Classification of cervical biopsy free-text diagnoses through linear-classifier based natural language processing

Jim Wei-Chun Hsu a, Paul Christensen a,b, Yimin Ge a,b, S. Wesley Long a,b,c,*

a Department of Pathology and Genomic Medicine, Houston Methodist Hospital, Houston, Texas, USA
b Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, New York, USA
c Houston Methodist Research Institute and Department of Pathology and Genomic Medicine, Houston Methodist Hospital, Houston, Texas, USA

ABSTRACT

Routine cervical cancer screening has significantly decreased the incidence and mortality of cervical cancer. As selection of proper screening modalities depends on well-validated clinical decision algorithms, retrospective review correlating cytology and HPV test results with cervical biopsy diagnosis is essential for validating and revising these algorithms to changing technologies, demographics, and optimal clinical practices. However, manual categorization of the free-text biopsy diagnosis into discrete categories is extremely laborious due to the overwhelming number of specimens, which may lead to significant error and bias. Advances in machine learning and natural language processing (NLP), particularly over the last decade, have led to significant accomplishments and impressive performance in computer-based classification tasks. In this work, we apply an efficient version of an NLP framework, FastText ™, to an annotated cervical biopsy dataset to create a supervised classifier that can assign accurate biopsy categories to free-text biopsy interpretations with high concordance to manually annotated data (>99.6%). We present cases where the machine-learning classifier disagrees with previous annotations and examine these discrepant cases after referee review by an expert pathologist. We also show that the classifier is robust on an untrained external dataset, achieving a concordance of 97.7%. In conclusion, we demonstrate a useful application of NLP to a real-world pathology classification task and highlight the benefits and limitations of this approach.

Key messages

Linear classifiers using natural language processing algorithms successfully assigned accurate biopsy categories to free-text biopsy interpretations, with a concordance greater than 99.6% on validation set data and 97.7% on untrained external data.

Introduction

In the past 40 years, the incidence and mortality of cervical cancer decreased significantly in the United States due to successful cervical cancer screening programs. Cytology is the mainstay of cervical cancer screening programs around the world; the practice of cytology informs crucial subsequent clinical management and further testing through well-validated clinical decision algorithms.1 Persistent infection with high-risk human papillomavirus (hrHPV) causes cervical cancer and precancerous lesions. Retrospective review of cervical cytology and HPV test results in correlation with follow-up cervical biopsy diagnoses may inform updates to clinical decision algorithms.2,3 Our group previously published several studies of this nature.4-9 As our dataset of cervical biopsy diagnoses increased in size, manual categorization of the free-text biopsy diagnosis into discrete categories proved increasingly challenging.

There is increasing interest in encoding pathology free-text information by extracting clinically relevant information from data-rich pathology reports.10 Some well-tested approaches have focused on using text-processing algorithms to identify and amend report defects in surgical pathology, such as errors in voice recognition11 and data tabulation.12 These approaches encompass a wide variety of text processing strategies, including regular expressions, term detection and tokenizing, support vector machines, statistical approaches, and formal parsing. The comparison of these approaches is difficult due to lack of validation studies, high-quality and annotated datasets with abundant textual diversity, and well-defined consensus approaches for performance evaluation.13

The field of natural language processing (NLP) consists of multiple models that have been used to increase classification accuracy in myriad applications ranging from quality improvement in factories, anomaly detection in national security, and photo recognition on social networks. These NLP models use algorithms to transform free-text data into encoded data, which is more reliable, less error-prone, and less costly to search than free text.14 Two primary approaches exist: rules-based approaches, which...
involves manual encoding of a set of rules designed to process a particular item using the grammar and heuristic rules appropriate for that dataset, and statistical-based approaches, which learn relationships present within existing data (training set) and apply it to previously unseen data (validation set). Rules-based approaches, which comprise word/phrase matching approaches such as regular expressions, SQL queries, and ontology matching, have wide-ranging applicability in several areas of pathology including automating tissue bank annotation, coding and retrieving surgical pathology reports, and extracting Cancer Registry data. However, robustness to natural language variation is a challenge for any rules-based approach. Additionally, depending on the specific rules-based approach used, the ordering of rules can have significant and unforeseen effects on the output, due to the fundamentally iterative nature of most rules-based approaches.

Natural language models using statistical-based approaches, introduced through advances in machine-learning algorithms and statistical inference, were first introduced in the late 1980s, but with exponential increases in computational power and memory, are increasingly practical and useful for real-world datasets. These are further subdivided into 2 dominant approaches. Representation models learn relationships between elements in the input through a series of transformations such as component and cluster analysis before a classification/prediction step. In contrast, deep learning-based models utilize neural networks to extract relevant features from large training sets to feed into multi-layered (feed-forward) perceptron models, which are simplified models of biological neurons. For natural language processing, these 2 approaches translate into linear classifiers and deep classifiers, respectively. Linear classifiers assign labels to words by extracting relevant features from the set of input sentences, which are fed to a shallow classification algorithm (support vector machine; SVM) to obtain a categorical output. Deep classifiers process the input sentences with several layers of feature extraction, and backpropagate the features to deeper layers of the network by convolution, which may offer better classification performance. However, deep classifiers can have significant computational costs, often requiring expensive graphics processing units (GPUs) or other massively parallel computing devices, and often require larger, more diverse training data for acceptable classification performance.

There is increasing focus within the field to develop more efficient linear classifiers with classification performance comparable to deep approaches while using much less computational power that can run even on inexpensive smartphones. A set of popular approaches uses distributed representations of word vectors that are indifferent to word ordering to create a statistical model of multiple word phrases, or “n-gram”, frequency that can be used in downstream classification tasks. Specifically, unlike simple “bag of words” models, multiple-word n-grams enable local word ordering to be captured, as exemplified by the different meanings of the phrases “work to live” and “live to work”. As originally described in the skip-gram model and further refined by biased sampling against frequent words (Word2Vec, Google) and capturing local word ordering (FastText, Facebook), these more efficient approaches have dramatically improved the scalability and usability of natural language processing in various portable applications, including internet search, image recognition, and content tagging. In particular, FastText, a library for efficient text classification and representation learning developed by Facebook (R) Research, builds on the improvements of linear classifiers while making significant improvements in classification accuracy through fast loss approximation (linearly decaying learning rate). This allows it to approximate the performance of conventional DNN classifiers (char-CNN, VDCNN) and comparable linear classifiers (Tagspace) with an order of magnitude of performance improvement.

In this project, we focus on implementing unsupervised, order-invariant word vectors to classify cervical biopsy free-text diagnoses into discrete pre-defined categories, and compare this approach with our previously implemented rules-based regular expression classifier. We also investigate discrepancies between categories generated by NLP and by the rules-based classifier, with referee pathologists determining the “ground-truth” classification of these cases.

### Methods

#### Dataset

The dataset is comprised of 59,751 free-text gynecologic biopsy diagnoses received from BioReference laboratories and interpreted at Houston Methodist Hospital from 2013 to 2018 (Fig. 1). These free-text diagnoses were labelled into discrete diagnosis categories using the rules-based classifier described below. An NLP classifier was created by training on a subset of the labelled data, then used to predict the label on the remaining data. Additionally, a dataset comprising 6672 free-text gynecologic biopsy diagnoses interpreted at Houston Methodist Hospital in 2020 was used for external validation (true holdout). This external dataset comprises interpretations from many different pathologists than the training and validation datasets, and was extracted from a different Laboratory Information System (LIS).

#### Rules-based classifier

Regular expressions were used to construct a rules-based classifier to generate preliminary discrete diagnosis labels. Briefly, the rules-based approach is a Java class file that accepts an input of diagnoses separated by newlines, lowercases all alphanumeric characters, applies a series of string-matching functions in an if/then loop to assign discrete categories (BiopsyClass), and converts the output into a separate text file.

The string-matching functions are sequenced to capture more significant diagnoses (e.g. invasive carcinoma, HSIL) before less significant diagnoses (e.g. LSIL, HPV effect) (Fig. 2). The classifier mapped the free-text biopsy diagnosis into a granular discrete class and a more general class using the BiopsyClass and MappedBiopsyClass vocabularies, respectively.

![Fig. 1.](image)
The BiopsyClass vocabulary is a more granular classifier that distinguishes between Cervical Intraepithelial Neoplasia (CIN) 2, CIN3, adenocarcinoma in situ and carcinoma, and the MappedBiopsyClass vocabulary is a more general classification label.

NLP classifier

Pre-processing

The free-text diagnoses were pre-processed to reduce uninformative variation. Newlines were removed. Punctuation was preserved, but tokenized (a space was placed before and after each punctuation symbol). The text was transformed into lowercase. Multiple training and validation sets were constructed by randomly sampling without replacement a proportion $r$ out of the 59,751 total diagnoses for the training set, where $r$ ranged from 0.01 (1% of dataset used for training) to 0.95 (95%) with a step size of 0.01, for 5 run replicates each, for a total of $n=480$ runs (see Bootstrapping below). For comparative performance evaluation with the rules-based approach and for discrepancy analysis, training was performed using a randomly sampled 80% (47,800 free-text diagnoses) of the randomly shuffled dataset using 20% (11,951 free-text diagnoses) as the validation set.

Model training

For each training data set, we created an NLP classification model using FastText 0.2.0 on a Linux (Ubuntu 18.04) virtual machine. Settings used included epochs of 5–50, a learning rate of 0.1, and $n$-grams of 1–2 (bigrams). Accuracy probabilities and prediction labels on the validation set were generated using the “predict-prob” function. Selection of invariant constants (hyperparameters) was assisted by use of the “autotune” feature in FastText on a subset of hyperparameters (number of epochs, learning rate, and $n$-grams). The BiopsyClass was used as the training and prediction label for the algorithm. To reduce pathologist interpretative variance as a confounding factor (e.g. CIN2 vs CIN3 for HSIL), the more general MappedBiopsyClass was used to compare the performance of both the rules-based and NLP algorithms. Unless otherwise specified, training was performed on 80% (47,800 free-text diagnoses) of the randomly shuffled dataset using 20% (11,951 free-text diagnoses) as the validation set.

Model validation

Bootstrapping

To assess the stability of classifier performance based on choosing discrete training set entries, sampling with replacement (bootstrapping) was performed. The algorithm was run for 5 randomly generated training and validation steps for each proportion $r$, where $r$ ranged from 0.01 to 0.95 with a step size of 0.01, for a total of $n=480$ runs (see Bootstrapping below). For comparative performance evaluation with the rules-based approach and for discrepancy analysis, training was performed using a randomly sampled 80% of the total dataset.

Hyperparameter optimization

FastText offers automatic hyperparameter optimization, allowing for semi-automated tuning of classifier hyperparameters such as learning rate, epoch size, and $n$-gram sizes. The hyperparameter optimization algorithm sequentially runs training and validation classification and iteratively selects the parameters that produce the highest F1-score (combined precision and recall). The default parameters are $n$-grams = 1 and epoch = 5. Learning rate, epoch size (5–50), $n$-gram length (from 1 to 5), array

Table 1

| BiopsyClass      | MappedBiopsyClass | Description                                      |
|------------------|-------------------|--------------------------------------------------|
| Benign           | Benign            | Benign lesion or normal tissue                   |
| CA:AD            | CA                | Carcinoma, adenocarcinoma                        |
| CA:ADIS          | HSIL+             | Adenocarcinoma in situ                           |
| CA:ENDOMETRIAL   | CA:ENDOMETRIAL    | Carcinoma, endometrial                           |
| CA:NOS           | CA                | Carcinoma, not otherwise specified              |
| CA:SC            | CA                | Carcinoma, squamous cell                         |
| Endometrium      | Endometrium       | Benign endometrium                               |
| HSIL:CIN2        | HSIL+             | High-grade squamous intraepithelial lesion (CIN2) |
| HSIL:CIN3        | HSIL+             | High-grade squamous intraepithelial lesion (CIN2) |
| LSIL             | LSIL              | Low-grade squamous intraepithelial lesion (CIN2)  |
| VaIN             | VaIN              | Vaginal squamous intraepithelial neoplasia        |

Fig. 2. Overview of the rules-based classifier. The rules-based classifier does basic pre-processing of interpretation text followed by a string matching approach that eliminates diagnoses based on severity (carcinoma, followed by HSIL, followed by LSIL). An example is provided for a sample text interpretation (inset box).
dimension size (1–100), and bucket size were selected as optimizable hyperparameters, and performed on a randomly sampled 80% training / 20% validation dataset. Optimized classification F1-score was achieved with an epoch size of 20, nGram length of 2, and default learning rate (0.1) with a dimension size of 100 and bucket size of 100 000.

K-fold cross validation

Cross validation using sequentially sampled portions of the data were used to assess classifier stability and check for spurious dependence on the sequence of the underlying data. The biopsy predictions were randomly shuffled, then sequentially partitioned into 5 folds, using 20% of the data as a hold-out set for each iteration. For each iteration, validation was performed on the remaining 80% of the dataset. Precision and recall @ 1 was evaluated with default (nGrams = 1, epoch = 5) and optimized (nGrams = 2, epoch = 20) FastText parameters, as described in the hyperparameter optimization section.

Analysis and discrepancies

Manual review of discrepant (NLP-predicted label differs from rules-based predicted label) cases was performed by 2 pathologists (PAC, YG), who assigned the correct (“ground-truth”) label for each discrepant case after mutual discussion. The results of the discrepancy review were designated as NLP (when only NLP was correct), Rules (when only the regular expression classifier was correct), or Neither (when neither NLP nor Rules were correct). Optimized NLP hyperparameters were used to train the NLP classifier on the training dataset with manually corrected “ground-truth” labels before external validation (see below).

External validation

As described in the discrepancy analysis above, free-text diagnoses were labelled into discrete diagnosis categories by rules-based classifier followed by manual review (to generate a “ground-truth” label) for each discrepant case after mutual discussion. The results of the discrepancy review were designated as NLP (when only NLP was correct), Rules (when only the regular expression classifier was correct), or Neither (when neither NLP nor Rules were correct). Optimized NLP hyperparameters were used to train the NLP classifier on the training dataset with manually corrected “ground-truth” labels before external validation (see below).

Results

NLP classification concordance with regular expressions

Overall classification concordance with rules-based regular expressions was excellent at 99.3%; however, there was considerable heterogeneity within each BiopsyClass, ranging from 100% of diagnosis correctly classified to 57.9% for vulvar intraepithelial neoplasia (VIN) specifically (Fig. 3).

Parameter stability and effect of training parameters

K-fold cross validation

Precision and recall @1 with default (nGrams = 1, epoch = 5) and optimized (nGrams = 2, epoch = 20) FastText parameters remained stable between 0.987–0.989 and 0.994–0.995, respectively (Fig. 4). Run-to-run variation in class distribution remained minimal.

N-gram analysis

In addition to automatic hyperparameter optimization, the performance of the NLP classifier at various N-gram lengths was assessed using manual parameters (nGrams 1 to 5). A nGram setting of 2 noticeably increased precision and recall accuracy, especially for “rare” diagnoses such as carcinoma (CA) and vulvar intraepithelial neoplasia (VIN) with no increases in the training set (Fig. 5A) compared to nGram of 1. With only 1% of sampled training data, the algorithm was able to achieve more than 95% concordance with rules-based regular expressions (Fig. 5B).

Discussion

In this study, we demonstrate that a fast, efficient NLP technology package (FastText) can be successfully trained on a large collection of cervical
biopsy free-text interpretation, and via supervised learning, classify biopsies with greater than 99% accuracy, which remains up to 98% on external untrained data. We also demonstrate the sensitivity of the algorithm to training parameters such as n-gram length and training epochs, with an n-gram length of 2 with concomitantly longer training intervals to optimize classification accuracy. We also demonstrate the informative nature of discrepancy analysis (identifying interpretations where the NLP predicted class does not match the rules-based generated class), and how delegating these cases to a referee decision can identify weaknesses of both models and further refine classification accuracy while only minimally increasing manual workload.

FastText offers a number of optimizable parameters “out of the box” that significantly affects classification performance, including learning rate; number of training epochs; and the use of bi-grams, tri-grams, or...
other n-grams to generate the classifier. Compared to baseline parameters, the choice of bi-grams (nGrams = 2) with longer training epochs (epoch = 20) improved classification accuracy markedly; however, n-grams greater than 2 and longer epoch settings did not significantly improve results, but contributed to longer processing times. A possible explanation could be that more text diversity (larger sample sizes or multi-institutional data) would be needed for the substantially higher token count with higher nGram settings to be adequately trained, or that bi-grams already sufficiently capture training text with limited input diversity (such as pathology reports of only gynaecologic tissue).

Overall, NLP can achieve excellent concordance with a manually iterated rules-based approach to categorization, and in certain instances can achieve superior performance for diagnoses with variation not captured by regular expressions, such as different syntax or phrasing. Of note, the NLP approach is robust to training with a very limited dataset (1% of all data), which achieved greater than 95% concordance with rules-based approaches (Fig. 5B), and is chiefly limited by missclassification of rarer diagnoses. In other words, for a dataset of size 10,000, examining only 100 entries was sufficient to capture the majority of within-dataset variation. Conversely, increasing the proportion of training set sampled beyond 80% did not improve classification performance, as it is likely that the algorithm is overfitted with such a high proportion of training data (Fig. 5A); this may explain why increasing epochs further or similar hyperparameter changes did not measurably affect classification accuracy. Diagnoses with low frequency pose challenges for rules-based approaches because they may not have a rule coded for that scenario; the machine learning-based models are also challenged because the labelled training data may not have representative cases included or labelled correctly, and these diagnostic challenges are reflected by the relatively low confidence that the NLP algorithm assigns. These results could likely be improved by biased or class-imbalanced sampling of less common diagnoses/tokens, as demonstrated by similar applications of NLP in non-pathology fields. For rules-based approaches, which require continuous revision and updating, the NLP algorithm is overfitted, or that bi-grams already sufficiently capture training text with limited input diversity (such as pathology reports of only gynaecologic tissue).

Conclusion

In this study, we demonstrated a successful implementation of NLP-based classifiers into study of a large-scale pathology dataset. These classifiers have significant advantages compared to rules-based classifiers such as classification performance, minimizing design bias in rule selection, and robustness to minor variations in free-text data. Integration of these classifiers into an interpretative workflow requires additional considerations for downstream interpretation tasks. These confidence scores are not available in the rules-based approach. Triaging diagnoses with lower confidence scores for human review, as well as incorporating these diagnoses into the model after retrospective review, can be an effective way to further improve model robustness. For example, re-running the model with corrected discrepant annotations for SMILe (stratified mucin intraepithelial lesion) and vulvar skin with condyloma acuminatum correctly produced the classifications “HSIL” and “VIN” respectively in a subsequent validation run (Fig. 6C), and this process could be iterated to further improve classification as discrepant annotations are manually reviewed. In addition, NLP-based approaches offer the unparalleled advantage of fast retraining by simply re-processing the revised “ground-truth” classifiers generated by referee pathologists to further improve performance for real-world use, which supports an iterative training approach to handle unexpected alterations in incoming data.

Evaluating the suitability of natural language processing technologies for the practice of pathology is multi-factorial. Compared to manual approaches, NLP effectively eliminates inter-observer variability, and compared to manually curated rule-based systems, also reduces “design bias”, a problem inherent to systems such as regular expressions. However, a successful implementation of this technology in a working pathology practice involves more than dataset selection, training, and validation; the context and method that NLP is integrated into daily workflow is also crucial to consider. The actual correctness of so-called “ground-truth” classifiers should also be considered; inter-observer variability among practicing pathologists can be significant for descriptive, complex, or rarely seen diagnoses, and low prediction confidence from NLP-based models can effectively signal the presence of these so-called “problematic cases” by automatically flagging them for post-prediction manual review.
such as LIS compatibility, flagging of problematic or “low confidence” bases, and refinement of the user experience during operation of the classifier, which remain the areas to be explored further.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2022.100123.

References

1. Nayar R, Wilbur DC. The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes. Springer. 2015.
2. Cuzick J, Myers O, Lee J-H, et al. Outcomes in women with cytology showing atypical squamous cells of undetermined significance with vs without human papillomavirus testing. JAMA Oncol 2017;3:1327–1334.
3. Landy R, Castanon A, Hamilton W, et al. Evaluating cytology for the detection of invasive cervical cancer. Cytopathology 2016;27:201–209.
4. Ge Y, Christensen P, Luna E, Armylagos D, Schwartz MR, Mody DR. Performance of Aptima and Cobas HPV testing platforms in detecting high-grade cervical dysplasia and cancer. Cancer Cytopathol 2017;125:652–657.
5. Samimi SA, Mody RR, Goodman S, et al. Do infection patterns of human papillomavirus affect the cytologic detection of high-grade cervical lesions on papanicolaou tests? Arch Pathol Lab Med 2018;142:347–352.
6. Fuller MY, Mody RR, Luna E, et al. Performance of Roche cobas high-risk human papillomavirus (hrHPV) testing in the two most common liquid-based Papanicolaou test platforms. J Am Soc Cytopathol 2018;7:142–148.
7. Goodman S, Mody RR, Coffey D, et al. Negative Pap tests in women with high-grade cervical lesions on follow-up biopsies: contributing factors and role of human papilloma-virus genotyping. Diag Cytopathol 2018;46:239–243.
8. Salazar KL, Zhou HS, Xu J, et al. Multiple human papilloma virus infections and their impact on the development of high-risk cervical lesions. Acta Cytol 2015;59:391–398.
9. Ge Y, Christensen PA, Luna E, et al. Age-specific 3-year cumulative risk of cervical cancer and high-grade dysplasia on biopsy in 9434 women who underwent HPV cytology cotesting. Cancer Cytopathol 2019;127:757–764.
10. Burger G, Aby-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016;69:949–955.