HealthManagement.org
The Journal
LEADERSHIP • CROSS-COLLABORATION • WINNING PRACTICES
VOLUME 20 • ISSUE 6 • 2020 • € 22
ISSN = 1377-7629

Cover Story

Smart Diagnostics

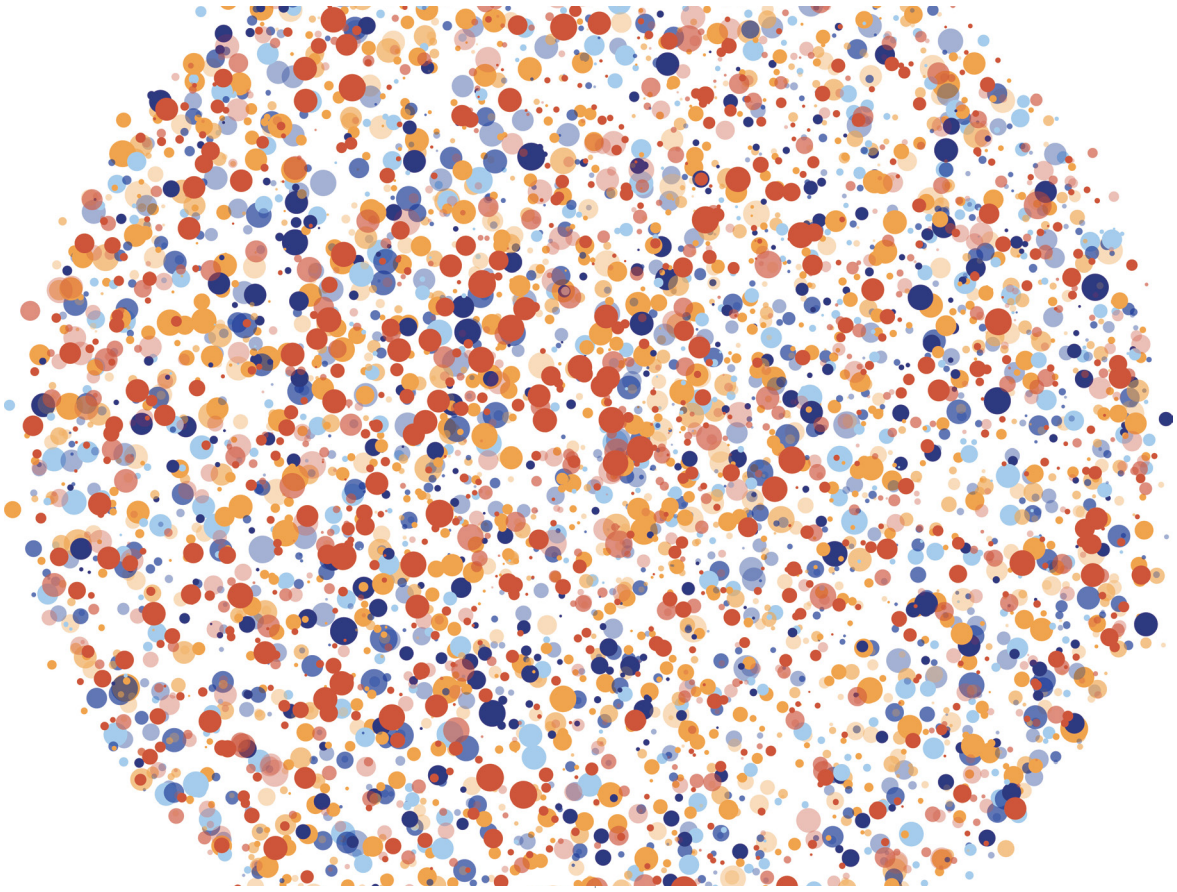
Deep Mining for Determining Cancer Biomarkers

Interviewee: Abeer Alzubaidi | School of Science and Technology | Nottingham Trent University | Nottingham, UK

Interviewee: Jonathan Tepper | CEO | Percepetronix Ltd. | UK

Interviewee: Ahmad Lotfi | Professor of Computational Intelligence | Nottingham Trent University | Nottingham, UK | Visiting Professor | Tokyo Metropolitan University

The use of omics data for knowledge discovery is an approach that can be used for personalised cancer medicine and for a better understanding of cancer genotype and phenotype. Three researchers have developed a deep feature learning model to discover biomarkers that are positively and negatively associated with cancer. HealthManagement.org spoke to Abeer Alzubaidi, Jonathan Tepper and Ahmad Lotfi to find out more about this new approach and its potential.



What do you think are the most significant limitations in harnessing omics data in the biomedical space?
Extracting knowledge from omics datasets is a serious challenge for the research community interested in understanding the cancer genotype and phenotype. Such datasets are characterised by high dimensionality and relatively small sample sizes with small signal-to-noise ratios. This significantly challenges existing machine learning-based solutions due to the so-called 'curse of dimensionality' where the addition of new input features typically requires an exponential number of input observations (which are commonly unavailable) to discover the underlying structure of the data that allows these models to generalise well to unseen cases. This also puts great pressure on data mining models that attempt to separate the signal from the noise in a bid to discover robust determinants.

Describe how the non-linear sparse auto-encoders in your deep learning model work?
The Sparse Compressed Auto-Encoder (SCAE) is simply a feedforward neural network trained with a variant of back-propagation to reproduce its input signal on its output layer, resulting in a hidden or latent feature layer of neurons representing the underlying transformation performed. The principle idea behind our SCAE model is to transform the original high dimensional omics data into a reduced feature space so that enough of the interesting complexity can be retained whilst not requiring additional observations to further constrain the model. This reduced description of the omics data is further realised through a regularisation technique within SCAE that maximises the likelihood of retaining important input signals describing much of the variance within the data, whilst filtering out the less important and noisy signals.

The Stacked Sparse Compressed Auto-Encoder (SSCAE) is composed of a sequence of SCAE trained in a dependent and co-operative manner, where the hidden feature layer of one model feeds as input to another. The underlying complexity of omics data is compactly represented with multiple levels of abstraction, therefore, we apply a greedy recursive approach to transforming the input signals containing tens of thousands of genes into a hidden representation of a lower dimension and higher abstraction, which is then provided as input to another SCAE, which encodes this further at a higher abstract level and so on. The resulting abstract hidden layer is then provided as input to the final layer of SSCAE (i.e. the output layer), which is a softmax classification layer trained to classify the input as belonging to either a patient with or without cancer.

In addition, we augmented a novel weight interpretation feature into SSCAE such that we were able to determine which original features on the input layer were most highly predictive, positively and negatively associated with the positive patient groups e.g. cancer, ER+/PR+. This method is simply based on computing the integrated weight score for each gene within the original input data that indicates its contribution to the latent representations formed within SSCAE during learning. This expands our deep learning model to include a feature selection method in addition to the feature extraction capacity already inherent within this paradigm. As a result, two smaller subsets of robust molecular markers are produced, one corresponding to those genes that are highly expressed for most of the patients from the positive group compared to the negatives; and the other subset refers to those genes that are highly expressed for most of the samples in the negative group compared to the positives. These subsets of robust biomarkers are then validated by training an independent classifier, such as a Support Vector Machine (SVM), to construct highly accurate classifier systems.

In what genotype and phenotype scenarios have you implemented your model and what outcomes have you detected?
It is well-known that much more accurate machine-learning methods are required to specify and measure phenotypes of complex diseases such as cancer. In particular, our focus has specifically been to reduce the amount of spurious or false positive associations within sophisticated classifier-based systems by intelligent feature selection and extraction. Moreover, if it is possible to identify robust biomarkers for cancer this will help standardise the definition of the disease to facilitate the interpretation and reproducibility of methods and results. However, we recognise this is against the challenging backdrop of data samples that are of very high dimensionality and relatively low sample sizes.

We utilised proteomic and genomic data sets to discover the phenotypes that underlie the variations apparent between the cancer and control patient groups. Fundamentally, two types of outcomes were revealed by our deep mining model, both indicating strong likelihoods of a patient having cancer. The first outcome indicated a subset of highly positively-weighted genes whereby the amplifications and gains in the gene expression levels were associated with the likelihood of a patient having cancer. Conversely, the second outcome revealed another subset of genes that were highly negatively-weighted and coincided with significant downregulation in the gene expression levels, and again indicated the strong likelihood of a patient having cancer.

How has your model ameliorated existing models of cancer biomarker identification?
As mentioned earlier, extracting knowledge from omics datasets is a serious challenge for machine learning-based solutions due to the 'curse of dimensionality.' Whilst some existing deep learning (neural network models with many hidden layers) approaches appear able to handle 'curse of dimensionality' issues and improve generalisability, this is typically at the expense of long training times, a need for substantial data to train the models, and lack of transparency in that it is not able to unambiguously state which input features are responsible for its behaviour.

To alleviate the limitations of existing approaches, we introduced SSCAE, a deep feature mining model with an explanatory technique that can be used for discovering robust high-level abstract representations from high dimensional small sample size omics datasets and reveal key determinants underlying these latent representations. Unlike other systems, SSCAE can perform deep classification whilst simultaneously revealing the key input features underlying its hidden representations. SSCAE's output decisions were further validated using appropriate evaluation metrics and independent model validations, thus providing significant confidence as to the relevance, robustness, and reproducibility of the discovered biomarkers.

How do you think medicine and research could collaborate more efficiently and effectively for better diagnosis of cancer?

A significant obstacle for biomarker discovery research remains the need for more effective interdisciplinary research environments, involving academics, clinicians and government working in a co-ordinated and prioritised manner. There are relatively few examples of situations where novel omics biomarkers originating from the cancer research community has found its way into routine clinical practice. Effective interdisciplinary research is therefore paramount if findings from state-of-the-art machine learning research is to be truly exploited and brought into the service of precision medicine e.g. data scientists should have clear routes of access to clinicians when evaluating genes identified by their machine learning methods. Similarly, clinicians must have fluid access to data scientists and bioinformaticians when requiring solutions to real-world problems, such as understanding the genetic make-up of new and emerging diseases or pandemics e.g., viruses such as COVID-19 and Ebola.

Based on the outcomes of your research, what do you think is the next stage? Is there scope to take your research further?

Moving forward, we will investigate the capacity of SSCAE to detect generic biomarkers for selected cancers across a range of independent high-quality genomic samples collected from different studies. This will further add confidence to the significance of the generic biomarkers already discovered by SSCAE and indicate which of these the academic and wider biomedical community should explore further. Also, with the rise of high quality integrated and multi-modal omics data, such as the TCGA database which contains a combination of genomic, epigenomic, proteomic, imaging and clinical data for matched patient groups, will enable us to develop sophisticated 'integrative models' that may reveal even more valuable indicators of disease. We feel this will provide a sound basis for the development of more effective diagnostic and prognostic systems in the future.

Extracting knowledge from omics datasets is a serious challenge for machine learning-based solutions due to the curse of dimensionality

How do you see your model being used in a real-world setting?

SSCAE could be realised as an essential software tool for bioinformaticians and clinicians. Bioinformaticians would use SSCAE for research and development purposes, evaluating various panels of genetic biomarkers for an array of different diseases with our deep mining model providing state-of-the-art analytics to further their research. Clinicians would use SSCAE optimised for specific diseases so that they can interrogate the software tool for biomarkers relating to specific cancers to establish whether or not patients have cancer and if so, to inform specific treatment patterns and protocols (subject to individual patient biomarkers being available to SSCAE).

Are there any particular cancers that your approach is most suitable for?

Our research provided much support for the strong association between gene expression and oestrogen and progesterone receptors and the development and treatment of breast cancer. However, our deep mining approach to feature extraction and selection is generic, and can therefore be applied to most, if not all cancers, where an underlying cause is believed to have a strong genetic component.

In our next paper, we will be presenting the outcomes of our experiments with our deep mining model for exploring the association between mRNA expression data and the positivity of both ER and PR receptors in breast cancer.

Conflict of Interest

None.

REFERENCES

Alzubaidi A, Tepper J, Lotfi A (2020) A Novel Deep Mining Model for Effective Knowledge Discovery from Omics Data. Artificial Intelligence in Medicine, 104, 101821.