A Multi-Label Classification Framework for Cancer

Markian Jaworsky (u1101991)

Supervisors:
Principal supervisor Xiaohui Tao
Associate supervisors Jianming Yong, Ji Zhang, Lei Pan (External), Shiva Pokhrel (External)

School of Mathematics, Physics, and Computing, University of Southern Queensland, Australia

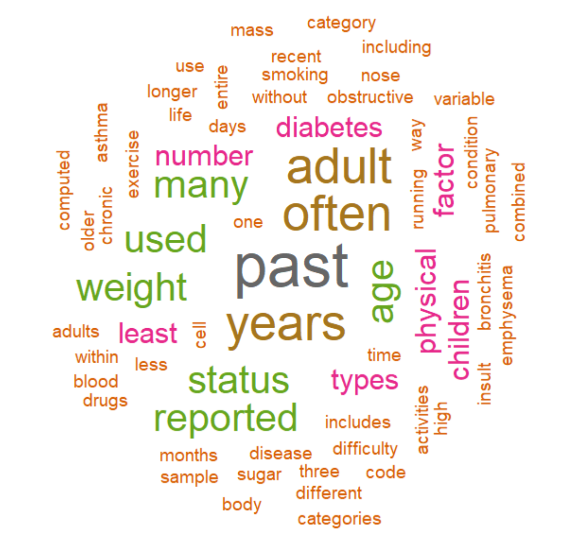
A presentation partially fulfil the requirements for the conformation of PhD candidature, August 21, 2022

Introduction

- A range of risk factors can be used as predictor variables in the likelihood of developing chronic illness.
- With awareness, patients can adapt their lifestyle in order to improve their chances of longer term survival.
- Risk factors can be categorized as being lifestyle, environmental or biomedical and can change over time.

Introduction

- Word cloud of text frequency variance between smoking and non-smoking lung cancer patients.

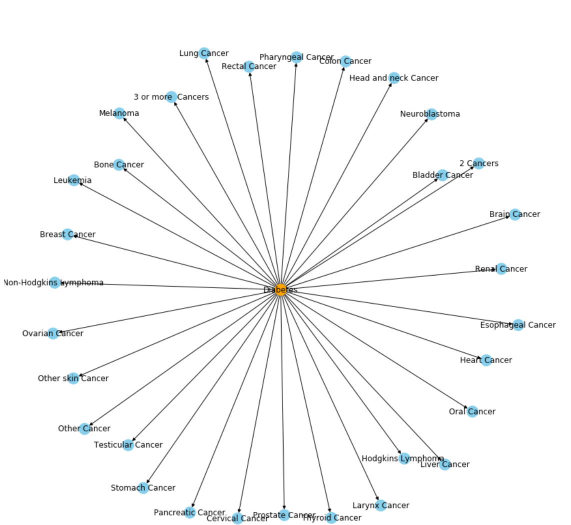


Introduction - Key Research Approach Change

- Most studies are typically focused on single disease datasets, however, to ensure that health advice is generalized and contemporary, the features that can predict the likelihood of many diseases can improve health advice effectiveness, when considering the point of view of the patient.
- 6 cancers correlate with diabetes according to the 2020 study of Wang *et al.* [1], suggesting that the cause of many cancers can be simplified to a combination of DNA damage and inflammation.
- Predictor variables with a 1-to-many cancer relationship can improve health advice of predictor variables with 1-to-1 relationships.

Research Scope - Goal

- A Framework for Handling the Data Challenges in the Multi-Label Classification of Multiple Cancer subtypes.
    - Providing current and evidental health advice.



Research Scope - Problem

- Individual diseases have individual data patterns, in order to predict multiple diseases, we must be able to overcome the many data challenges that can occur.

Research Scope - Questions

- How can we introduce innovation and provide a framework for the identification of data patterns in the classification of cancer subtypes?
- Using agile methodology how do we develop a new framework for the construction of predictive models of multiple cancers, and multiple cancer occurrences?

Research Scope - Objectives

- Develop a robust approach to using health surveys for the construction of predictive models of multiple chronic illnesses.
- Identify the optimum subset of health survey predictor variables to classify the largest subset of multiple cancer subtypes.

Research Scope - Limitations and Assumptions

- This study is focused on the technical challenges of data, the topic of data privacy is outside the scope of the reviewed literature.

Research Scope - Expected Research Outcome 1

- A subset of predictor variables that can be robustly used to predict a level of risk in multiple cancer subtypes.

Research Scope - Expected Research Outcome 2

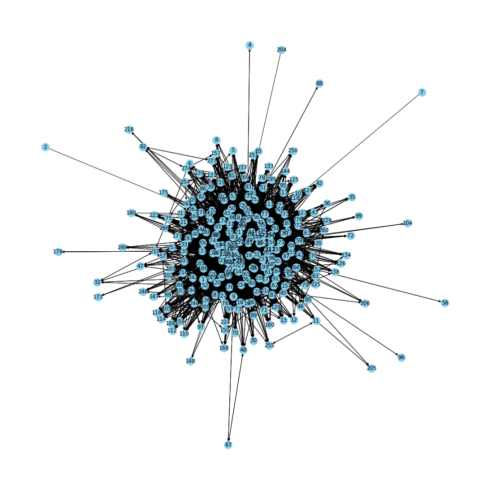
- A Framework for Handling the Data Challenges in the Multi-Label Classification of Multiple Chronic Illnesses.

Research Concept - Knowledge Graphs

- Our review of knowledge graph studies indicates that most research depends on the manual selection of articles to be used in constructing a knowledge graph.
- By identifying a complete source of knowledge we can automate the construction of a complete knowledge graph.
- Using a knowledge-based method to select features gives a predictive model assurance against detecting spurious correlations.

Research Scope - Contribution 1

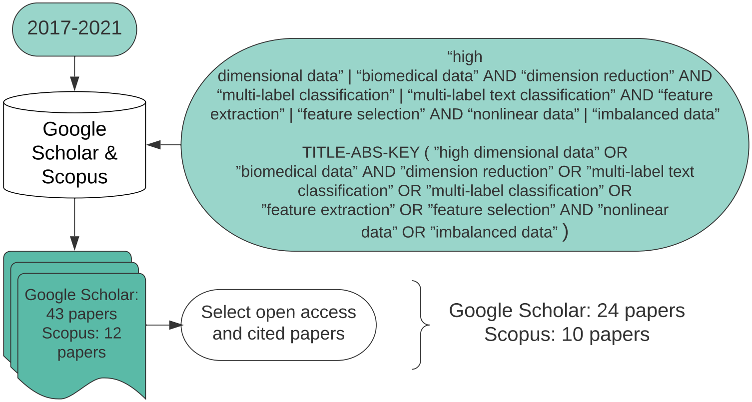
- Novel automated chapter ranking significance-based knowledge graph to identify inter-feature relations.
- Our automation uses the 26 chapters of WHO ICD.



Research Scope - Potential Contribution 2

- Further improve contribution 1 with prioritization of linear features over nonlinear.
- 2020 BRFSS Health Survey consists of 279 questions, of which 40 (86%) have nonlinear properties. Eg. Yes, No, Unsure.
- Only 39 (13%) questions can be considered to have linear variable answers. Eg. age, height, weight.

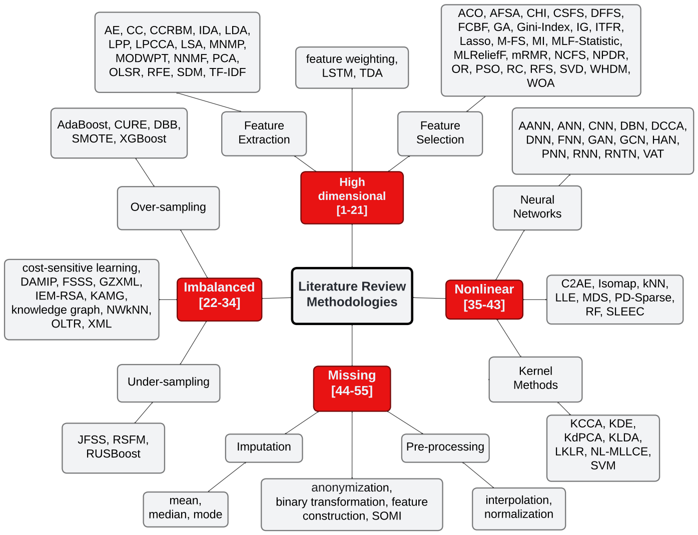
Literature Review - Scope



Literature Review - Data Challenges

- High Dimensional Datasets [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]
- Imbalanced Class Datasets [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35]
- Multiple and Nonlinear Datasets [36, 37, 38, 39, 40, 41, 42, 43, 44]
- Datasets with Missing and Erroneous Values [45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56]

Literature Review

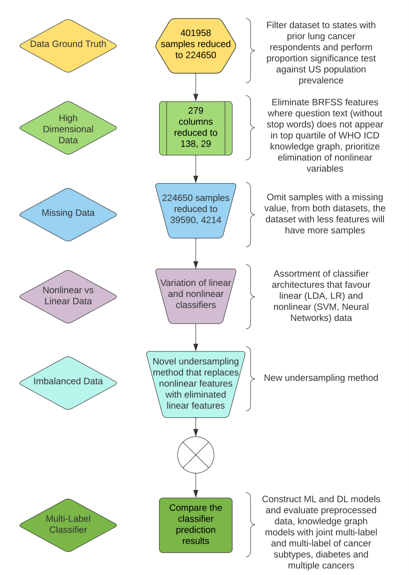


- There is not an existing single widely-adopted framework for uncovering data patterns to support the construction of multi-label chronic illness classifier.

Literature Review

- Therefore we are unable to satisfactorily state that RQ1 is adequately supported and proposed to conduct further research in order to answer RQ2.

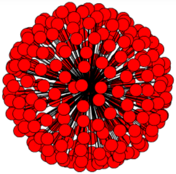
Research Design - Data Preparation



Research Design - Goal

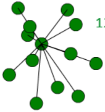
- Reproducible health survey analysis and predictive model construction

Research Design



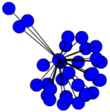
278 Health Survey Predictor Variable Candidates

Research Design:
Find the subset of predictor variables and algorithm that best classifies the largest subset of response variables



12 ML/DL Algorithm Candidates

34 Health Survey Response Variable Candidates



Research Design - Health Survey Classifiers

- We validate our general health advice by testing our feature selected dataset with the baseline dataset using these classifiers
- AdaBoost [57]
- K-nearest neighbours [57, 58]
- Linear Discriminant Analysis [59]
- Logistic Regression [57, 58]
- Multinomial Naive Bayes [57]
- Random Forest [57, 58]
- RUSBoost [57]
- Support Vector Machine [57]
- TensorFlow Convolutional Neural Networks [57]

Research Design - Parameter Tuning Optimization

- To be both optimal and reproducible, a tuning strategy is required
- Hyperopt [60, 61]
- Grid Search [62]
- ULMfit [63]

Research Design - Evaluation Metrics

- Macro Average F1-Score [15, 19, 18, 12, 50, 23, 27, 30, 26, 44, 42, 52, 53, 9, 64, 65, 62, 66, 55, 67, 31, 68, 48, 69, 49, 70, 71, 72, 73, 8, 74, 75, 76, 77, 78]
- Precision at Top K (P@k) [18, 40, 41, 79, 80, 81, 82, 32, 64, 33, 35, 83, 84, 85, 61, 86, 74, 77, 87]

Research Design - Dataset

- The United States CDC make available an anonymized annual Behavioral Risk Factor Surveillance System (BRFSS) survey data, which is free to the public domain and may be copied and distributed without permission.

From: Garvin, William S. (CDC/DDNID/NCCDPHP/DPH) <wsg1@cdc.gov>
Sent: 03 September 2021 23:02
To: Markian Jaworsky <Markian.Jaworsky@usq.edu.au>
Cc: Barrett, Drue H. (CDC/DDPHSS/OS/OSI) <dhb1@cdc.gov>
Subject: RE: Question on statement of CDC BRFSS Annual Questionnaire Data Ethics Approval

Dear Markian,

The Behavioral Risk Factor Surveillance System (BRFSS) is approved by US Office of Management and Budget (OMB) to collect data from the US general population under OMB Control number 0920-1061.

The CDC Human Research Protection Office has determined that this research activity (BRFSS data collection) remains exempt under 45 CFR 46.101(b)(2).

The BRFSS is a state based survey conducted in partnership with the participating state-health departments. A common core questionnaire and standardized optional modules are approved by the states and CDC programs each year. The state health departments implement the survey and oversee the ongoing data collection for their state, whether through a contracted data collector or in-house data collection. The states and data collectors have institutional review boards which review state-specific questionnaire content and determine what is applicable for inclusion in the BRFSS for a given state. The questionnaire and basic BRFSS data collection protocol are covered in the BRFSS Overview documentation released with the public use data set each year on the BRFSS website.

Hopefully this provides the information needed for explanation of the approval process of the BRFSS.

If you have additional questions or concerns please let us know.

Thank you,

Bill Garvin
Survey Operations
Population Health Surveillance Branch
Division of Population Health
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
(770) 488-4621

Research Design - Baseline Dataset

- The baseline dataset to be used first applies a filter of the geographical state of the respondent, states eliminated from the dataset have been identified by searching through prior years of surveys and deselecting states with no lung cancer candidates.
- In order to establish a dataset that is representative of the ground truth, the BRFSS annual survey can be filtered by states until an insignificant proportion of lung cancer, a chronic illness with a high prevalence amongst both male and female patients, such that it is comparable to the prevalence of lung cancer in the population of the USA.

Research Plan - Phase 1

- Knowledge Graph-Based Feature Selection for Multi-Label Classification on Health Survey Data

Research Plan - Phase 2

- Automated Knowledge Graph Construction for Healthcare Domain

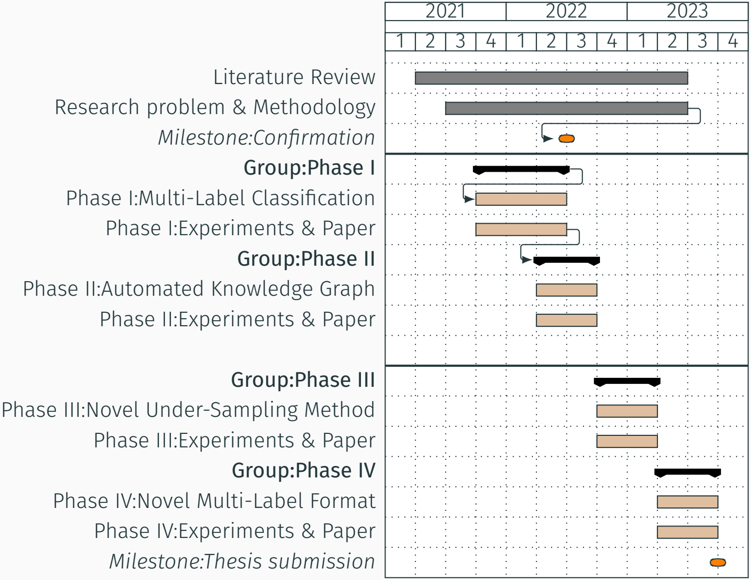
Research Plan - Phase 3

- Nonlinear with Linear Feature Replacement in Multi-Label Classification

Research Plan - Phase 4

- Construct a Multi-Label format which best classifies cancer subtypes

Timeline - Gantt Chart



Timeline - Dissemination plan

- Confirmation of Candidature Draft Submission: 02/08/2022
- Science Direct Artificial Intelligence in Medicine Submission: 30/04/2023

- Biomedical Engineering Review IEEE Submission: 25/04/2022
- IEEE International Conference on Data Mining (ICDM 2022) Submission: 11/06/2022
- Springer International Conference on Health Information Systems Acceptance: 16/08/2022

Other Issues - Ethics Approval

- USQ HREC ID: H21REA222
- Approval date: 15/10/2021
- Expiry date: 15/10/2024
- USQ HREC status: Approved

Other Issues - Publication Acceptance

- Part-time study
- Publication acceptance time

Conclusion 1

- Imbalanced class datasets require training sets to have balanced class data, or appropriate metrics to assess the precision of minority class predictions.
- Features with missing values can still be transformed into meaningful information, even in a binary format.

Conclusion 2

- The widely adopted oversampling method SMOTE only supports single label data samples, as such an ensemble of linear, nonlinear, and majority class undersampling multi-label classifiers can provide the best coverage for identifying and predicting true positive minority class samples.
- It was also observed that the usage of linear variables can be more valuable than nonlinear when a dataset has imbalanced classes.

Conclusion 3

- We can automate the construction of a knowledge graph using a Wilcox rank significance test and in order to assist the selection of features that contain the highest and lowest levels of inter-feature relations, using a knowledge base such as the WHO ICD.

References

[1] Mina Wang, Yingying Yang, and Zehuan Liao. “Diabetes and cancer: Epidemiological and biological links”. In: *World journal of diabetes* 11.6 (2020), p. 227.

[2] Marziyeh Arabnejad et al. “Nearest-Neighbor Projected Distance Regression for Epistasis Detection in GWAS With Population Structure Correction”. In: *Frontiers in Genetics* 11 (2020). DOI: 10.3389/fgene.2020.00784. URL: https://doi.org/10.3389/fgene.2020.00784.

[3] Alberto Cano, Sebastián Ventura, and Krzysztof J Cios. “Multi-objective genetic programming for feature extraction and data visualization”. In: *Soft Computing* 21.8 (2017), pp. 2069–2089. DOI: 10.1007/s00500-015-1907-y. URL: https://doi.org/10.1007/s00500-015-1907-y.

[4] Francisco Chinesta et al. “Virtual, digital and hybrid twins: a new paradigm in data-based engineering and engineered data”. In: *Archives of computational methods in engineering* 27.1 (2020), pp. 105–134. DOI: 10.1007/s11831-018-9301-4. URL: https://doi.org/10.1007/s11831-018-9301-4.

[5] Travers Ching et al. “Opportunities and obstacles for deep learning in biology and medicine”. In: *Journal of The Royal Society Interface* 15.141 (2018), p. 20170387. DOI: 10.1098/rsif.2017.0387. URL: https://doi.org/10.1098/rsif.2017.0387.

[6] Juying Dai et al. “Signal-based intelligent hydraulic fault diagnosis methods: Review and prospects”. In: *Chinese Journal of Mechanical Engineering* 32.1 (2019), pp. 1–22. DOI: 10.1186/s10033-019-0388-9. URL: https://doi.org/10.1186/s10033-019-0388-9.

[7] Fei Deng et al. “Predict multicategory causes of death in lung cancer patients using clinicopathologic factors”. In: *Computers in Biology and Medicine* 129 (2021), p. 104161. DOI: 10.1016/j.compbiomed.2020.104161. URL: https://doi.org/10.1016/j.compbiomed.2020.104161.

[8] Rania M Ghoniem, Nawal Alhelwa, and Khaled Shaalan. “A Novel Hybrid Genetic-Whale Optimization Model for Ontology Learning from Arabic Text”. In: *Algorithms* 12.9 (2019), p. 182. DOI: *10.3390/a12090182*. URL: *https://doi.org/10.3390/a12090182*.

[9] Hai Huang and Huan Liu. “Feature selection for hierarchical classification via joint semantic and structural information of labels”. In: *Knowledge-Based Systems* 195 (2020), p. 105655. DOI: *10.1016/j.knosys.2020.105655*. URL: *https://doi.org/10.1016/j.knosys.2020.105655*.

[10] Kwang-Ho In et al. “Lung cancer patients who are asymptomatic at diagnosis show favorable prognosis: a Korean Lung Cancer Registry Study”. In: *Lung cancer* 64.2 (2009), pp. 232–237. DOI: *10.1016/j.lungcan.2008.08.005*. URL: *https://doi.org/10.1016/j.lungcan.2008.08.005*.

[11] He Jiang. “Sparse estimation based on square root nonconvex optimization in high-dimensional data”. In: *Neurocomputing* 282 (2018), pp. 122–135. DOI: *10.1016/j.neucom.2017.12.025*. URL: *https://doi.org/10.1016/j.neucom.2017.12.025*.

[12] Kyoungok Kim. “An improved semi-supervised dimensionality reduction using feature weighting: application to sentiment analysis”. In: *Expert Systems with Applications* 109 (2018), pp. 49–65. DOI: *10.1016/j.eswa.2018.05.023*. URL: *https://doi.org/10.1016/j.eswa.2018.05.023*.

[13] Li Ma and Suohai Fan. “CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests”. In: *BMC bioinformatics* 18.1 (2017), pp. 1–18. DOI: *10.1186/s12859-017-1578-z*. URL: *https://doi.org/10.1186/s12859-017-1578-z*.

[14] Mahla Mokhtia, Mahdi Eftekhari, and Farid Saberi-Movahed. “Feature selection based on regularization of sparsity based regression models by hesitant fuzzy correlation”. In: *Applied Soft Computing* 91 (2020), p. 106255. DOI: *10.1016/j.asoc.2020.106255*. URL: *https://doi.org/10.1016/j.asoc.2020.106255*.

[15] Kwang Ho Park et al. “Deep Learning Feature Extraction Approach for Hematopoietic Cancer Subtype Classification”. In: *International Journal of Environmental Research and Public Health* 18.4 (2021), p. 2197. DOI: *10.3390/ijerph18042197*. URL: *https://doi.org/10.3390/ijerph18042197*.

[16] Julliano Trindade Pintas, Leandro AF Fernandes, and Ana Cristina Bicharra Garcia. "Feature selection methods for text classification: a systematic literature review". In: *Artificial Intelligence Review* (2021), pp. 1–52. DOI: 10.1007/s10462-021-09970-6. URL: https://doi.org/10.1007/s10462-021-09970-6.

[17] Wenbin Qian et al. "Mutual information-based label distribution feature selection for multi-label learning". In: *Knowledge-Based Systems* 195 (2020), p. 105684. DOI: 10.1016/j.knosys.2020.105684. URL: https://doi.org/10.1016/j.knosys.2020.105684.

[18] Wissam Siblini, Pascale Kuntz, and Frank Meyer. "A review on dimensionality reduction for multi-label classification". In: *IEEE Transactions on Knowledge and Data Engineering* 33.3 (2019), pp. 839–857. DOI: 10.1109/TKDE.2019.2940014. URL: https://doi.org/10.1109/TKDE.2019.2940014.

[19] Abdellah Tebani, Carlos Afonso, and Soumeya Bekri. "Advances in metabolome information retrieval: turning chemistry into biology. Part II: biological information recovery". In: *Journal of inherited metabolic disease* 41.3 (2018), pp. 393–406. DOI: 10.1007/s10545-017-0080-0. URL: https://doi.org/10.1007/s10545-017-0080-0.

[20] Lokeswari Venkataramana, Shomona Gracia Jacob, and Rajavel Ramadoss. "A parallel multilevel feature selection algorithm for improved cancer classification". In: *Journal of Parallel and Distributed Computing* 138 (2020), pp. 78–98. DOI: 10.1016/j.jpdc.2019.12.015. URL: https://doi.org/10.1016/j.jpdc.2019.12.015.

[21] Xinzheng Xu et al. "Review of classical dimensionality reduction and sample selection methods for large-scale data processing". In: *Neurocomputing* 328 (2019), pp. 5–15. DOI: 10.1016/j.neucom.2018.02.100. URL: https://doi.org/10.1016/j.neucom.2018.02.100.

[22] Mingwei Zhang, Ziqi Ji, and Zebo Dong. "Classification based on label semantic characteristic analysis". In: *2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*. IEEE. 2017, pp. 78–82. DOI: 10.1109/ACIRS.2017.7986069. URL: https://doi.org/10.1109/ACIRS.2017.7986069.

[23] Meng Liu et al. "Cost-sensitive feature selection via f-measure optimization reduction". In: *Thirty-First AAAI Conference on Artificial Intelligence*. 2017. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10770.

[24] FY Chin, CA Lim, and KH Lem. "Handling leukaemia imbalanced data using synthetic minority oversampling technique (SMOTE)". In: *Journal of Physics: Conference Series*. Vol. 1988. IOP Publishing. 2021, p. 012042. DOI: 10.1088/1742-6596/1988/1/012042. URL: https://doi.org/10.1088/1742-6596/1988/1/012042.

[25] Risky Frasetio Wahyu Pratama, Santi Wulan Purnami, and Santi Puteri Rahayu. "Boosting support vector machines for imbalanced microarray data". In: *Procedia computer science* 144 (2018), pp. 174–183. DOI: 10.1016/j.procs.2018.10.517. URL: https://doi.org/10.1016/j.procs.2018.10.517.

[26] Xuchun Wang et al. "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier". In: *BMC medical informatics and decision making* 21.1 (2021), pp. 1–14. DOI: 10.1186/s12911-021-01471-4. URL: https://doi.org/10.1186/s12911-021-01471-4.

[27] Barbara Pes. "Learning from High-Dimensional and Class-Imbalanced Datasets Using Random Forests". In: *Information* 12.8 (2021), p. 286. DOI: 10.3390/info12080286. URL: https://doi.org/10.3390/info12080286.

[28] Aliaksandr Barushka and Petr Hajek. "Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks". In: *Applied Intelligence* 48.10 (2018), pp. 3538–3556. DOI: 10.1007/s10489-018-1161-y. URL: https://doi.org/10.1007/s10489-018-1161-y.

[29] Xuedong Li et al. "Improving rare disease classification using imperfect knowledge graph". In: *BMC medical informatics and decision making* 19.5 (2019), pp. 1–10. DOI: 10.1186/s12911-019-0938-1. URL: https://doi.org/10.1186/s12911-019-0938-1.

[30] Patrícia Gonzalez-Dias et al. "Methods for predicting vaccine immunogenicity and reactogenicity". In: *Human vaccines & immunotherapeutics* 16.2 (2020), pp. 269–276. DOI: 10.1080/21645515.2019.1697110. URL: https://doi.org/10.1080/21645515.2019.1697110.

[31] Linchao Zhu and Yi Yang. "Inflated episodic memory with region self-attention for long-tailed visual recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4344–4353. DOI: 10.1109/CVPR42600.2020.00440. URL: https://doi.org/10.1109/CVPR42600.2020.00440.

References v

[32] Jingjing Li et al. "On both cold-start and long-tail recommendation with social data". In: *IEEE Transactions on Knowledge and Data Engineering* 33.1 (2019), pp. 194–208. DOI: 10.1109/TKDE.2019.2924656. URL: https://doi.org/10.1109/TKDE.2019.2924656.

[33] Tharun Medini, Beidi Chen, and Anshumali Shrivastava. "SOLAR: Sparse Orthogonal Learned and Random Embeddings". In: *arXiv preprint arXiv:2008.13225* (2020). URL: https://arxiv.org/abs/2008.13225.

[34] Wei-Cheng Chang et al. "Pre-training tasks for embedding-based large-scale retrieval". In: *arXiv preprint arXiv:2002.03932* (2020). URL: https://arxiv.org/abs/2002.03932.

[35] Shanshan Wu et al. "Learning a compressed sensing measurement matrix via gradient unrolling". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6828–6839. URL: https://arxiv.org/abs/1806.10175.

[36] Mingda Li, Weiting Gao, and Yi Chen. "A Topic and Concept Integrated Model for Thread Recommendation in Online Health Communities". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 765–774. DOI: 10.1145/3340531.3411933. URL: https://doi.org/10.1145/3340531.3411933.

[37] Betty van Aken et al. "Clinical outcome prediction from admission notes using self-supervised knowledge integration". In: *arXiv preprint arXiv:2102.04110* (2021). URL: https://arxiv.org/abs/2102.04110.

[38] Stephen L France and Sanjoy Ghose. "Marketing analytics: Methods, practice, implementation, and links to other fields". In: *Expert Systems with Applications* 119 (2019), pp. 456–475. DOI: 10.1016/j.eswa.2018.11.002. URL: https://doi.org/10.1016/j.eswa.2018.11.002.

[39] Xinghao Yang et al. "A survey on canonical correlation analysis". In: *IEEE Transactions on Knowledge and Data Engineering* (2019). DOI: 10.1109/TKDE.2019.2958342. URL: https://doi.org/10.1109/TKDE.2019.2958342.

[40] Wenjie Zhang et al. "Deep extreme multi-label learning". In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 2018, pp. 100–107. URL: https://arxiv.org/abs/1704.03718.

[41] Weiwei Liu and Xiaobo Shen. "Sparse extreme multi-label learning with oracle property". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4032–4041. URL: https://proceedings.mlr.press/v97/liu19d.html.

[42] Chao Tan and Genlin Ji. "LKLR: A local tangent space-alignment kernel least-squares regression algorithm". In: *Tsinghua Science and Technology* 24.4 (2019), pp. 389–399. DOI: 10.26599/TST.2018.9010120. URL: https://doi.org/10.26599/TST.2018.9010120.

[43] Jia Chen, Gang Wang, and Georgios B Giannakis. "Nonlinear dimensionality reduction for discriminative analytics of multiple datasets". In: *IEEE Transactions on Signal Processing* 67.3 (2018), pp. 740–752. DOI: 10.1109/TSP.2018.2885478. URL: https://doi.org/10.1109/TSP.2018.2885478.

[44] Shu-Kai S Fan et al. "Defective wafer detection using a denoising autoencoder for semiconductor manufacturing processes". In: *Advanced Engineering Informatics* 46 (2020), p. 101166. DOI: 10.1016/j.aei.2020.101166. URL: https://doi.org/10.1016/j.aei.2020.101166.

[45] Omogbai Oleghe. "A predictive noise correction methodology for manufacturing process datasets". In: *Journal of Big Data* 7.1 (2020), pp. 1–27. DOI: 10.1186/s40537-020-00367-w. URL: https://doi.org/10.1186/s40537-020-00367-w.

[46] Michael P Bancks et al. "Epidemiology of diabetes phenotypes and prevalent cardiovascular risk factors and diabetes complications in the National Health and Nutrition Examination Survey 2003–2014". In: *Diabetes research and clinical practice* 158 (2019), p. 107915. DOI: 10.1016/j.diabres.2019.107915. URL: https://doi.org/10.1016/j.diabres.2019.107915.

[47] Emmanouil Antonios Platanios et al. "Learning from Imperfect Annotations: An End-to-End Approach". In: (2019). URL: https://openreview.net/forum?id=rJlVdREKDS.

[48] Stefano Melacci et al. "Domain Knowledge Alleviates Adversarial Attacks in Multi-Label Classifiers". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). DOI: 10.1109/TPAMI.2021.3137564. URL: https://doi.org/10.1109/TPAMI.2021.3137564.

[49] Nanyang Wang et al. "Pixel2mesh: 3d mesh model generation via image guided deformation". In: *IEEE transactions on pattern analysis and machine intelligence* (2020). DOI: 10.1109/TPAMI.2020.2984232. URL: https://doi.org/10.1109/TPAMI.2020.2984232.

[50] Youqiang Zhang et al. "A novel ensemble method for k-nearest neighbor". In: *Pattern Recognition* 85 (2019), pp. 13–25. DOI: 10.1016/j.patcog.2018.08.003. URL: https://doi.org/10.1016/j.patcog.2018.08.003.

[51] Bertrand De Meulder et al. "A computational framework for complex disease stratification from multiple large-scale datasets". In: *BMC systems biology* 12.1 (2018), pp. 1–23. DOI: 10.1186/s12918-018-0556-z. URL: https://doi.org/10.1186/s12918-018-0556-z.

[52] Bain Khusnul Khotimah, Miswanto Miswanto, and Herry Suprajitno. "Optimization of feature selection using genetic algorithm in naïve Bayes classification for incomplete data". In: *Int. J. Intell. Eng. Syst* 13.1 (2020), pp. 334–343. DOI: 10.22266/ijies2020.0229.31. URL: https://doi.org/10.22266/ijies2020.0229.31.

[53] Muhammad Adil et al. "LSTM and bat-based RUSBoost approach for electricity theft detection". In: *Applied Sciences* 10.12 (2020), p. 4378. DOI: 10.3390/app10124378. URL: https://doi.org/10.3390/app10124378.

[54] Sakib Mahmud Khan et al. "Multi-class twitter data categorization and geocoding with a novel computing framework". In: *Cities* 96 (2020), p. 102410. DOI: 10.1016/j.cities.2019.102410. URL: https://doi.org/10.1016/j.cities.2019.102410.

[55] Khan Md Hasib et al. "A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service". In: *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE. 2021, pp. 450–455. DOI: 10.1109/ICICT4SD50815.2021.9396879. URL: https://doi.org/10.1109/ICICT4SD50815.2021.9396879.

[56] Doyen Sahoo, Chenghao Liu, and Steven CH Hoi. "Malicious URL detection using machine learning: A survey". In: *arXiv preprint arXiv:1701.07179* (2017). URL: https://arxiv.org/abs/1701.07179.

[57] R Prashanth and Sumantra Dutta Roy. "Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning". In: *Neurocomputing* 305 (2018), pp. 78–103.

[58] Fikrewold H Bitew et al. "Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey". In: *Genus* 76.1 (2020), pp. 1–16.

[59] Carlo Ricciardi et al. "Linear discriminant analysis and principal component analysis to predict coronary artery disease". In: *Health informatics journal* 26.3 (2020), pp. 2181–2192.

[60] Ilias Chalkidis et al. "Large-scale multi-label text classification on EU legislation". In: *arXiv preprint arXiv:1906.02192* (2019). DOI: 10.18653/v1/P19-1636. URL: https://doi.org/10.18653/v1/P19-1636.

[61] Ilias Chalkidis et al. "Extreme multi-label legal text classification: A case study in EU legislation". In: *arXiv preprint arXiv:1905.10892* (2019). DOI: 10.18653/v1/W19-2209. URL: https://doi.org/10.18653/v1/W19-2209.

[62] Louis Létinier et al. "Artificial intelligence for unstructured healthcare data: application to coding of patient reporting of adverse drug reactions". In: *Clinical Pharmacology & Therapeutics* (2021). DOI: 10.1002/cpt.2266. URL: https://doi.org/10.1002/cpt.2266.

[63] Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. "Large scale legal text classification using transformer models". In: *arXiv preprint arXiv:2010.12871* (2020). URL: https://arxiv.org/abs/2010.12871.

[64] Dezhao Song et al. "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training". In: *Information Systems* (2021), p. 101718. DOI: 10.1016/j.is.2021.101718. URL: https://doi.org/10.1016/j.is.2021.101718.

[65] Kathryn Annette Chapman and Günter Neumann. "Automatic ICD Code Classification with Label Description Attention Mechanism.". In: *IberLEF@ SEPLN*. 2020, pp. 477–488. URL: https://www.semanticscholar.org/paper/Automatic-ICD-Code-Classification-with-Label-Chapman-Neumann/af3247725d9327857d922f17078b73cd1cba3f49.

[66] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. "Generalized zero-shot recognition based on visually semantic embedding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2995–3003. URL: https://arxiv.org/abs/1811.07993.

[67] Jian Wang et al. "Deep metric learning with angular loss". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2593–2601. URL: https://arxiv.org/abs/1708.01682.

[68] Muhammad Ali Ibrahim et al. "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification". In: *Journal of Biomedical Informatics* 116 (2021), p. 103699. DOI: 10.1016/j.jbi.2021.103699. URL: https://doi.org/10.1016/j.jbi.2021.103699.

[69] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. "Learning classifiers for target domain with limited or no labels". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7643–7653. URL: https://proceedings.mlr.press/v97/zhu19d.html.

[70] Lihi Dery. "Multi-label Ranking: Mining Multi-label and Label Ranking Data". In: *arXiv preprint arXiv:2101.00583* (2021). URL: https://arxiv.org/abs/2101.00583.

[71] Pedro Ruas et al. "LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents.". In: *IberLEF@ SEPLN*. 2020, pp. 422–437. URL: https://www.researchgate.net/publication/344429519_LasigeBioTM_at_CANTEMIST_Named_Entity_Recognition_and_Normalization_of_Tumour_Morphology_Entities_and_Clinical_Coding_of_Spanish_Health-related_Documents.

[72] Jiayu Wu. "Leveraging Label Information in Representation Learning for Multi-label Text Classication". PhD thesis. UCLA, 2019. URL: https://escholarship.org/uc/item/3870d965.

[73] Donna Xu et al. "Survey on multi-output learning". In: *IEEE transactions on neural networks and learning systems* 31.7 (2019), pp. 2409–2429. DOI: 10.1109/TNNLS.2019.2945133. URL: https://doi.org/10.1109/TNNLS.2019.2945133.

[74] Qian Li et al. "A survey on text classification: From shallow to deep learning". In: *arXiv preprint arXiv:2008.00364* (2020). URL: https://arxiv.org/abs/2008.00364.

[75] Willem Waegeman, Krzysztof Dembczyński, and Eyke Hüllermeier. "Multi-target prediction: a unifying view on problems and methods". In: *Data Mining and Knowledge Discovery* 33.2 (2019), pp. 293–324. DOI: 10.1007/s10618-018-0595-5. URL: https://doi.org/10.1007/s10618-018-0595-5.

[76] Rafael Leal. "Unsupervised zero-shot classification of Finnish documents using pre-trained language models". PhD thesis. University of Helsinki, 2020. URL: http://urn.fi/URN:NBN:fi:hulib-202012155147.

References x

[77] Kalina Jasinska-Kobus et al. "Probabilistic label trees for extreme multi-label classification". In: *arXiv preprint arXiv:2009.11218* (2020). URL: https://arxiv.org/abs/2009.11218.

[78] Jinseok Nam. "Learning Label Structures with Neural Networks for Multi-label Classification". PhD thesis. Technische Universität, 2019. URL: https://tuprints.ulb.tu-darmstadt.de/id/eprint/8738.

[79] Elham J Barezi, James T Kwok, and Hamid R Rabiee. "Multi-Label learning in the independent label sub-spaces". In: *Pattern Recognition Letters* 97 (2017), pp. 8–12. DOI: 10.1016/j.patrec.2017.06.024. URL: https://doi.org/10.1016/j.patrec.2017.06.024.

[80] Bingyu Wang et al. "Ranking-based autoencoder for extreme multi-label classification". In: *arXiv preprint arXiv:1904.05937* (2019). URL: https://arxiv.org/abs/1904.05937.

[81] Abhilash Gaure et al. "A probabilistic framework for zero-shot multi-label learning". In: *The Conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 1. 2017, p. 3. URL: https://www.semanticscholar.org/paper/A-Probabilistic-Framework-for-Multi-Label-Learning-Gaure-Rai/9c259f81257355f1e6a386d5f9ea5e4fe7744447.

[82] Nilesh Gupta et al. "Generalized Zero-Shot Extreme Multi-label Learning". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 527–535. DOI: 10.1145/3447548.3467426. URL: https://doi.org/10.1145/3447548.3467426.

[83] Sihong Xie and Philip S Yu. "Active zero-shot learning: a novel approach to extreme multi-labeled classification". In: *International Journal of Data Science and Analytics* 3.3 (2017), pp. 151–160. DOI: 10.1007/s41060-017-0042-5. URL: https://doi.org/10.1007/s41060-017-0042-5.

[84] Anthony Rios and Ramakanth Kavuluru. "EMR coding with semi–parametric multi–head matching networks". In: *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2018. NIH Public Access. 2018, p. 2081. DOI: 10.18653/v1/N18-1189. URL: https://doi.org/10.18653/v1/N18-1189.

[85] Kalina Jasinska-Kobus et al. "Online probabilistic label trees". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1801–1809. URL: https://arxiv.org/abs/2007.04451.

[86] Purvi Prajapati and Amit Thakkar. “Performance improvement of extreme multi-label classification using K-way tree construction with parallel clustering algorithm”. In: *Journal of King Saud University-Computer and Information Sciences* (2021). DOI: *10.1016/j.jksuci.2021.02.014*. URL: *https://doi.org/10.1016/j.jksuci.2021.02.014*.

[87] Steven CH Hoi et al. “Online learning: A comprehensive survey”. In: *Neurocomputing* 459 (2021), pp. 249–289. DOI: *10.1016/j.neucom.2021.04.112*. URL: *https://doi.org/10.1016/j.neucom.2021.04.112*.

Thanks

Questions?

Markian Jaworsky
Markian.Jaworsky@usq.edu.au