Guest Editorial for Selected Papers from BIOKDD 2018 and DMBIH 2018

Da Yan, Xin Gao, Samah J. Fodeh, and Jake Y. Chen

BIOKDD 2018 Overview. The International Workshop on Data Mining in Bioinformatics (BIOKDD), held in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining for 17 years, has successfully established an annual forum for researchers and practitioners to present and discuss advances in data mining techniques that primarily target biological data.

The 17th BIOKDD workshop (BIOKDD 2018) was held on August 20, 2018 in London, UK. A prominent change from previous BIOKDD workshops is that SIGKDD 2018 initiated a special theme day call Health Day including 3 workshops and BIOKDD 2018 was one of them. The SIGKDD Health Day promotes the idea of combining Healthcare with Data Science, and encourages the interaction among SIGKDD attendees in health-related fields with joint events.

To better fit the Health Day's theme on Data Science, where healthcare applications should benefit from analytics over a diversified source of both biological data and clinical data, BIOKDD 2018 called for contributions that also target clinical data besides biological data. This turned out to be a big success, where BIOKDD 2018 brought together a broader audience in Biomedical Sciences and Health Informatics to introduce their research and to exchange their ideas. All BIOKDD attendees gave their poster presentation in the Health Day's poster session in addition to oral presentation of their research within the workshop, and 5 quality posters were selected for presentation in SIGKDD main conference's poster session to allow exposure to the broader audience of SIGKDD. Our general chair Jake was also on the Health Day's expert panel discussing the impact of AI and machine learning in making healthcare more predictive, preventive and proactive; Jake also gave the BIOKDD keynote talk entitled "Network medicine: from multi-omics analysis to drug repositioning".

Altogether 11 submissions were accepted by BIOKDD 2018 featuring analytics over diversified data sources including medical images and videos, EEG and EOG signals, case narratives, social media data, in addition to the various biological data. Another interesting observation is that 5 out of the 11 accepted works adopt deep learning, including novel applications over biological data in addition to the more straightforward applications over clinical data.

Given these observations, BIOKDD will continue to accept papers from clinical data mining, and the excitement about deep learning techniques is expected to continue.

Special Issue Overview. This special issue of *TCBB* features the extended versions of 4 quality papers presented in BIOKDD 2018. In addition, the special issue also invited 3 papers extended from the quality papers presented in the DMBIH 2018 workshop co-organized by our organizer Samah. The 6h Workshop on Data Mining in Biomedical Informatics and Healthcare (DMBIH 2018) was held on November 17, 2018, in Singapore, and provided a forum for data miners, informacists, data scientists, and clinical researchers to share their latest investigations in applying data mining techniques to biomedical and healthcare data. The theme of both workshops align well with each other.

Each of the 7 invited papers was reviewed by at least 2 additional reviewers invited by the *TCBB* guest editors and the workshop reviews were shared with the new reviewers. The papers also went through 1 to 2 rounds of revisions.

The first paper invited from BIOKDD 2018, "Deep learning for automated feature discovery and classification of sleep stages," by Michael Sokolovsky, Francisco Guerrero, Sarun Paisarnsrisomsuk, Carolina Ruiz, and Sergio A. Alvarez [1] explores the use of a deep CNN architecture for automated sleep stage classification of human sleep EEG and EOG signals, which is useful for the diagnosis of sleep disorders. The task currently relies on highly trained human technicians, making the process time-consuming, and yields results that are subject to error, subjectivity and variation. In this study, they demonstrated that the performance gains achieved by their network rely mainly on network depth rather than using more signal channels. Also, performance of their approach is on par with human expert inter-scorer agreement. Finally, they examined the internal activation levels of their CNN and found that it spontaneously discovers signal features such as sleep spindles and slow waves.

The second paper invited from BIOKDD 2018, "Deep learning benchmarks on L1000 gene expression data," by Matthew McDermott, Jennifer Wang, Wen Ning Zhao, Steven D. Sheridan, Peter Szolovits, Isaac Kohane, Stephen J. Haggarty, and Roy H. Perlis [2] fills the gap of lacking published benchmark tasks and well-characterized baselines for testing and comparing machine learning (esp. deep learning) models over gene expression data. They defined 3 biologically motivated benchmarking tasks over 2 curated views of the public L1000 LINCS dataset as well as 1 privately produced gene expression dataset. On each task, they profiled k nearest neighbor classifiers, decision trees, random forests, linear classifiers, and two neural clas-

sifiers: feed-forward artificial neural networks (FF-ANNs) and graph convolutional neural networks (GCNNs). They found that GCNNs can be highly performant with large datasets as GCNNs utilize the regulatory relationships between pairs of genes, while FF-ANNs consistently perform well. Among non-neural classifiers, linear models and KNN classifiers dominate.

The third paper invited from BIOKDD 2018, "A data-driven approach to predict and classify epileptic seizures from brain-wide calcium imaging video data," by Jingyi Zheng, Fushing Hsieh and Linqiang Ge [3] studies the prediction of epileptic seizures using the calcium imaging video data which images the whole brain-wide neurons activities with electrical discharge recorded by calcium fluorescence intensity (CFI). Using the zebrafish's brain-wide calcium image video data, they proposed a data-driven approach to effectively detect the systemic change-point, and to further predict the epileptic seizures. Unlike the previous two works, this work did not use deep learning; rather, they explored the macroscopic patterns of epileptic and control cases, and extracted features based on the pattern difference for constructing prediction models. Their results showed that their approach could effectively predict the time range of future epileptic seizure. They also proposed a new method to discretize related features, and combined with hierarchical clustering to better visualize and explain the pattern difference between epileptic and control cases.

The fourth paper invited from BIOKDD 2018, "Computerized classification of prostate cancer gleason scores from whole slide images," by Hongming Xu, Sunho Park, and Tae Hyun Hwang [4] presented an automatic technique for Gleason grading of tumor patterns which is one of the most powerful prognostic predictors in prostate cancer. They used H&E stained whole slide pathology images, and divided an image into a set of small image tiles; salient tumor tiles with high nuclei densities were selected to extract texture features that characterize different Gleason patterns for training a multi-class support vector machine (SVM) classifier that labels patient slides with different Gleason scores. Their experiments showed that their approach achieved superior performances over state-of-the-art texture descriptors as well as deep learning models for prostate cancer Gleason grading.

The first paper invited from DMBIH 2018, "Post-structuring radiology reports of breast cancer patients for clinical quality assurance," by Shreyasi Pathak, Jorit van Rossen, Onno Vijlbrief, Jeroen Geerdink, Christin Seifert, and Maurice van Keulen [5] studies the problem of quality assurance of patient reports to ensure that the clinicians conform to the protocols set by hospitals. They presented a machine-learning-based natural language processing (NLP) system for automatic quality assurance of radiology reports on breast cancer. Their approach first identifies the top-level structure (headings) from free-text reports and classifies the report content into the top-level headings, and then automatically structures reports using the BI-RADS standard. Top level structure and content of reports were predicted with an F1 score of 0.97 and 0.94, respectively, using Support Vector Machine (SVM) classifiers; while for automatic structuring, their proposed hierarchical Conditional Random Field (CRF) outperformed the baseline CRF with an F1 score of 0.78 vs 0.71. The determined structure of a report is represented in semi-structured XML format of the free-text report to facilitate visualizing the conformance of the findings to the protocols.

The second paper invited from DMBIH 2018, "Extracting inter-sentence relations for associating biological context with events in biomedical text," by Enrique Noriega-Atala, Paul Douglas Hein, Shraddha Satish S. Thumsi, Zechy Wong, Xia Wang, Sean Michael Hendryx, and Clayton Thomas Morrison [6] studies the problem of identifying biological contexts in biomedical texts and associating them with biochemical events described in texts. They cast the problem as an inter-sentential relation extraction problem where related entities can be separated by a significant distance. They presented a new corpus of open access biomedical texts that have been annotated by biology subject matter experts to highlight context-event relations, where they focus on biological context as descriptions of the species, tissue type and cell type that are associated with biochemical events. Using this corpus, they evaluated several classifiers for context-event association along with a detailed analysis of the impact of a variety of linguistic features on the classifier performance. They found that gradient tree boosting performs by far the best.

The third paper invited from DMBIH 2018, "Classification of patients with coronary microvascular dysfunction," by Samah Fodeh, Taihua Li, Haya Jarad, and Basmah Safdar [7] studies the leverage of structured and unstructured narratives in clinical notes to detect patients with coronary microvascular dysfunction (CMD), which is a major cause of ischemia but very challenging to diagnose due to lack of CMD-specific screening measures. Using machine learning, they have shown that structured data are not sufficient to detect CMD and integrating unstructured data in the computational model boosts the performance significantly.

Acknowledgement. As guest editors of this special issue, we would like to thank the contributing authors, BIOKDD 2018 and DMBIH 2018 program committee, the TCBB reviewers who reviewed papers in this special issue, and the TCBB staff for the supported to make this special issue possible.

REFERENCES

[1] Michael Sokolovsky, Francisco Guerrero, Sarun Paisarnsrisomsuk, Carolina Ruiz, and Sergio A. Alvarez. "Deep learning for automated feature discovery and classification of sleep stages," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

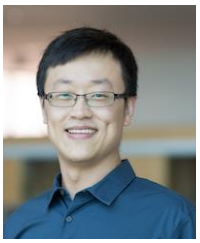
[2] Matthew McDermott, Jennifer Wang, Wen Ning Zhao, Steven D. Sheridan, Peter Szolovits, Isaac Kohane, Stephen J. Haggarty, and Roy H. Perlis. "Deep learning benchmarks on L1000 gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[3] Jingyi Zheng, Fushing Hsieh and Linqiang Ge. "A data-driven approach to predict and classify epileptic seizures from brain-wide calcium imaging video data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[4] Hongming Xu, Sunho Park, and Tae Hyun Hwang. "Computerized classification of prostate cancer gleason scores from whole slide images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

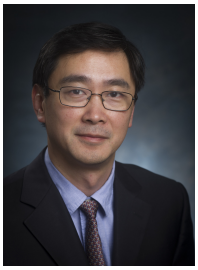
[5] Shreyasi Pathak, Jorit van Rossen, Onno Vijlbrief, Jeroen Geerdink, Christin Seifert, and Maurice van Keulen. "Post-structuring radiology reports of breast cancer patients for clinical quality assurance," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.
[6] Enrique Noriega-Atala, Paul Douglas Hein, Shraddha Satish S. Thumsi, Zechy Wong, Xia Wang, Sean Michael Hendryx, and Clayton Thomas Morrison. "Extracting inter-sentence relations for associating biological context with events in biomedical text," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.
[7] Samah Fodeh, Taihua Li, Haya Jarad, and Basmah Safdar. "Classification of patients with coronary microvascular dysfunction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

Da Yan is currently an Assistant Professor at the Department of Computer Science, the University of Alabama at Birmingham. He is the sole winner of Hong Kong 2015 Young Scientist Award in Physical/Mathematical Science. Dr. Yan regularly publishes in 1st-tier conferences and journals like SIGMOD, PVLDB, SIGKDD, ICDE, WWW, TKDE, TPDS, SoCC, EuroSys, PPoPP, etc. He also regularly serves as the reviewers of top journals including TODS, VLDBJ, TKDE, TPDS, etc., and serves in the program committees of top conferences such as SIGMOD 2019 and 2020, PVLDB 2018, IJCAI 2017, ICPP 2018, etc. Dr. Yan's research focuses on Big Data and Data Science.



Xin Gao is an Associate Professor, and group leader of Structural and Functional Bioinformatics (SFB) group in Computational Bioscience Research Center, and in Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST). He received the BS and PhD degrees in computer science from Tsinghua University, China, and University of Waterloo, Canada, respectively. Before joining KAUST, he was a Lane Fellow in Lane Center for Computational Biology at Carnegie Mellon University, US. His research focuses on the intersection between computer science and biology. His group works on building computational models, developing machine learning techniques, and designing efficient and effective algorithms to solve key open problems along the path from genome-scale sequence analysis, to protein structure prediction/determination, to function annotation, to understanding and controlling molecule behaviors in complex biological systems, and to biomedicine and healthcare. He has published more than 170 articles in leading journals and conferences in the fields of bioinformatics and machine learning.

Samah J. Fodeh is an Assistant Professor in the Department of Emergency Medicine and the Yale Center for medical Informatics at the Yale School of Medicine. She is also affiliated with the Veterans Administration (VA) Connecticut Healthcare Systems. She has a wealth of experience developing algorithms that exploit complementary data modalities to enhance knowledge discovery. Dr. Fodeh's research focuses on developing unsupervised machine learning methods to transform big clinical text data into a structured representation useful for subsequent analysis such as prediction and visualization. Dr. Fodeh has experience working with social media data as well. She published articles that leverage data from social media outlets to understand more about critical healthcare problems such as suicide and opioid overdose events. At the VA Center of Innovation Pain Research, Informatics, Medical co-morbidities, and Education (PRIME) Center, she investigates different approaches to improve understanding of the complex interactions between pain, opioids, and associated chronic disease and behavioral health factors and to develop efficacious interventions. Dr. Fodeh obtained her training with a Ph.D. as a computer scientist from Michigan State University.



Jake Y. Chen is the Chief Bioinformatics Officer and a tenured Professor of Genetics, Computer Science, and Biomedical Engineering at the University of Alabama at Birmingham. Previously, he was the founding director of Indiana Center for Systems Biology and Personalized Medicine. He has over 20 years of R&D experience in biological data mining and systems biology, with over 150 peer-reviewed publications. He is currently President-elect of the Midsouth Computational Biology and Bioinformatics Society. He also serves on the editorial boards of BMC Bioinformatics and JAMIA. He was recently listed as "Top 100 AI Leaders in Drug Discovery and Healthcare".