Title of Work:
Autism Spectrum Disorder Detection from Semi-Structured and Unstructured Medical Data

Conference:
EURASIP Journal on Bioinformatics and Systems Biology

Rationale to ensure venue quality:
The EURASIP Journal on Bioinformatics and Systems Biology is a peer-reviewed open access journal published by SpringerOpen. It has been available for over a decade and offers a common platform for articles from a variety of areas, including
"signal processing, bioinformatics, statistics, biology and medicine." From what I can tell, Springer is an established research group that appears to be associated with many high-quality, peer-reviewed research articles. With regards to the article itself, the article has multiple citations and nearly 1,000 downloads since it's online publishing in February 2017.

Problem Statement:
The process of diagnosing Autism Spectrum Disorder (ASD) is traditionally very time-consuming, labor-intensive, and requires extensive expertise. However, medical experts agree that early identification of ASD can be very beneficial. Despite this, no efficient way of automating evaluation processes or programmatically assisting diagnosing physicians exists. To that end, Yuan et al. develop a tool capable of assisting physicians in digitizing, identifying, and classifying semi-structured and unstructured hand-written clinical data for the purpose of generating a recommendation on an ASD Diagnosis.

Paper Synopsis:
Autism Spectrum Disorder (ASD) is a classification of a range of disorders associated with neurological development. Currently, no laboratory test for ASD exists and diagnosing it has proven to be very complex and labor intensive. Unfortunately, given the complexity and resources required, this process can take a great deal of time, and can lead to, in some cases, a year-long waiting list just to begin the diagnosing process. This can have a negative impact on the child's health and can delay access to treatment. Therefore, there is an established need for quickening or automating some part of this process.

To that end, Yuan et al. examine the feasibility of a method for identifying children who potentially have ASD by applying Natural Language Processing and machine learning to a set of unstructured and semi-structured handwritten physician notes. Before examining their process, Yuan et al. examine some of the difficulties associated with working with biomedical data. The data cannot be evaluated with crowdsourcing techniques due to privacy requirements. Similarly,

the resources and data itself is both limited and private, making it difficult to accumulate a sufficiently large dataset. Finally, the data is not stored in a usable format, consisting of hand written data with a significant amount of noise. To address this, Yuan et al. propose a machine learning-based tool that converts semi-structured and unstructured natural language text into a digital format, mines relevant information from these documents, automatically de-identifies text to comply with privacy requirements, and allows for automated ASD detection through NLP-based extraction.

The first phase of the tool involves data collection and preprocessing. All medical forms were scanned into a digital format (.tif files). An entropy-based de-skewing algorithm was applied to any tilted or misaligned images to straighten the data into a useable format. Additionally, de-identification was automatically performed on semi-structured data using predefined form boundaries unique to each type of form. After this, handwriting recognition software including Omnipage Capture SDK, Captricity, and ABBYY were used to scan and digitize the results. Next, a series of lexical features were extracted to identify trends for machine learning. Features extracted include a Bag-of-words and n-gram models (focusing on bigrams and trigrams), term frequency-inverse document frequency (tf-idf), Latent Dirichlet Allocation (LDA), and Distributed Representation (Doc2vec), which were used to train a Support Vector Machine (SVM) for ASD Detection.

18,962 Lexical features were extracted across the entire dataset, but only 386 were found to have a non-zero weight on ASD diagnoses. These were identified through a 7-fold cross validation for evaluation and a 5-fold cross-validation for optimization of parameters during training. Upsampling was used to accommodate for the dataset having a small number of positive results (a correct diagnosis of ASD), which universally increased the Precision, Recall, and F2 Scores. After sufficient training, the system achieved an accuracy of 83.4% and a recall of 91.1%.

Future Work:

The obvious area for future work is to continue the experiment across larger datasets as they become available. Ideally, a dataset with enough positive results to not require upsampling would become available for the purposes of training and evaluating the study. Additionally, Yuan et al. noted that the weights of linguistic features by the classifier were all mostly very similar, meaning that a robust analysis based purely on NLP techniques is not feasible with the current study. Further investigation into these linguistic features could allow for a more robust and automated system. Finally, I'd like to see an analysis on the effectiveness of structured data (potentially from some sort of curated medical database with ASD data) in a system such as this. Allowing a comparison across the various data types could offer insight into what features are relevant for NLP extraction.