible way to create structured reports from free text ones, in order to standardize previously diagnosed cases for monitoring, benchmarking and benefit analyses in public health management. Second, it can allow to retrieve similar cases in proprietary databases [16], enabling pathologists to navigate repositories of images for clinical decision support and teaching. In such situations, the comparison with visually similar cases is fundamental to reduce the risk of misinterpretations in the diagnosis and provide high quality teaching guidelines. Finally, it can allow faster preparation of multi-center and population-level studies, which require a single agreed international and evidence-based standard to ensure interoperability and comparability [6].

Manually extracting information from free-text pathology reports is an expensive and time-consuming process since they vary widely between institutions, usually contain noise and do not have a standard structure. Still, manual extraction is the most common practice when structured reports are not available, since free-text reports are in most cases extremely noisy and the design and creation of tools that automatically extract information from pathology reports is not straightforward [12, 3, 18]. With the advent of digital pathology and structured reporting, there is an increasing interest in automatic analysis of pathology reports [2, 23, 14, 7, 22].

Natural Language Processing (NLP) tools are used extensively to analyze clinical health records automatically [23, 11, 7]. While there has been an increase in the use of recurrent neural networks [7] and word2vec embeddings [17] to represent the content of reports, the use of deep learning techniques has not yet fully penetrated clinical NLP [21]. Particularly, with the recently proposed distributed representations of words and documents [8], and transformer networks that have outperformed traditional NLP approaches in many NLP tasks and benchmarks [19], the evidence on the applicability to clinical and pathology text remains under-explored.



THE CANCER GENOME ATLAS

FINAL DIAGNOSIS:
PART 1: LYMPH NODES, RIGHT PELVIC, EXCISION –

SURGICAL PATHOLOGY
DIAGNOSIS

DIAGNOSIS:
A. Soft tissue, right lateral margin of prostatectomy, biopsy:
No malignancy identified.

B. Lymph nodes, pelvic, resection; negative for malignancy, nine nodes.

C. Prostate gland, radical prostatectomy:
Tumor Characteristics:
1. Histologic type: Adenocarcinoma, conventional type.
2. Prostate size: 35.7 grams.
3. Tumor quantitation: Malignancy is present peripherally and in the apex, bilaterally, a
4. Gleason grade:
a. Primary pattern: 3/5.
b. Secondary pattern: 4/5.
c. Total Gleason score: 7/10.
5. Extraprostatic extension: Not identified.
6. Seminal vesicle involvement: Not identified.
7. Lymphovascular space invasion: Not identified.
8. High grade PIN: Present focally.
9. Perineural invasion: Present, and extensive.
9. Treatment effect: Not identified.
Surgical Margin Status:
1. Malignancy extends to the inked left posterior apical margin.
2. Malignancy extends to less than 0.05 cm in multiple additional areas, bilaterally.
Lymph Node Status:
1. See part B.
Other:
1. pTNM stage: pT2c, N0.

NOSIS:
SP: Prostate and seminal vesicles; radical prostatectomy:
Tumor Type:
Adenocarcinoma
Gleason's Grade:
Primary Gleason grade:3
Secondary Gleason grade:3
Total Gleason score:6
Tertiary Gleason grade: 4
Tumor Location:
Involves Right posterior
Involves Left posterior
Involves Left anterior
Dominant tumor mass located in: left posterolateral mid-
cular Invasion:
Not Identified

Diagnosis
1. and 3. Tumor-free pelvic lymph nodes on both sides. [redacted] TCGA – CH – 5748
2. Tumor-free pelvic fibrolipomatous tissue.
4. Prostate with extensive infiltrates of an adenocarcinoma Gleason score 3+4.
Tumor infiltration of both lobes from the apex to the base, somewhat more extensive on the right than on the left.
Apical areas show tumor infiltration of the periprostatic tissue, extending in parts to the margin.
Tumor-free seminal vesicles.
5. Partially resected right seminal vesicle tumor-free. [redacted]

SURGICAL PATHOLOGY
DIAGNOSIS
DIAGNOSIS:
A. Soft tissue, right lateral margin of prostatectomy, biopsy:
No malignancy identified.
B. Lymph nodes, pelvic, resection; negative for malignancy, nine nodes.
C. Prostate gland, radical prostatectomy:
Tumor Characteristics:
1. Histologic type: Adenocarcinoma, conventional type.
2. Prostate size: 35.7 grams.
3. ~~Tumor quantitation: Malignancy is present~~ peripherally and in the apex, bilaterally, a
4. Gleason grade:
a. Primary pattern: 3/5.
b. Secondary pattern: 4/5.
c. Total Gleason score: 7/10.
5. ~~Extraprostatic extension: Not identified.~~
6. Seminal vesicle involvement: Not identified.
7. Lymphovascular space invasion: Not identified.
8. High grade PIN: Present focally.

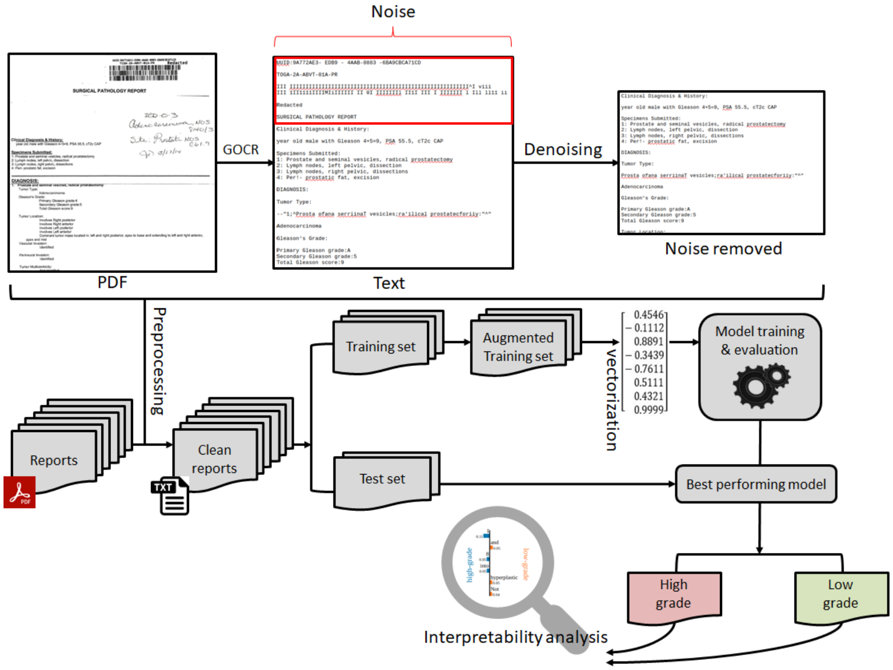
Fig. 1. Snippets of three pathology reports from the TCGA-PRAD repository. The variation in the diagnosis text makes it difficult to manually develop specified rules to extract these important parts of the report automatically.

2 Related Work

There are machine learning and NLP approaches in the literature that classify and extract clinical information from pathology reports [7] automatically. The tasks' performance varies widely and depends mainly on the database's size of the database and how structured the reports are. In the work of Yala et al. [23], each of the sentences in a large dataset of more than 90,000 breast cancer pathology reports is represented with an n-gram. Each report is classified independently into 20 categories, with an average accuracy of 97%. In the work of Qiu et al. [14], the authors use a CNN to automatically extract ICD-O-3 topographic codes from a corpus of breast and lung cancer pathology reports, obtaining a micro-F score of 0.811 and outperforming conventional NLP strategies. Gao et al. [7] used hierarchical attention networks to model free-text pathology reports and extract from them information, including primary tumor sites and histological grades, obtaining macro F-scores of 0.852 and 0.708 respectively in a set of 942 pathology reports. In the work of Alawad et al. [1], the authors used a multitask CNN for classifying histological grade, type, laterality, and primary cancer site in a dataset of 95231 pathology reports, achieving a macro-F measure of 0.766 in the grading task. Similar studies usually lack an in-depth analysis of the classifier's more relevant words, besides reporting the model's quantitative performance. This paper investigates the use of paragraph vectors to represent and classify high and low-grade prostate pathology reports. We encode the reports using distributed representations of sentences and compare it to standard NLP techniques. Our results show that our approach is better than conventional and TF-IDF by 0.23 in AUC, reaching an F-score of 0.858 and an AUC of 0.854. We also analyze the more relevant words qualitatively for classifying the reports in each class, finding them similar to the words that pathologists use the diagnosis of Gleason grading.

3 Methods

This section describes the pathology report corpus used, the pre-processing steps and the classification approach. Figure 2 gives an overview of our approach used to automatically classify pathology reports into high-grade vs. low-grade prostate cancer.



```
UUID:9A772AE3- EDB9 - 4AAB-8883 -6BA9CBCA71CD

TOGA-2A-ABVT-01A-PR

III IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII^I viii
III 1II1111IIIIMIiIIIIIII II 0I IIIIIIII II1I III I IIIIIII 1 I11 1111 i1

Redacted

SURGICAL PATHOLOGY REPORT
Clinical Diagnosis & History:

year old male with Gleason 4+5=9, PSA 55.5, cT2c CAP

Specimens Submitted:
1: Prostate and seminal vesicles, radical prostatectomy
2: Lymph nodes, left pelvic, dissection
3: Lymph nodes, right pelvic, dissections
4: Per!- prostatic fat, excision

DIAGNOSIS:

Tumor Type:

--"1;^Prosta ofana serriinaT vesicles;ra'ilical prostatecforiiy:"^"

Adenocarcinoma

Gleason's Grade:

Primary Gleason grade:A
Secondary Gleason grade:5
Total Gleason score:9
```

```
Clinical Diagnosis & History:

year old male with Gleason 4+5=9, PSA 55.5, cT2c CAP

Specimens Submitted:
1: Prostate and seminal vesicles, radical prostatectomy
2: Lymph nodes, left pelvic, dissection
3: Lymph nodes, right pelvic, dissections
4: Per!- prostatic fat, excision

DIAGNOSIS:

Tumor Type:

Prosta ofana serriinaT vesicles;ra'ilical prostatecforiiy:"^"

Adenocarcinoma

Gleason's Grade:

Primary Gleason grade:A
Secondary Gleason grade:5
Total Gleason score:9

Tumor Location:
```

Fig. 2. Our approach

3.1 Corpus

The approach described in this paper uses publicly available prostate adenocarcinoma clinical pathology reports from The Cancer Genome Atlas (TCGA) PanCancer dataset [3]. The clinical report corpus originally consisted of 494 [4]

[3] 1https://portal.gdc.cancer.gov/projects/TCGA-PRAD

[4] http://www.cbioportal.org/study/clinicalData?id=prad_tcga_pan_can_atlas_2018

reports out of which 404 non-empty reports were selected for further analysis. These reports were varying in length (see Figure 5), unstructured free-text (see Figure 1) scanned copies of the original documents available as Portable Document Format (PDF). An unstructured report in contrast to a structured report is not divided into self-explanatory sections. The corpus documents were manually labelled with two class labels: high-grade (Gleason Score > 7) and low-grade (Gleason Score $< 6,7$) using the diagnosis information from them. This separation has clinically relevant patient stratification [10]. After manual classification, 171 reports were identified as high-grade prostate cancer and 233 reports were identified as low-grade prostate cancer creating a slight class imbalance (refer Figure 5).

3.2 Corpus Preprocessing

Any text corpus requires thorough preprocessing before it can be used for any downstream NLP task. Text preprocessing primarily includes 1) conversion of immutable text documents to machine readable, 2) filtering of useless and noisy parts from the data, and 3) removal of uninformative filler words. The following preprocessing steps were performed before the feature extraction.

1. PDF to text: The PDF documents were converted into editable and searchable text files using GOCR, an open-source optical character reader (OCR) data suite [5].

2. Fixed-pattern noise removal: Next, the most apparent noise elements following a known pattern, like a trail of hyphens (-), pipes (|), asterisks (*), patient identifiers (for e.g., Patient ID: QUID : 70DD94DF - 1301 . 40FC-A52B - 43E2229563E3), sample identifiers (for e.g., TCGA- ZG-A9NI - 91A-PR), and HEX NULL characters (e.g. <0x0C>, <0x0F>) were automatically removed. Fully automatic filtering can miss some noise. Denoising these documents was an important preprocessing step.

3. Stop-word removal: The most frequent, noisy tokens were automatically removed using a set of predefined English language stop-words provided by NLTK (Natural Language ToolKit) [6] along with the corpus-specific stop-words and punctuation, listed in Table 1.

3.3 Data Augmentation

Class imbalance often reduces the classifier performance, so in the present work it was addressed by oversampling through back-translation text augmentation technique. Augmentation and oversampling using back-translation process involves augmenting the minority class by translating a document to a language

[5] `http://jocr.sourceforge.net/`

[6] `https://www.nltk.org/api/nltk.tokenize.html`

Table 1. List of the additional corpus-specific stop-words removed from the pathology reports.

Stop-words			Punctuation				
report	reviewed	surgical	;	#	[]	‘	.
electronically	approved	signed	:	()	?	/	&
pathology	page	redacted	,	_	!	\	”

other than the source language and then translating it back to the source language [20]. Here the documents were translated from the source language English to German and back using the Google translate python package [7]. German was used as a target language for augmentation because it has a high lexical similarity with English (a similarity coefficient of 0.60) thereby adding variability to the oversampled text without altering its meaning [5].

The corpus was split into training and test sets. After splitting, the training set was oversampled to equally learn both the classes during training. Table 2 gives a summary of the corpus used in our experiments.

Table 2. Number of reports per class after train-test split and oversampling.

Partition/Class	Train	Test
High-Grade	186	47
Low-Grade	186	34

3.4 Document Representation

In natural language classification problems, it is important to represent the text documents in machine understandable form. Text representation or vectorization methods convert text documents into fixed-length numeric vectors understood by NLP systems. Two types of numeric representations were extracted from the documents, each one encoding different levels of information. These text representation methods were: I) Count-based vectors, and II) Semantic vectors.

Count vectors: Count vectors encode text as word counts or frequency. Term Frequency - Inverse Document Frequency (TF-IDF) is weighted, sparse, word frequency encoding for numeric text representation. TF-IDF is a multiplication between term frequency (TF) matrix and inverse document frequency (IDF) matrix. Term frequency of a word W is defined as the word count of W for the document D divided by the number of words N in D. IDF of a word W is defined as logarithm of the total number of documents divided by the number of documents containing W. tf-idf increases weight for the meaningful words in the corpus and reduces the weight for filler words like a, an, the, in, if, *etc* [4].

[7] `https://pypi.org/project/googletrans/`

Semantic vectors: Count vectors not only lose the word order and semantic information but also suffer high-dimensionality and sparsity. To take into account semantics of the text, document-level, semantic, dense paragraph vectors were extracted. Paragraph vectors are generated in an unsupervised manner and learn a distributed representation for pieces of text along with distributed representation for the individual words. These vectors learn to associate words with document identifiers rather than with the other words in the context. This work used two kinds of paragraph vectors: 1) a distributed memory model of paragraph vectors (PV-DM) and 2) a distributed bag of words model of paragraph vectors (PV-DBOW) [8]. PV-DM and PV-DBOW vectors were generated for the training documents on fly during the experiments using the gensim functionality [8].

3.5 Document Classification

After feature extraction for each document, L2 normalization for each feature vector was computed. All the labelled documents were used to train and evaluate multiple classifiers (Logistic Regression (LR), Support Vector Machines (SVMs) with linear kernel and K-nearest neighbour (KNN)) in order to separate the reports into high *vs.* low-grade prostate cancer. Grid search was used to explore and identify best performing parameters for these classifiers and feature vector combinations. The model performance was evaluated on an independent held-out test set.

3.6 Experimental Setup

Twelve experiments were conducted each combining the above-mentioned feature vector-classifier combination. Grid search was used to identify the best feature vector, hyperparameters and classifier combination in the training set using ROC AUC (Receiver Operating Curve; Area Under Curve) as a guiding metric. The macro-F1 score, Precision, Recall and ROC AUC measures were used. Random seed for the experiments was set to 42.

Hyperparameter space for count vectors: The tf-idf vectors were extracted and an n-gram space with n ranging from 1 to 10 was explored. Too frequently or too infrequently appearing terms were controlled.

Hyperparameter space for semantic vectors: For the semantic paragraph vectors, vector dimensions of 100, 300, and 500 with window sizes 2, 3 and 5. The paragraph vectors were trained for epochs 20, 30 and 50 along with the above vector dimension and window size combination.

[8] `https://radimrehurek.com/gensim/models/doc2vec.html`

4 Results

Table 3 reports classification results for the best feature vector classifier combination. Paragraph vectors capture better discriminatory information between the classes compared to the count vectors as seen from the ROC AUC scores. PV-DBOW - logistic regression has the best ROC AUC score of 85.4% compared to the other feature vector classifier combinations. Compared to SVM classifier combined with paragraph vectors, LR offers gains in precision by 2.6%, while the recall values for both the classifiers remain identical (79.4%). Paragraph vectors achieve this best ROC AUC score for the denoised, augmented and class-balanced training documents, but training with noisy documents leads to a massive drop in ROC AUC by 11.5%. Training these noisy documents without any oversampling leads to a further drop in ROC AUC by 2.6%. The confusion matrix for both the best performing feature vector - classifier combination is shown in Figure 3. The hyperparamters for the best performing feature vectors are shown in Table 4. For the tf-idf vectors, KNN classifier offers overall better macro precision and recall compared to the other classifiers.

Table 3. The table shows the results for the best classifier-feature vector combination (see section 3.6).

	class "High-grade"			Macro average			
Feature-Classifier	Precision	Recall	F1	Precision	Recall	F1	ROC AUC
tf-idf-LR	0.630	0.500	0.554	0.657	0.644	0.645	0.645
tf-idf-SVM	0.655	0.559	0.603	0.683	0.683	0.675	0.673
tf-idf-KNN	0.739	0.500	0.596	0.723	0.686	0.689	0.686
PV-DBOW-KNN	0.826	0.559	0.667	0.784	0.737	0.743	0.737
PV-DBOW-SVM	0.844	**0.794**	0.818	0.850	0.844	0.847	0.843
PV-DBOW-LR							
- Denoised oversampled	**0.870**	**0.794**	**0.830**	**0.866**	**0.854**	**0.859**	**0.854**
- noisy reports	0.847	0.579	0.688	0.767	0.739	0.732	0.739
- no oversampling	0.735	0.658	0.694	0.716	0.714	0.713	0.713

Table 4. The table shows the parameter settings used for the best feature vector-classifier combination (refer section 3.6).

Features	Classifier	Vector parameters
tf-idf	KNN	n-gram 10, max df 0.7, min df 0.0
PV-DBOW	LR	window size 5, vector dimension 300, epochs 20

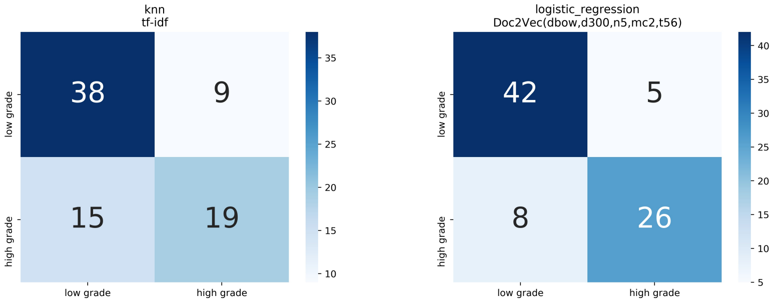


Fig. 3. Confusion matrix for best count vector classifier combination vs. best semantic vector classifier combination

5 Discussion & Analysis

The corpus characteristics: The corpus with pathology reports consisted of highly variable length reports. These are unstructured reports without any explicit demarcation or order for different sections like patient history, diagnosis, conclusion and summary. Figure 5 shows a histogram of the number of reports against report length measured in number of characters. The smallest report had 485 characters while the longest one had 11440. Additionally, except diagnosis section, not all reports were complete and comprehensive lacking one or the other above mentioned sections. The reports originated from heterogeneous sources and were available in the PDF format. Upon conversion from PDF to text format using open-source OCR software, further added to the noise to the already heterogeneous corpus. All these issues made the preprocessing and classification process rather challenging.

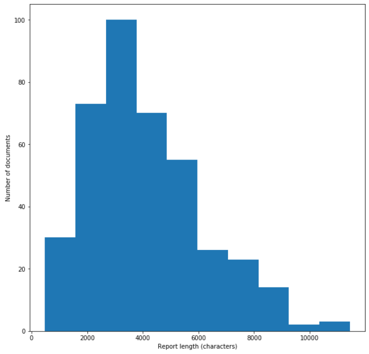


Fig. 4. The graph shows report length for non-empty reports in terms of characters.

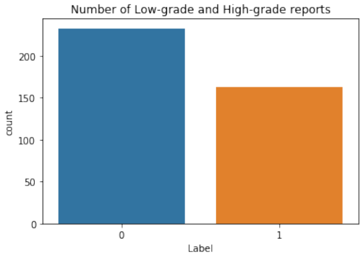
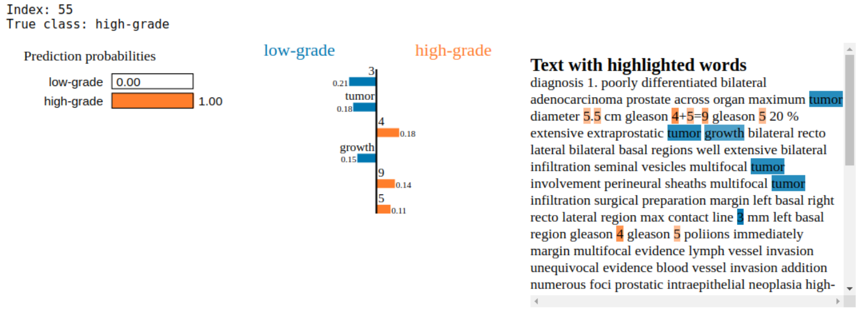


Fig. 5. Histogram showing class imbalance in the corpus. “0” corresponds to class low-grade prostate cancer and “1” correspond to class high-grade prostate cancer.

Interpretation: Paragraph vectors (refer Paragraph 3.4) used in our work are representations generated in unsupervised manner using shallow neural networks. Such representations are not interpretable and are considered as black-box representations. In order to inspect if the best performing paragraph vector representations did capture relevant hints for classifying the documents into high *vs.* low-grade prostate cancer, we used LIME (Local Interpretable Model-Agnostic Explanations) [15]. LIME abstracts the behavior of a black-box system around individual predicted instances in the form of interpretable natural language units (bag of words) i.e. the words. It generates explanations by training a locally-faithful interpretable linear model for individual predicted instances by tweaking its feature values and observing the output. LIME was used to extract six explanations for each class from the individual reports in the test set. Figures 6, 7, 8, and 9 show LIME explanations for the high-grade and low-grade prostate cancer classes on the test instances.

Figure 6 shows a high-grade instance with LIME explanations for both the classes. LIME picks up the numbers "4", "5", and "9" from the text which forms the part of gleason score phrase. Total gleason grading score of "9" is one of the strongest clues to classify the diagnosed prostate cancer into high-grade.



```
Index: 55
True class: high-grade
```

Text with highlighted words

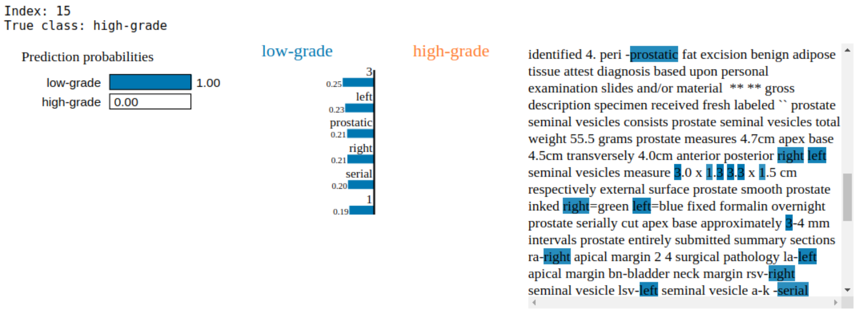
diagnosis 1. poorly differentiated bilateral adenocarcinoma prostate across organ maximum tumor diameter 5.5 cm gleason 4+5=9 gleason 5 20 % extensive extraprostatic tumor growth bilateral recto lateral bilateral basal regions well extensive bilateral infiltration seminal vesicles multifocal tumor involvement perineural sheaths multifocal tumor infiltration surgical preparation margin left basal right recto lateral region max contact line 3 mm left basal region gleason 4 gleason 5 poliions immediately margin multifocal evidence lymph vessel invasion unequivocal evidence blood vessel invasion addition numerous foci prostatic intraepithelial neoplasia high-

Fig. 6. LIME explanation for a high-grade report instance from the test set.

In Figure 7, we present a contradictory example where for a high-grade prostate cancer report, LIME confidently picks up rather irrelevant terms (right, left, prostatic, etc.) for the low-grade class instead.

Figure 8 shows low-grade instance with LIME explanations for the low-grade class. One of the relevant explanation for classification into low-grade are the numbers "3", "4", and "7". These numbers form part of the total gleason score term for low-grade prostate cancer. The model, however, does not pick any word explanations from the class high-grade and also picks other irrelevant explanations like the term "prosectomy" and "excision".

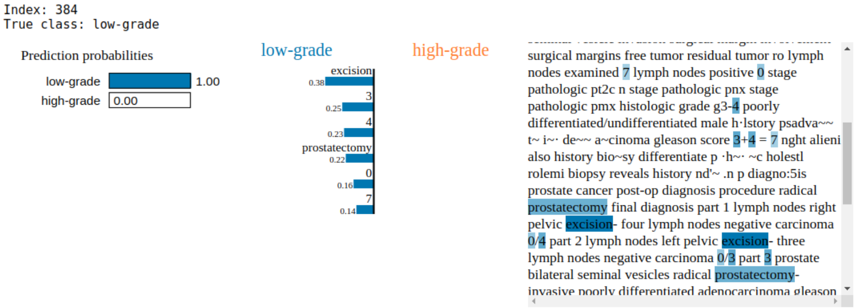
Figure 9 shows a low-grade instance with several strong explanations. For the class low-grade, several relevant clues are picked up; the tumor histologic grade term "g3" and the numbers "3" and "4". It can be noticed that these numbers



```
Index: 15
True class: high-grade
```

identified 4. peri -prostatic fat excision benign adipose tissue attest diagnosis based upon personal examination slides and/or material ** ** gross description specimen received fresh labeled `` prostate seminal vesicles consists prostate seminal vesicles total weight 55.5 grams prostate measures 4.7cm apex base 4.5cm transversely 4.0cm anterior posterior right left seminal vesicles measure 3.0 x 1.3 3.3 x 1.5 cm respectively external surface prostate smooth prostate inked right=green left=blue fixed formalin overnight prostate serially cut apex base approximately 3-4 mm intervals prostate entirely submitted summary sections ra-right apical margin 2 4 surgical pathology la-left apical margin bn-bladder neck margin rsv-right seminal vesicle lsv-left seminal vesicle a-k -serial

Fig. 7. LIME explanation for a high-grade report instance from the test set.

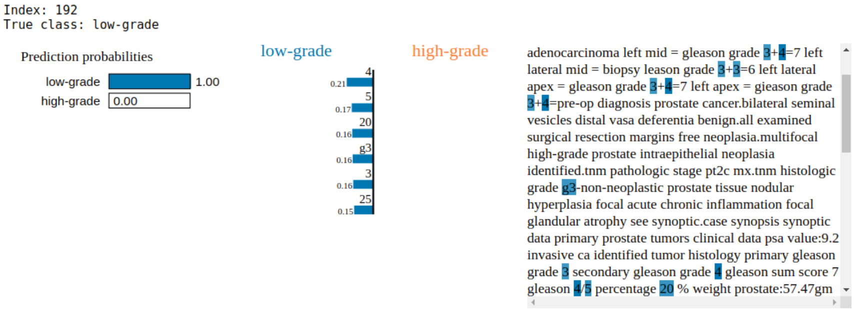


```
Index: 384
True class: low-grade
```

surgical margins free tumor residual tumor ro lymph nodes examined 7 lymph nodes positive 0 stage pathologic pt2c n stage pathologic pnx stage pathologic pmx histologic grade g3-4 poorly differentiated/undifferentiated male h·lstory psadva~~ t~ i~· de~~ a~cinoma gleason score 3+4 = 7 nght alieni also history bio~sy differentiate p ·h~· ~c holestl rolemi biopsy reveals history nd'~ .n p diagno:5is prostate cancer post-op diagnosis procedure radical prostatectomy final diagnosis part 1 lymph nodes right pelvic excision- four lymph nodes negative carcinoma 0/4 part 2 lymph nodes left pelvic excision- three lymph nodes negative carcinoma 0/3 part 3 prostate bilateral seminal vesicles radical prostatectomy-invasive poorly differentiated adenocarcinoma gleason

Fig. 8. LIME explanation for a low-grade report instance from the corpus.

form part of the Gleason grade phrases and are present in the reports as "gleason grade 3+3", "gleason grade 3+4", "primary gleason grade 3", "secondary gleason grade 4" depicting an overall low-grade Gleason score.



```
Index: 192
True class: low-grade
```

adenocarcinoma left mid = gleason grade 3+4=7 left lateral mid = biopsy leason grade 3+3=6 left lateral apex = gleason grade 3+4=7 left apex = gieason grade 3+4=pre-op diagnosis prostate cancer.bilateral seminal vesicles distal vasa deferentia benign.all examined surgical resection margins free neoplasia.multifocal high-grade prostate intraepithelial neoplasia identified.tnm pathologic stage pt2c mx.tnm histologic grade g3-non-neoplastic prostate tissue nodular hyperplasia focal acute chronic inflammation focal glandular atrophy see synoptic.case synopsis synoptic data primary prostate tumors clinical data psa value:9.2 invasive ca identified tumor histology primary gleason grade 3 secondary gleason grade 4 gleason sum score 7 gleason 4/5 percentage 20 % weight prostate:57.47gm

Fig. 9. A strong LIME explanation for a low-grade report instance from the corpus.

From each example LIME explanation demonstrated here, it has to be noted that the paragraph representation model picks several strong informative terms for each class but also picks up many irrelevant words with high confidence and misses out on rather strong diagnosis information terms. LIME explanations warrant further inspection with respect to the hyper-parameters like the number of explanations generated, number of neighbouring samples used to generate explanations, random seed used, and distance metric [9].

6 Conclusions & Future Work

We presented an approach for classification of noisy, heterogeneous pathology reports corpus into high-grade prostate cancer and low-grade prostate cancer using two levels of textual information. Semantic information proved to be more discriminatory between the classes compared to the count information. These pathology reports included unstructured complex information that is spread over long segments of text. Our results and interpretability analysis suggest not only the feasibility, but also reliability of using paragraph vectors to represent and classify prostate pathology reports into high- vs. low-grade prostate cancer.

Each report consists of multiple tumor staging terms, clinical measurements, prostrate tissue anatomy information and their combination with negation terms. We hypothesize that this problem of extracting information from pathology reports might be better suited as an entity recognition problem. Extracting semantically inclined entities could help fine-grained classification of these rather noisy, heterogeneous reports. The noisy and denoised prostrate cancer pathology report corpus, the source code to reproduce our experiments and the python notebook to explore interpretability analysis can be found on Github: `https://github.com/anjani-dhrangadhariya/pathology-report-classification.git`.

7 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825292 ExaMode. Sebastian Otálora thanks Minciencias through the call 756 for PhD studies.

References

1. Alawad, M., Gao, S., Qiu, J.X., Yoon, H.J., Blair Christian, J., Penberthy, L., Mumphrey, B., Wu, X.C., Coyle, L., Tourassi, G.: Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. Journal of the American Medical Informatics Association **27**(1), 89–98 (2020)
2. Baranov, N.S., Nagtegaal, I.D., van Grieken, N.C., Verhoeven, R.H., Voorham, Q.J., Rosman, C., van der Post, R.S.: Synoptic reporting increases quality of upper gastrointestinal cancer pathology reports. Virchows Archiv **475**(2), 255–259 (2019)

3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
4. Dhrangadhariya, A., Jimenez-del Toro, O., Andrearczyk, V., Atzori, M., Müller, H.: Exploiting biomedical literature to mine out a large multimodal dataset of rare cancer studies. In: Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. vol. 11318, p. 113180A. International Society for Optics and Photonics (2020)
5. Eberhard, D.M., Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the world. twenty-third edition. dallas, texas: Sil international (2020), `https://www.ethnologue.com/language/de`
6. Ellis, D., Srigley, J.: Does standardised structured reporting contribute to quality in diagnostic pathology? the importance of evidence-based datasets. Virchows Archiv **468**(1), 51–59 (2016)
7. Gao, S., Young, M.T., Qiu, J.X., Yoon, H.J., Christian, J.B., Fearn, P.A., Tourassi, G.D., Ramanthan, A.: Hierarchical attention networks for information extraction from cancer pathology reports. Journal of the American Medical Informatics Association **25**(3), 321–330 (2018)
8. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196 (2014)
9. Madhyastha, P., Jain, R.: On model stability as a function of random seed. arXiv preprint arXiv:1909.10447 (2019)
10. Narain, V., Bianco Jr, F.J., Grignon, D.J., Sakr, W.A., Pontes, J.E., Wood Jr, D.P.: How accurately does prostate biopsy gleason score predict pathologic findings and disease free survival? The Prostate **49**(3), 185–190 (2001)
11. 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 **11**(9), 455 (2020)
12. Otálora, S., Atzori, M., Khan, A., Jimenez-del Toro, O., Andrearczyk, V., Müller, H.: A systematic comparison of deep learning strategies for weakly supervised gleason grading. In: Medical Imaging 2020: Digital Pathology. vol. 11320, p. 113200L. International Society for Optics and Photonics (2020)
13. Pena, G.P., Andrade-Filho, J.S.: How does a pathologist make a diagnosis? Archives of pathology & laboratory medicine **133**(1), 124–132 (2009)
14. Qiu, J.X., Yoon, H.J., Fearn, P.A., Tourassi, G.D.: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE journal of biomedical and health informatics **22**(1), 244–251 (2017)
15. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
16. Schaer, R., Otálora, S., Jimenez-del Toro, O., Atzori, M., Müller, H.: Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. Journal of pathology informatics **10** (2019)
17. Jimenez-del Toro, O., Otálora, S., Atzori, M., Müller, H.: Deep multimodal case–based retrieval for large histopathology datasets. In: International Workshop on Patch-based Techniques in Medical Imaging. pp. 149–157. Springer (2017)
18. del Toro, O.J., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., Müller, H.: Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In: Medical Imaging

2017: Digital Pathology. vol. 10140, p. 101400O. International Society for Optics and Photonics (2017)

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
20. Wang, Y., Liu, F., Verspoor, K., Baldwin, T.: Evaluating the utility of model configurations and data augmentation on clinical semantic textual similarity. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. pp. 105–111 (2020)
21. Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al.: Deep learning in clinical natural language processing: a methodical review. Journal of the American Medical Informatics Association **27**(3), 457–470 (2020)
22. Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association **25**(10), 1419–1428 (2018)
23. Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J.M., Coopey, S.B., Polubriaginof, F., et al.: Using machine learning to parse breast pathology reports. Breast cancer research and treatment **161**(2), 203–211 (2017)