EXAMINING INTEROBSERVER VARIABILITY IN HISTOPATHOLOGY REPORTS FOR LUNG ADENOCARCINOMA

Sardar Elias[1], Manit Zaveri[1], and Jay Shah[1]

[1]University of Waterloo

December 19, 2018

Abstract

The difficulty in visually interpreting complex lung tumor patterns in histopathology images results in interobserver variability among pathologists, leading to diagnosis and treatment delays. This can be avoided by implementing a digital system that mediates disagreements by generating informative captions for pathology images based on feature extraction and pattern recognition. Extracting key diagnostic terms from historical pathology reports is the preliminary step to developing this system. This paper proposes an algorithm that generates relevant diagnostic terms from a pathology report based on a short list of known keywords using cosine similarity.

INTRODUCTION

Lung carcinoma is the leading cause of cancer-related death worldwide. In most cases, the chances of survival for lung cancer patients are dramatically reduced due to late diagnoses and limited treatment interventions. Furthermore, complications during the diagnosis of a lung cancer patient add to the overall treatment inefficiency. One of these complications arise due to conflicting diagnostic opinions among pathologists upon reviewing the lung pathology reports. This is known as interobserver variability. The intended result of the ongoing research on reducing interobserver variability is a pattern recognition system that analyzes a lung tumor histopathology image and generates a caption consisting of key diagnostic terms to describe the tumor, e.g., the histologic subtype, tumor stage and other relevant diagnostic information. This system can serve as a mediator and help pathologists with their diagnosis when visual pattern recognition of complex tumors is difficult. The design stages of this system are outlined below:

1. Extract keywords from existing pathology reports
2. Extract features from corresponding pathology images
3. Assign the keywords to the image features and create a database
4. Pathology image of a new patient will be matched feature-to-feature with those in the database, generating the assigned keywords based on feature similarities.

The purpose of this paper is to study the archive of existing lung cancer histopathology reports of previous patients and propose an algorithm design that extracts relevant keywords from the reports. This will complete Stage 1 of the process and provide the preliminary tools required in Stage 2.

BACKGROUND AND RELATED WORK

Following a biopsy or surgery of lung tumor, tissue samples are sent for laboratory analysis to determine the prognostic features, which may take up to three days. Furthermore, the process may entail sending the specimen to other specialized laboratories for further study to help determine appropriate treatment measures, resulting in additional delays by days or even weeks. Tumors vary greatly from patient to patient and are usually difficult to diagnose; their complex nature requires the specialty of multiple medical professionals. Several pathologists review the lab results before the final pathology report is sent to an oncologist who can evaluate and determine a treatment plan. An accurate and verified pathology report provides all the necessary details on the patient, tumor, prognosis and current state, and is based on pathologists' visual interpretation of the tumor pattern, as seen on the histopathology image. As lung carcinoma consists of many subtypes and stages, these patterns can be very complicated, even to the trained eye of a pathologist. If interobserver variability among pathologists exists, the treatment process is severely delayed.

Lung carcinomas are mainly divided into two groups; Non-Small Cell Lung Carcinoma (NSCLC) and Small-Cell Lung Carcinoma (SCLC), the former consisting primarily of Adenocarcinoma (over 50% of cases) and Squamous Cell Carcinoma (over 30% of cases). These classes are further categorized into histologic subtypes based on their molecular profiles. A comprehensive classification of lung carcinoma subtypes and their molecular pathology has been outlined in the *2015 WHO Classification* [1]. For instance, the primary histologic subtypes of lung adenocarcinoma include lepidic, acinar, papillary, micropapillary, and solid patterns. Most lung adenocarcinomas demonstrate a mixture of different histologic patterns/subtypes; however, each subtype has its prognostic significance, with lepidic pattern harboring the best course of prognosis while micropapillary and solid patterns demonstrating a more aggressive behavior. Depending on the size, dominance and aggressiveness of each pattern, the tumor can be classified as invasive or non-invasive. Below are some histology image samples of the patterns [2]:

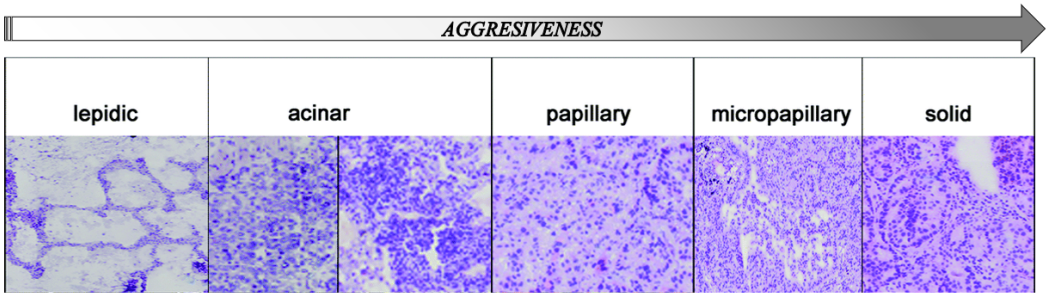

Figure 1: Histologic Subtypes of Lung Adenocarcinoma based on pattern recognition. Different subtypes have prognostic significance. Increasing gradient of aggressive tumor behavior, going from Lepidic to Solid patterns.

Evidently, visually recognizing the patterns and distinguishing between different types of tumor based on their molecular pathology structures is an intricate task to perform. In a study conducted by Gelfman, Nelson, et al., each of five pathologists were asked to give two independent readings on 50 different specimens, which were a mixture of well and poorly differentiated adenocarcinoma. When the readings for each specimen were compared, the results showed significant disagreement among the pathologists; 2% to 5% variability for well differentiated adenocarcinomas, 23% to 25% for undifferentiated large cell and small cell carcinomas and 40% to 42% for poorly differentiated adenocarcinomas. Moreover, the aggregate disagreement between two readings of the same sample by the same pathologist was up to 20% for poorly differentiated adenocarcinoma, proving that variability not only exists between different pathologists but also between readings of the same pathologist when the image is too complex [3].

METHODS

In our proposed method to extract relevant diagnostic terms from lung cancer histopathology reports, we developed a Python script that takes advantage of mathematical algorithms and text properties to find and rank certain words in a document as desired. The first step of the process was to manually "clean" the archive of histopathology reports. The KIMIA lab database of medical documents consists of thousands of diagnostic reports of previous patients, of which 995 reports pertain to the histopathology of lung cancer. Most of these documents are scanned copies of diagnostic results with sections of redacted information, handwritten notes, illegible texts and characters that were subject to incorrect automatic text transcription. Therefore, it was necessary to ensure that all the documents were free of formatting and spelling errors, and that every diagnostic word from a report was transcribed to a text format as accurately as possible. Once the cleanup process was complete, the pathology reports were downloaded as a collection of ".txt" files.

The next step involves creating two separate text files – a "PROFILE" file containing a list of important and relevant diagnostic key terms we are looking for, such as all the cancer subtypes, patterns, cancer stages, tumor location, etc., and a "COMMON WORDS" file that contain a list of 'stopwords' (and, the, this, when, etc.,) as well as common terms that are of no significance, not relevant to our query and do not provide us with any information regarding the tumor. The list of common words need not be exhaustive; however, some degree of completeness is expected. Longer the list of common words, more precise the outcome. The 'stopwords' can be downloaded as a list from any online source, and additional words may be added at discretion. It is important that no word from the "PROFILE" list appears on the "COMMON WORDS" list. The Python code requires that both files are in the same folder as the downloaded reports.

Methodology of the Python Algorithm

Each report's file name is specified at the end of the script, along with the "PROFILE" and "COMMON WORDS" list files. It is this main function that is called to run the mentioned file and generate keywords from the document based on the code's algorithm.

main('TCGA-05-4382.952c0f32-1472-49e1-8334-b0f1de4ac921.txt','profile.txt','common.txt')

The code imports data from the specified document in the directory, along with all other data in the corpus. Basic operations are performed to eliminate unnecessary text complications and maintain data similarity. For example, string_data_clean.lower() converts all uppercase data to lowercase, str_data = str_data.replace function is used to replace all punctuations to periods ("."), str_data.split is used to split sentences into new lines and separate words, etc. The first section of the script generates a "bag of words" from the imported file to be used in upcoming functions.

At this point, the textual data is converted to vectors in a vector space model. This is an important step because in order to calculate a numerical value for text, such as frequency scores, cosine values, etc., the data needs to be in the form of a matrix to allow for computational interpretation. The resulting vector space model represents terms, sentences and documents in the form of vectors (Figure 2).

The section of the script under the line of code, get_word_frequency(string_data_clean, test_words, common_words),processes the parsed data, calculates the vector magnitude and generates the term frequency for every word in the document.

Cosine Similarity calculates the cosine of the angle between two document vectors. This tells us the orientation of the two vectors and hence generates a metric that provides information on how similar the two documents are. To build the cosine similarity equation, the following dot product formula is implemented into the Python script:

$$\cos\theta = \frac{\vec{P} \cdot \vec{Q}}{||\vec{P}||||\vec{Q}||}$$

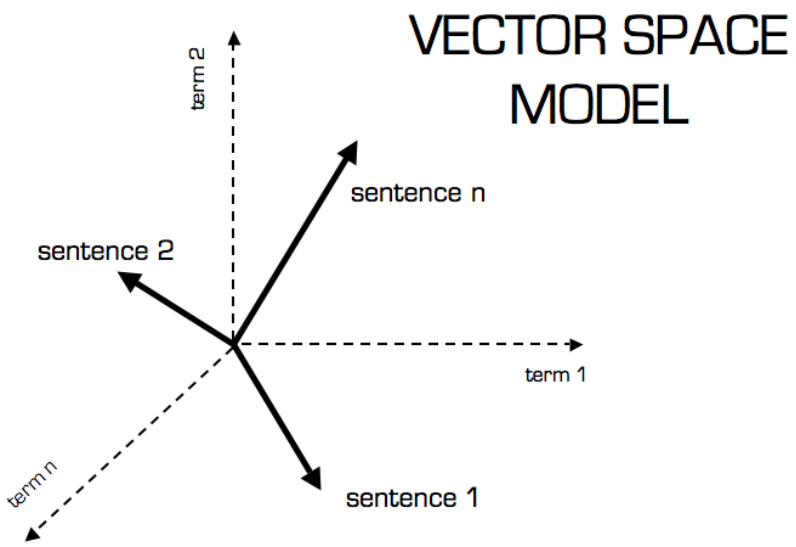

Figure 2: Words, sentences and documents converted to individual vectors and mapped on to a Vector Space Model [4]

where P and Q are documents in the vector space. The count_denom_p and count_denom_q lines under the profile_calculation section of the script makes use of the cosine similarity equation to find the magnitudes and Euclidean distances among all the documents in the vector space of the corpus. The term frequency for every word in the "PROFILE" list is then weighed against the cosine similarity values to generate a list of terms extracted from the specified document that are closest to the "PROFILE" words. A score closer to '1' indicates high similarity, whereas scores of '0' indicate no similarity.

EXPERIMENTAL RESULTS

# of Terms	Key "PROFILE" terms
23	unifocal, acinar, mucigenous, well, poorly, moderately, differentiated, adenocarcinoma, non-mucigenous, pulmonary, squamous, papillary, micropapillary, lepidic, solid, sarcoma, lymphoma, sclerosis, invasive, non-invasive, metastasis, t1, t2, t3, t4, nx, n0, n1, n2, n3, mx, m0, m1, pt0, pto, pt1, pt2, pt3, pt4, pnx, pn0, pn1, pmx, pm0, pm1

Figure 3: Table 1: key test-words on "PROFILE" list

To test our code, we randomly chose 5 reports from the pathology reports archive to run on Python. Table

Report Name	# of Words in Report	# of Output Terms	Output Terms
"TCGA-21-1070…txt"	817	23	grade, differentiated, iii, upper, lobe, lung, chest, wall, -poorly, squamous, cell, carcinoma, 9cm, invading, overlying, rlbs, pt3j, vascular, invasion, present, poorly, histologic, classification, non-keratinizing
"TCGA-44-8120…txt"	435	24	lung, upper, lobectomy, invasive, differentiated, adenocarcinoma, histologic, grade, moderately, microscopic, description, distant, pm, pmx, regional, lymph, nodes, pn, sampled, mediastinal, negative, metastatic, carcinoma, metastasis
"TCGA-39-5021…txt"	968	10	8, lung, upper, lobe, lobectomy, squamous, cell, carcinoma, differentiated, poorly
"TCGA-66-2765…txt"	506	83	immediate, intraoperative, evaluation, lymph, nodes, massive, pigment, storage, inclusion, tiny, anisotropic, particles, fibrosis, differing, degrees, hyaline, thickening, 1, primary, diagnosis/diagnoses, resected, lung, tissue, shows, peripheral, focus, differentiated, predominantly, large-cell, partly, clear-cell, nonkeratinizing, squamous, cell, carcinoma, formation, large, central, lacuna, region, pleura, 2, cm, area, whitish, discoloration, unclear, demarcation, depression-like, puckering, retraction, solid, pale, brownish-white, tumor, max, poorly, remark/addendum, limited, carcinomatous, epithelium, questionable, whether, arose, result, resorption, moderately, assumption, view, size, focal, evidence, showing, minimal, invasion, overlying, classified, pn01, r0, stage, 113
"TCGA-NJ-A4YQ…txt"	802	52	lung, upper, lobe, wedge, resection, invasive, differentiated, adenocarcinoma, grade, 3, measuring, 2, poorly, frozen, lobe wedge, non-small, cell, carcinoma, favor, cm, histologic, status, margins, bronchial, negative, vascular, parenchymal, pleural, visceral/parietal, invasion, absent, lymphatic, present, lymph, node, levels, 5, 10, 12, 6, extension, structures, non-neoplastic, pathology, na, tmn, stage, pt1bn0, pl0, gross, description

Figure 4: Table 2: Results of output terms generated by the Python algorithm

1 shows the set of 23 key test-words that were used as our "PROFILE" list. These are a mixture of lung carcinoma subtypes, cancer stages and other relevant diagnostic terms to be used a basis for the desired output terms. A separate list of roughly 200 "COMMON WORDS" were specified in another text file; most of which were downloaded from open-source platforms. Both files were placed in the same folder as the script and all the reports. Finally, the code was run separately for each of the 5 files. As mentioned in the Methods section, the file name must be copied and pasted inside the main function at the end of the script for each iteration. This makes it easy for the user to extract terms for a specific file at a time.

The results are shown in Table 2. The first column shows the first 10 characters of the filename – this is unique for each report, so the test can be reproduced for verification of results. The second and third columns show the number of words in the original report and the number of output words extracted by the algorithm, respectively. Finally, the output terms are shown on the last column as a comma-separated string. As evident, the output terms not only reflect the available "PROFILE" words from a report but also words that are similar. This is a result of the Euclidean distance calculation performed by the cosine similarity matrix.

CONCLUSION

[Significance of Results?]

[What can be done to improve results? Elaborate on these: Improve list of common words, improve list of profile words, "loop" the code so it can run multiple reports at once (maybe run whatever files are placed in the directory folder?), etc]

[Talk about next steps. Explain some medical imaging techniques that can be used in future stages of this research]

Free text and summary style reports are the current major formats of pathology reports. They remain the norm for reporting many non-neoplastic diseases. In the last two decades, with an increase in the amount

of information required to assess/detect cancer, concise reports have become more popular in reporting cancer cases. A mixture of both free text and concise style reports are used by most, but they are at most times lengthy and time consuming to compose. In some cases, digital images have been used for the pathology report in many centers worldwide, but their value and contribution are often questioned.

There is an overwhelming need for adopting a standardized concise report style especially for cancer cases [5]. The methods outlined in this paper can help us move closer to achieving standard.

REFERENCES

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5581350/ - Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification- Kentaro Inamura

[2] http://www.pathologyoutlines.com/topic/lungtumoradenoclass.html

[3] Gelfman, Nelson, et al. "Observer Variability in the Histopathologic Diagnosis of Lung Cancer." ATS Journals, www.atsjournals.org/doi/abs/10.1164/arrd.1970.101.5.671

[4] http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074674/ - Trends and Challenges in Pathology Practice: Choices and necessities - Hassan MH Kamel