Identification of Important Biological Pathways for Ischemic Stroke Prediction through a Mathematical Programming Optimisation Model —DIGS

Yongnan Chen
Dept of Informatics, Faculty of Natural & Mathematical Sciences, King's College London, London, UK
yongnan.chen@kcl.ac.uk

Konstantinos Theofilatos
School of Cardiovascular Medicine & Science, King's College London, London, UK
konstantinos.theofilatos@kcl.ac.uk

Lazaros G Papageorgiou
Dept of Chemical Engineering, Faculty of Engineering Science, University College London, London, UK
l.papageorgiou@ucl.ac.uk

Sophia Tsoka
Dept of Informatics, Faculty of Natural & Mathematical Sciences, King's College London, London, UK
sophia.tsoka@kcl.ac.uk

ABSTRACT
Stroke ranks second after heart disease as a cause of disability in high-income countries and as a cause of death worldwide. Identifying the biomarkers of ischemic stroke is possible to help diagnose stroke cases from non-stroke cases, as well as advancing the understanding of the underlying theory of the disease. In this study, a mathematical programming optimisation framework called DIGS is applied to build a phenotype classification and significant pathway inference model using stroke gene expression profile data. DIGS model is specifically designed for pathway activity inference towards supervised multi-class disease classification and is proved has great performance among the mainstream pathway activity inference methods. The highest accuracy of the prediction on determining stroke or non-stroke samples reaches 84.4% in this work, which is much better than the prediction accuracy produced by currently found stroke gene biomarkers. Also, stroke-related significant pathways are inferred from the outputs of DIGS model in this work. Taken together, the combination of DIGS model and expression profiles of stroke has better performance on the discriminate power of sample phenotypes and is capable of effective in-depth analysis on the identification of biomarkers.

CCS Concepts
•Applied computing→ Life and medical sciences→ Bioinformatics

Keywords
Acute ischemic stroke; Microarray; Gene expression profile; Biological pathway; Machine learning; Mathematical programming; MILP optimisation

1.INTRODUCTION
Acute ischemic stroke (AIS) is a dangerous disease worldwide, which has multiple complications and hard to cure [1]. According to [2], stroke is still the second leading cause of death and the third leading cause of disability after years of clinical treatment and basic researches. Also, it is well known that the economic costs of treatment and post-stroke care of stroke patients are substantial. Therefore, looking for an effective way for diagnostic or pathogenesis of AIS is important for both scientific researches and clinical practice [3].

Microarray technology has become a popular methodology in deriving comprehensive view from gene expression data of certain conditions. Based on the development of the microarray technology, several researches have identified molecular biomarkers from AIS blood samples [4]. However, most of these simple and efficient biomarker deriving approaches focused on independent genes and adopt basic statistical approaches. Therefore they were suffering from low prediction accuracy and the difficulty of biological interpretation. However, most of these simple and efficient biomarker deriving approaches treated genes independently and adopted basic statistical approaches. Therefore they were suffering from low prediction accuracy and the difficulty of biological interpretation [3, 5]. Following the principle that genes do not work isolated but work in concert, in recent years, independent gene editing therapeutic methods are increasingly replaced by simultaneously considering functional gene groups. Biological pathways are one of the representative kinds of these functional gene sets, which are available from public databases, for example, Reactome [6], Kyoto Encyclopedia of Genes and Genomes (KEGG) [7] and Gene Ontology (GO) [8]. Biological pathways provide the possibility of analysing groups of genes that belongs to same pathways and identifying the target-relevant pathways as biomarkers [5].

In [5], a novel multi-class disease classification method, Differential Gene Signature (DIGS), which infers pathway activity in a supervised manner, is proposed. DIGS is a MILP mathematical programming formulation that consists of a linear objective function and several linear constraints. The general idea of DIGS is using weighted linear summation of the constitute genes expression values from same pathway as the pathway activity evaluation of that sample, where the weights of constitute genes are decided by the optimisation model so that the

constructed pathway activity can optimally distinguish samples from different phenotype. DIGS has been tested on Psoriasis, Breast Cancer, Prostate Cancer and diffuse large B-cell lymphoma (DLBCL), and showed good performance on distinguishing their sub-phenotypes and detecting biomarkers of these diseases.

In this work, we apply DIGS on three acute ischemic stroke microarray profile datasets (Section 2), aiming to reach high prediction accuracy between stroke and non-stroke phenotypes and deriving relative functional pathways as biomarkers of AIS (Section 3).

2. METHODOLOGY

2.1. Data Acquisition and Preparation

Three publicly available gene expression profile datasets GSE22255 [4], GSE16561 [9] and GSE58294 [10] were obtained from Gene Expression Omnibus (GEO). All microarrays experiments of these three GEO series were conducted on Affymetrix Human Genome U133 Plus 2.0 Array Platform. GSE22255 [4] contains 20 stroke and 20 control (non-stroke) peripheral blood mononuclear cells (PBMCs); GSE16561 [9] contains 39 stroke and 24 control peripheral whole blood samples; GSE58294 [10] contains 23 control whole blood samples and 69 stroke samples. Because the 69 stroke samples of GSE58294 were analysed at three time points: less than 3 hours, 5 hours, and 24 hours following the onset of stroke, only samples collected at 3 hours time points were used in our work. In total, there are 82 stroke peripheral blood samples from stroke patients and 55 from control patients.

These three datasets have been combined, preprocessed and normalised according to the pipeline described in [3] in order to remove experimental biases, normalise based on sample type, filter genes, impute missing values and detect outliers.

Table 1. Processed Dataset Summary

Disease	Samples	Genes	Phenotypes
AIS	137	13243	Stroke: 82
			Control: 55

Pathways data were acquired from MsigDB [11] KEGG C2 functional gene sets, which include 186 curated pathways with in total 5267 genes.

2.2. Pathway Activity Inference: DIGS [5]

The overview of the computational procedure of DIGS, which is developed for pathway-based sample phenotype classification, is illustrated in Figure 1. Pathway specific gene expression matrices G is consists of the standardised gene expression values of sample s across gene m. The amount of gene expression matrices is equal to the amount of pathways. For each matrix G, DIGS model assigns pathway activity score for each sample and pathway activity range for each phenotype. The purpose of the model is to make as many pathway activity scores as possible fall into its corresponding phenotype range.

The mathematical programming based formulation of DIGS contains 10 constraints and an objective function. The first part of the formulations define how pathway activity values are calculated. For each sample s, pathway activity pa_s is defined as the summation of the gene expression values Gsm multiples gene weight $(rp_m\text{-}rn_m)$, where rp_m represents the positive weight of gene m and rn_m represents the negative weight. Then, the first limitation (set by the second and third constraints) is applied on a pair of

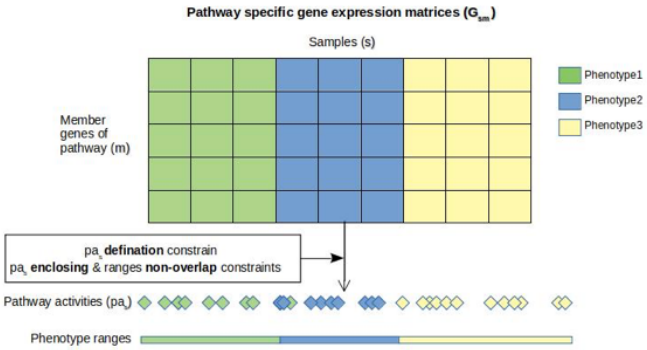

Figure 1. Schematic flow chart of DIGS pathway activity inference approach

positive variables, rp_m and rn_m. For each m, neither rp_m nor rn_m can take positive value, which means one of them is forced to be zero. Here a binary variable L_m is introduced to complete these constraints.

The second limitation (set by the fourth constraint) restricts on the number of genes that can be "active" genes among all member genes in a pathway. Active genes are defined as genes that gain non-zero weights, while keep the rest non-active genes' weights equal to zero. A binary variable W_m is introduced to indicate whether a gene m is "active". When W_m takes 1, the gene m is an active gene and its weight $(rp_m\text{-}rn_m)$ would be token between -1 and 1. Also, a user defined variable NoG (used in the fifth constraints) is introduced to restrict the maximum number of active genes. For normalization purpose, the summation of absolute gene weights is equal to 1, as defined as the sixth constraint.

The following part of formulations set restrictions on the phenotype ranges. According to the seventh and eighth constraints, the range for a phenotype c is defined by two continuous variables, lower bound LO_c and upper bound UP_c. A binary variable E_s is adopted to indicate whether the pa_s value of a sample s falls within the LO_c and UP_c of its corresponded phenotype.

Other last two constraints are introduced to guarantee that, for each pair of phenotypes (c, k), the ranges are not overlap. Here a binary variable Y_{kc} ensures this requirement. When $Y_{kc} = 1$, the relationship between c and k is $k<c$ and UP_k is lower than the LO_c; while $Y_{kc} = 0$ means the otherwise conditions ($c < k$, $UP_c < LO_k$). Also, ε, an arbitrarily small positive number, is designed to ensure pair-wise classes do not share borders.

Finally, the objective function of this optimization problem can be defined by minimising the number of miss-classified samples ($1 - E_s$).

In conclusion, all constraints of DIGS model are linear with a linear objective function and multiple binary or continuous variables. Therefore, DIGS is defined as a mixed integer linear programming (MILP) model that can be solved to reach global optimal with standard algorithms [5].

2.3. Implementation and Validation Scheme

The implementation procedure of pathway activity-based disease classification is illustrated in Figure 2. To gain robust and objective prediction results, all samples of stroke dataset are randomly split into 70% training set and 30% testing set. This procedure was repeated 10 times to produce 10 training/testing

sets. During model training process, testing samples are always blind to the training procedures to ensure no information leakage. For every training gene expression matrix of 10 training/testing sets, gene sets from KEGG pathways are integrated with the gene expression matrix to create individual pathway specific expression matrices. In total, 1860 pathway specific expression matrices are generated and DIGS models are trained on them. From model solving results, composite features, which summarise the expression patterns of member genes *m* into a new feature pa_s, are constructed for samples in a pathway matrix. The pathway activity value pa_s represents the activity levels or deregulation degrees of a pathway on samples. After profiling activity vectors from all pathway specific expression matrices independently, pathway activity matrices for each training set are formed by the ensemble of corresponded 186 pathway activity vectors. In the next step, classifiers are going to be trained on these pathway activity matrices.

In parallel, for each training/testing set and each pathway, gene weights (rp_m and rn_m) are extracted when solving DIGS models on training samples. Then the gene weights are applied to testing samples to construct pathway activity vectors for testing set. Similarly, pathway activity vectors from testing samples are combined into pathway activity matrix and classifiers are tested on it to produce prediction outputs.

Mathematical details of DIGS model used in this work is a reproduction of the model in [5], which also provides example input files and user guide at www.ucl.ac.uk/~uceclap/DIGS. The DIGS model is implemented by General Algebraic Modelling System (GAMS) [12] using the CPLEX MILP solver. According to the sensitivity analysis for parameter *NoG* in [5], DIGS model is robust with respect to *NoG* in range of 5 to 20. In this work, *NoG* is set as 10, which means allowing 10 genes per pathway to participate in pathway activity inference. The optimal gap is set as 0.00 for the attempting of getting globally optimal solutions. However, as the computation time limit for solving per DIGS model is set as 200 seconds by [5], the solving status of DIGS models includes both global optimal solution and optimal solution.

Overall, the above procedure produced 10 sets of training/testing pathway activity matrices corresponding to original random division of training/testing set on stroke dataset. Six commonly used machine learning classifiers, K-nearest-neighbour (KNN), Logistic Regression (LR), Random Forest, Neural Network (NN), Naive Bayes and Support Vector Machine (SVM), were employed in our study using Python package Sklearn version 0.22.1 to produce the classification accuracies on pathway activity matrices. Six classifiers were trained on 10 training pathway activity matrices and tested on testing pathway activity matrices with the following parameters: for NN, hidden layer is 2, learning rate is 0.1, training time is 10000; for KNN, the number of clusters is 5. For the other classifiers, default settings were retained.

3. RESULTS AND ANALYSIS

3.1. Evaluation of Prediction performance

To rigorously evaluate the prediction accuracy of various implemented classification approaches, prediction accuracy are averaged over 10 randomly developed 70% training sets and 30% testing sets for each classifier. Due to the inherent problem of unbalanced numbers of samples across two phenotypes (Table 1), both classification accuracy (ACC) and area under curve (AUC) [13] are used as metrics to measure the prediction accuracy of a classification model. Higher AUC values corresponded to better prediction performance, with AUC of 1 indicates perfect prediction, while 0.5 indicates the performance is equal to random. Overall, Table 2 shows the averaged ACC and AUC values across 10 testing sets produced by 6 different classifiers on stroke dataset.

Generally, all six classifiers have produced relatively high accuracies (~83.5%) and high AUC scores (~91.5%) towards the classification on stroke and control samples. Among six

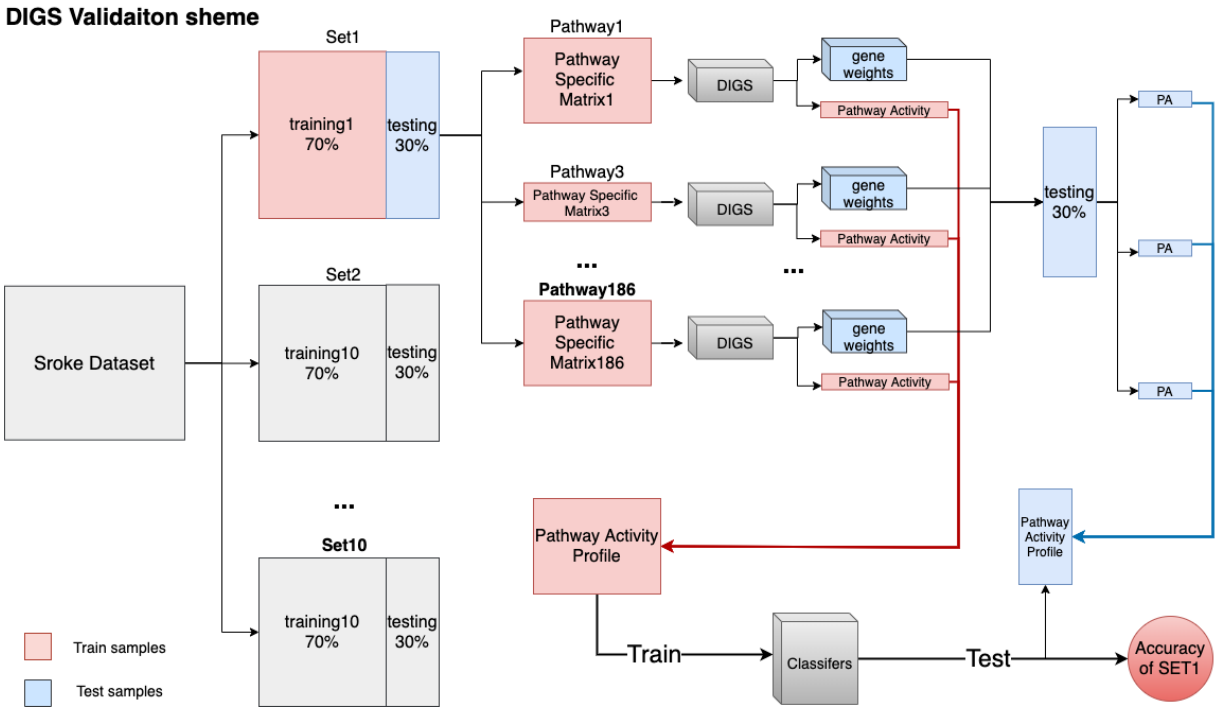

Figure 2. Overview of DIGS validation scheme from Microarray gene expression Profile to Phenotype Classification

Table 3. Significant pathways

	Pathway Name	Coef.
1	CELL_CYCLE	0.782
2	B_CELL_RECEPTOR_SIGNALING_PATHWAY	0.744
3	UBIQUITIN_MEDIATED_PROTEOLYSIS	0.743
4	LEISHMANIA_INFECTION	0.739
5	PYRIMIDINE_METABOLISM	0.739
6	SPLICEOSOME	0.737
7	CELL_ADHESION_MOLECULES_CAMS	0.736
8	RNA_DEGRADATION	0.727
9	TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	0.725
10	EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.721

classification methods, 5-Nearest-Neighbours reached the highest prediction accuracy (84.4%) and Logistic Regression got the best AUC score (93.2%).

In order to further validate the superiority of DIGS, other three widely used pathway activity inference methods were implemented on stroke dataset for comparison. In overview, these three methods are: i) Mean method [14] that take the mean gene expression values of all genes within a pathway for each sample. More specifically, Mean method derives the pathway activity vector from the pathway specific matrix by calculation the mean expression values across all member genes for each sample; ii) the second method, referred as Median method [15], has exactly same procedure as Mean method, only by replacing the mean expression values across genes with the median expression values across genes; and iii) the third method is called PCA method, built by [16], which uses the first principal component of the pathway specific expression matrix as representation of pathway activity scores for each sample. To make the prediction results comparable, the validation scheme for these other three pathway

Table 2. Mean Prediction accuracy of Stroke dataset

Classifier	ACC	AUC
5-NN	0.844	0.895
Logistic Regression	0.833	0.932
Random Forest	0.836	0.923
Neural Network	0.840	0.917
SVM	0.830	0.927
Naive Bayes	0.843	0.907

activity inference method is same as DIGS. The ten training/ testing sets used for DIGS were applied to Mean, Median and PCA method too. The arrangement of the resulting 10 pathway activity matrices and same classifier training procedures were adopt. The output prediction accuracy for theses three methods were also averaged across 10 testing sets and all results are plotted in Figure 3.

In Figure 3, x-axis is labeled with six classification approaches and y-axis represents the prediction accuracy values for each pathway activity inference methods across each classifier. From the figure, it is obvious that DIGS-based classification approach achieves higher classification rates than other pathway inference methods. The performance of Mean and Median methods are similar (accuracies range form 60% to 80%), and PCA methods gets the lowest prediction accuracies (range from 50% to 60%). It can be concluded that DIGS is the most effective method among

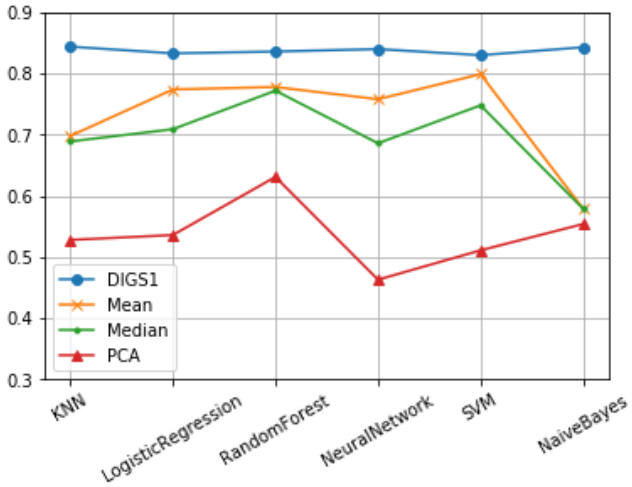

Figure 3.Classification accuracy comparison of four pathway activity inference methods

four methods for deriving pathway activity score towards phenotypes detection.

3.2. AIS Relevant Pathway Identification

3.2.1. Pathway Relevance Ranking

Not only promising classification rates can be achieved by DIGS model, but also a number of pathways are identified that may indicate pathway biomarkers. To rank the pathways, Point-biserial correlation coefficient ranking method in Python SciPy package (Version 1.3.0) is employed in this work.

To gain an ultimate pathway activity value for each pair of sample and pathway, 10 pathway activity matrices (combination of the corresponded training samples and testing samples of each training/testing set) were merged into one pathway expression matrix by averaging operation. Then, for each pathway , the Point-biserial Correlation Coefficient is calculated using the pathway activity vector across all samples and the phenotype vector that consists of sample phenotypes (stroke or control). Point-biserial Correlation Coefficient is a statistical measure of the relationship between a binary variable and a continuous variable, and it is mathematically equivalent to the Pearson correlation. In Machine Learning, Point-biserial correlation can be used to calculate the similarity between features and categories. In other words, it is adopt to judge whether the extracted features are positively correlated, negatively correlated or not correlated with the responded categories. The range of Point-biserial Coefficient is [-1, 1] and the greater the absolute value is, the stronger the correlation is. Therefore, the absolute value of the calculated correlation metrics for each pathway were ranked in descending order and top 10 pathways were selected as the most discriminative pathways.

The selected ten discriminative pathways are listed in Table 3. Apart from pathways that have obvious links to cancer pathways, for example the well-known signalling pathway (B cell receptor signalling pathway), and pathways involve in the cell metabolism procedures and genetic information processing (Cell cycle, Pyrimidine metabolism, RNA degradation and Spliceosome), we note a research concluded that the immunoblockade or genetic deletion of adhesion molecules showed to reduce infarct volume, edema, behavioural deficits and/or mortality in different animal models of ischemic stroke [17]. Also, [18] indicates that the ubiquitin-mediated proteolysis pathway, especially TRAF6, may be the most vital molecules among TLR downstream pathways in

incidences of ischemic stroke, which proves that two of our top 10 pathways (Ubiquitin mediated proteolysis and Toll-like receptor signalling pathway) are strongly related to the diagnosis of ischemic stroke.

Accordingly, these DIGS found top ranked pathways are highly related to AIS, and thereby can be treated as pathway biomarkers. More amount of significant pathways can be inferred by DIGS and their relationships with AIS are worth being studied in further researches.

3.2.2. Top Ranked Pathways Evaluation

To intuitively display the performance of significant pathways, two heat-maps were drawn using gene expression data and significant pathway activity data. In Figure 4 and Figure 5, rows are gene names and pathway names respectively and columns are sample phenotypes. Samples are hierarchically clustered based on similarity. In horizontal colour bar, different colours (green and blue) represents 'stoke' and 'control' respectively. From the comparison between Figure 4 and Figure 5, it is clear that most of the samples belonging to the same phenotype outcomes are indeed assigned into same clusters with the significant pathways data found by DIGS (compared with the randomly distributed sample clusters in Figure 4). This phenomenon is also consistent with the good prediction performance of DIGS inferred pathway activity that is illustrated in section 3.2.1.

To Further explain to what extent the significancy has been reached by the top ranked pathways, 10 box plots in Figure 6 show the distribution of pathway activity values of the two different sample phenotypes for each significant pathway. The left box plot in each subplot represents the activity values of 'stoke' samples and the right box plot represents for the 'control' samples. It is obvious that the ranges between upper quartiles and lower quartiles of the two phenotypes are perfectly separated in each subplot. That is to say, even using only one of theses top discriminate pathway to classify the phenotypes, the accuracy would be at least 75%. Similar analysis was done in [3] too, where the 10 most statistically significant differentially expressed genes, illustrating the extent of Fold Change and the separation of median expression levels between the two phenotypes, were selected from the stroke dataset and their logarithmic relative expression values were plotted. The box plots of these significant differentially expressed genes in [3] presents the distribution of the expression ranges of 'stroke' and 'control' phenotypes. However, the differences with the Figure 6 is obvious because the ranges between upper quartiles and lower quartiles of the two phenotypes in these box plots of the selected differential expressed genes are overlap, which indicates the prediction accuracy of the significant genes are less than the top ranked pathways. Also, according to their study, the prediction accuracy produced by log FC expression level significance method, which employed 557 genes, is less than 71%.

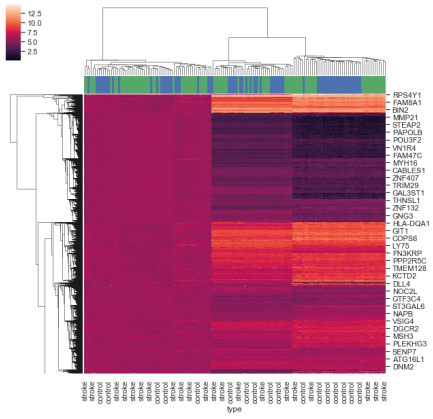

Figure 4. Heatmap: gene expression profile of Stroke dataset

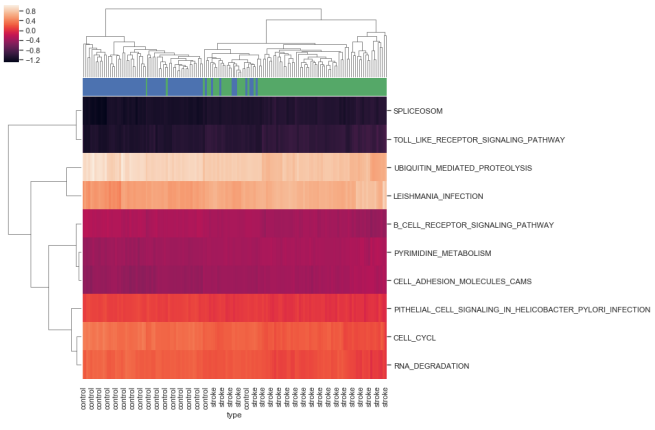

Figure 5. Heatmap: Significant pathway activity of Stroke dataset

It can be concluded that DIGS found pathways have stronger discriminate power on separating stroke samples from non-stroke samples than current found significant genes biomarkers of stroke, and also these results have proven the idea that the way of regarding genes belong to a same functional group as a whole for the analysis towards phenotypes on gene expression profiles is better than using single genes independently [16].

4. CONCLUSION

This work applies a mathematical programming optimisation method DIGS on three stroke gene expression profiles for the purpose of inferring pathway activity values for acute ischemic stroke samples. The prediction results towards stroke phenotypes gave promising accuracy rates (~83.5%) and relatively high AUC values (~91.5%) on testing sets. To authors' best knowledge, the classification accuracy reached by DIGS is higher than most current stroke phenotype prediction researches. Also, DIGS model identified a mount of biological pathways that are proved related to the cause of AIS can be seen as pathway biomarkers of AIS.

One of the improvements, compared with former gene-based stroke studies, is incorporating biological pathways with gene expression profiles. The advancement of combining pathways with gene expression profiles is approved by the higher classification accuracy of separating stoke samples and non-stroke samples in this work. Besides, DIGS provides a more flexible way for inferring stroke related biomarkers. By modifying the participate pathways for constructing pathway specific expression matrix or changing the number of active genes inside DIGS model, different aspects and levels of biological information can be extracted from the running outputs.

5. REFERENCES

[1] Wang, P.L., Zhao, X.Q., Yang, Z.H., Wang, A.X., Wang, C.X., Liu, L.P., Wang, Y.L., Wang, X.G., Yi, J.U., Chen, S.Y. and Chen, Q.D., 2012. Effect of in-hospital medical complications on case fatality post-acute ischemic stroke: data from the China National Stroke Registry. Chinese medical journal, 125(14), pp.2449-2454.

[2] Johnson, C.O., Nguyen, M., Roth, G.A., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H.N.,

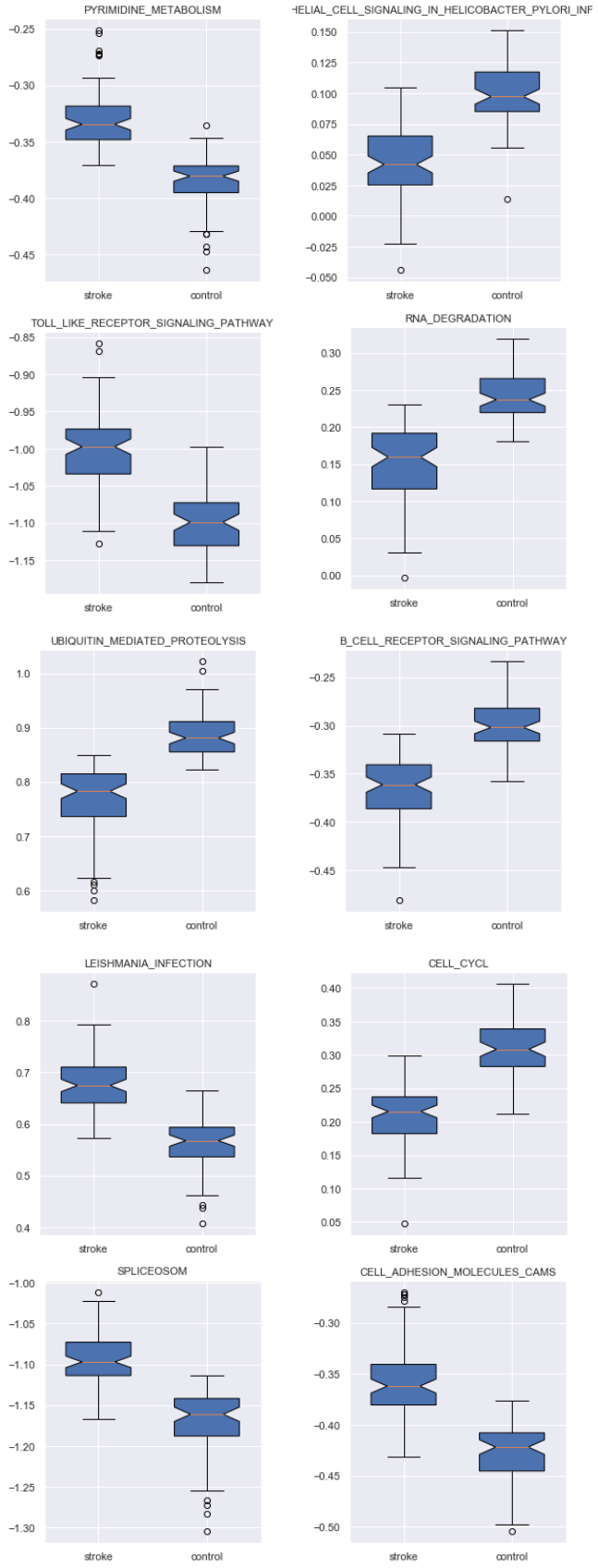

Figure 6. Boxplots: Ranges of pathway activity of different phenotypes for top pathways

Abu-Rmeileh, N.M. and Adebayo, O.M., 2019. Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(5), pp.439-458.

[3] Theofilatos, K., Korfiati, A., Mavroudi, S., Cowperthwaite, M.C. and Shpak, M., 2019. Discovery of stroke-related blood biomarkers from gene expression network models. *BMC medical genomics*, *12*(1), p.118.

[4] Krug, T., Gabriel, J.P., Taipa, R., Fonseca, B.V., Domingues-Montanari, S., Fernandez-Cadenas, I., Manso, H., Gouveia, L.O., Sobral, J., Albergaria, I. and Gaspar, G., 2012. TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. Journal of Cerebral Blood Flow & Metabolism, 32(6), pp.1061-1072.

[5] Yang, L., Ainali, C., Tsoka, S. and Papageorgiou, L.G., 2014. Pathway activity inference for multiclass disease classification through a mathematical programming optimisation framework. BMC bioinformatics, 15(1), p.390.

[6] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L. and Lewis, S., 2005. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, *33*(suppl_1), pp.D428-D432.

[7] Kanehisa, M., 2002, January. The KEGG database. In *Novartis Foundation Symposium* (pp. 91-100). Chichester; New York; John Wiley; 1999.

[8] Gene Ontology Consortium, 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, *32*(suppl_1), pp.D258-D261.

[9] Barr, T.L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A. and Matarin, M., 2010. Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology*, *75*(11), pp.1009-1014.

[10] Stamova, B., Jickling, G.C., Ander, B.P., Zhan, X., Liu, D., Turner, R., Ho, C., Khoury, J.C., Bushnell, C., Pancioli, A. and Jauch, E.C., 2014. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. *PloS one*, *9*(7).

[11] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), pp.1739-1740.

[12] GAMS Development Corporation, 2013. General Algebraic Modeling System (GAMS), rel. 24.2. 1.

[13] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), pp.1145-1159.

[14] Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., Zhu, J., Wang, H., Wang, C., Topol, E.J. and Wang, Q., 2005. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC bioinformatics*, *6*(1), p.58.

[15] Diao, H., Li, X., Hu, S. and Liu, Y., 2012. Gene expression profiling combined with bioinformatics analysis identify biomarkers for Parkinson disease. *PloS one*, *7*(12).

[16] Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. and Lee, D., 2008. Inferring pathway activity toward precise disease classification. *PLoS computational biology*, *4*(11).

[17] Yilmaz, G. and Granger, D.N., 2008. Cell adhesion molecules and ischemic stroke. *Neurological research*, *30*(8), pp. 783-793.

[18] Wu, D., Lee, Y.C.G., Liu, H.C., Yuan, R.Y., Chiou, H.Y., Hung, C.H. and Hu, C.J., 2013. Identification of TLR downstream pathways in stroke patients. *Clinical biochemistry*, *46*(12), pp.1058-1064.

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors' background

Your Name	Title*	Research Field	Personal website
Yongnan Chen	Phd candidate	Computational genome analysis, Optimisation model,	https://kclpure.kcl.ac.uk/portal/en/persons/yongnan-chen(0baed3c1-3dc6-482e-9322-06eae331dcac).html
Konstantinos Theofilatos	lecturer	Bioinformatics, Cardiovascular Informatics	https://kclpure.kcl.ac.uk/portal/konstantinos.theofilatos.html
Lazaros G Papageorgiou	Full professor	Process Systems Engineering, Systems Biology	https://www.ucl.ac.uk/chemical-engineering/people/prof-lazaros-papageorgiou
Sophia Tsoka	associate professor	Bioinformatics, Systems Biology	https://kclpure.kcl.ac.uk/portal/en/persons/sophia-tsoka(af6326b4-5a17-40f8-82ae-d4ab73260227).html

***This form helps us to understand your paper better, the form itself will not be published.**

***Title can be chosen from: master student, Phd candidate, assistant professor, lecturer, senior lecturer, associate professor, full professor, research, senior research**