Automated Text Mining to Improve the Curation of Genes Associated with Complex Disease

Michael Superdock[1]; Alper Uzun, PhD[1,2]; Indra Neil Sarkar, PhD[1], MLIS; James Padbury, MD[1,2]

[1]The Warren Alpert Medical School and Center for Biomedical Informatics, Brown University; [2]Department of Pediatrics, Women & Infants Hospital of Rhode Island

ABSTRACT

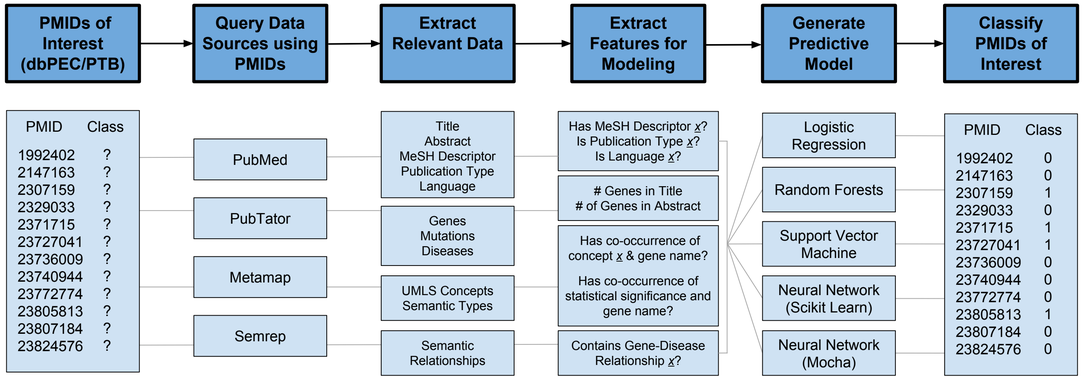
Manual curation of primary literature is a common, time-intensive approach for identifying genes associated with a disease of interest. This project aims to minimize the workload of manual curation for genetic studies by semi-automating the curation process. A computational pipeline was created using text-mining techniques to extract genetic data and other distinguishing features from articles. Five predictive models were trained on these features to classify articles as “considered” or “not considered” for later review by curators. The models were evaluated against manual classifications of curated papers from the Database for Preeclampsia (dbPEC) and the Database for Preterm Birth (dbPTB). A Random Forest classifier performed best for both datasets, with an AUC of 0.825 for dbPEC articles and an AUC of 0.918 for dbPTB articles. This classifier had results consistent with a 32.5% workload reduction for the curation of dbPEC articles and a 79.6% workload reduction for the curation of dbPTB articles, while still capturing over 95% of validated genes.

INTRODUCTION

Complex diseases often have heritable contributions from multiple genetic loci. One standard approach to identifying these gene-disease associations is the manual curation of primary literature. This requires a costly investment of time and resources, but it can yield a gene set that is immensely valuable for focusing subsequent genetic studies. Automated approaches to more efficiently identify gene-disease associations have long been a focus of informatics researchers.[1] A variety of fully-automated tools exist in this space, including DigSee[2], DISEASES[3], and DisGeNET[4]. Nevertheless, these tools provide gene sets less robust than manually curated databases that exist for the same diseases. This is consistent with the commonly held view that fully-automated curation systems are not accurate enough to replace manual curation.[5]

This project takes a semi-automated approach—a computational pipeline that uses previous curation data to evaluate new articles based on their relevance for the genetic study of a particular disease. This work showcases an integration of text-mining and machine learning on biomedical literature, which has relevance beyond identifying only gene-disease relationships.

METHODS



RESULTS

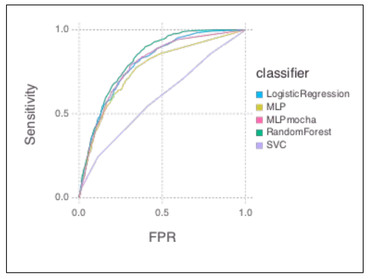


Figure 1. Receiver Operator Characteristic curve based on 5-fold cross validation results from 2134 papers curated for dbPEC.

dbPEC	
Classifier	**AUC**
Log. Regression	0.801
MLP	0.770
MLP mocha	0.801
Random Forest	**0.825**
SVC	0.598

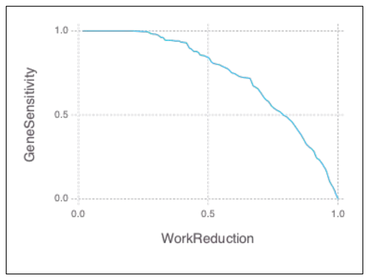


Figure 2. Random Forest workload savings plotted against the percentage of database-accepted genes found in classifier-accepted papers. Values are based on 5-fold cross validation of 2134 papers curated for dbPEC.

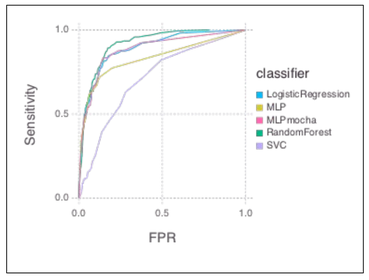


Figure 3. Receiver Operator Characteristic curve based on 5-fold cross validation results from 1056 papers curated for dbPTB.

dbPTB	
Classifier	**AUC**
Log. Regression	0.894
MLP	0.857
MLP mocha	0.895
Random Forest	**0.918**
SVC	0.718

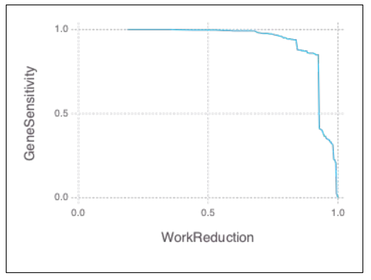


Figure 4. Random Forest workload savings plotted against the percentage of database-accepted genes found in classifier-accepted papers. Values are based on 5-fold cross validation of 1056 papers curated for dbPTB.

DISCUSSION

(1) These results indicate that a supervised machine learning approach is effective at classifying articles to be “considered” or “not considered” for the genetic study of our conditions of interest—preterm birth and preeclampsia—and that these classifications can be made using only metadata, titles, and abstracts.
(2) A random forest classifier outperformed all other classifiers for preterm birth and preeclampsia datasets, suggesting that it may be the best algorithm for this task.
(3) The consistency of this computation pipeline for both preterm birth and preeclampsia datasets may be indicative of its generalizability to other diseases.

FUTURE DIRECTIONS

(1) Compare the accuracy of this system to Abstrackr, a tool that similarly uses machine learning to classify articles for systematic reviews
(2) Vary training set size to determine how many curated papers are necessary to generate accurate classifications.
(3) Introduce additional training features and algorithm parameters.

CONCLUSION

This computational pipeline demonstrates that supervised machine learning can be used reliably to “consider” or “not consider” papers for curation, effectively reducing curator workload without significant loss of relevant genetic information. This tool may be useful to a researcher interested in curating primary literature to better understand genetic contributions to human disease.

References

1. Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform, 8*(5), 358-375. doi:10.1093/bib/bbm045
2. Kim, J., Kim, J. J., & Lee, H. (2017). An analysis of disease-gene relationship from Medline abstracts by DigSee. *Sci Rep, 7*, 40154. doi:10.1038/srep40154
3. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., & Jensen, L. J. (2015). DISEASES: text mining and data integration of disease-gene associations. *Methods, 74*, 83-89. doi:10.1016/j.ymeth.2014.11.020
4. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., . . . Furlong, L. I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res, 45*(D1), D833-D839. doi:10.1093/nar/gkw943
5. Karp, P. D. (2016). Can we replace curation with information extraction software? *Database (Oxford), 2016*. doi:10.1093/database/baw150

Acknowledgments

I would like to thank Paul Stey for assisting with data recovery for this project. This work was funded in part by the Scholarly Concentration Program in The Warren Alpert Medical School of Brown University and National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.