Computational and Structural Biotechnology Journal 20 (2022) 5547–5563

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

# Particle swarm optimization artificial intelligence technique for gene signature discovery in transcriptomic cohorts

Ross G. Murphy [a], Alan Gilmore [b], Seedevi Senevirathne [a], Paul G. O'Reilly [c], Melissa LaBonte Wilson [a], Suneil Jain [a], Darragh G. McArt [a,*]

[a] Movember FASTMAN Centre of Excellence, Patrick G Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast BT9 7AE, UK
[b] Patrick G Johnston Centre for Cancer Research, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast BT9 7AE, UK
[c] Sonrai Analytics Ltd, Whitla Medical Building, Health Sciences Campus, Belfast BT9 7BL, UK

ARTICLE INFO

*Article history:*
Received 18 March 2022
Received in revised form 22 September 2022
Accepted 22 September 2022
Available online 26 September 2022

*Keywords:*
Particle Swarm Optimization
Artificial Intelligence
Machine Learning
Cancer
Transcriptomics
Biomarker Discovery

ABSTRACT

The development of gene signatures is key for delivering personalized medicine, despite only a few signatures being available for use in the clinic for cancer patients. Gene signature discovery tends to revolve around identifying a single signature. However, it has been shown that various highly predictive signatures can be produced from the same dataset. This study assumes that the presentation of top ranked signatures will allow greater efforts in the selection of gene signatures for validation on external datasets and for their clinical translation. Particle swarm optimization (PSO) is an evolutionary algorithm often used as a search strategy and largely represented as binary PSO (BPSO) in this domain. BPSO, however, fails to produce succinct feature sets for complex optimization problems, thus affecting its overall runtime and optimization performance. Enhanced BPSO (EBPSO) was developed to overcome these shortcomings. Thus, this study will validate unique candidate gene signatures for different underlying biology from EBPSO on transcriptomics cohorts. EBPSO was consistently seen to be as accurate as BPSO with substantially smaller feature signatures and significantly faster runtimes. 100% accuracy was achieved in all but two of the selected data sets. Using clinical transcriptomics cohorts, EBPSO has demonstrated the ability to identify accurate, succinct, and significantly prognostic signatures that are unique from one another. This has been proposed as a promising alternative to overcome the issues regarding traditional single gene signature generation. Interpretation of key genes within the signatures provided biological insights into the associated functions that were well correlated to their cancer type.
© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

Gene signature development towards prognostic and predictive abilities on cancer data types is crucial for delivering personalized medicine [1]. One of the most popular signatures includes 70 genes for breast cancer prognosis which has been approved for clinical use [2]. The decreasing cost associated with sequencing the gene expression profile of multiple patient clinical cohorts has led to what is being defined as the big data era within bioinformatics [3]. This is reflected in the amount of publicly available gene expression datasets from published research associated with clinical trial cohorts and massive databases such as the Gene Expression Omnibus (GEO) [4], ArrayExpress [5], and The Cancer Genome Atlas (TCGA) [6]. Amongst research using these datasets towards gene signature generation, there appears to be little overlap of associated genes and their associated biology [7].

Gene signature discovery tends to revolve around statistical methods for feature selection and machine learning for classification tasks. Due to the highly dimensional nature of biological omics data types, feature selection is a crucial step in identifying gene signatures. Additionally, they provide faster and more cost-effective models by using less features to consider in a classification task [8]. Feature selection methods are split into three categories which is based on how they select features in combination with the classification model. These include filter techniques, wrapper techniques, and embedded techniques. The most popular of these methods amongst gene expression microarray analysis is univariate filtering techniques. These methods are fast and easy to interpret individually as features are ranked based on a statisti-

* Corresponding author.
*E-mail address:* d.mcart@qub.ac.uk (D.G. McArt).

https://doi.org/10.1016/j.csbj.2022.09.033
2001-0370/© 2022 Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

cal metric such as a p-value. Wrapper and embedded feature selection techniques are an alternative to multivariate analysis which incorporates classification bias for greater accuracy. The wrapper approach evaluates how well each feature subset performs with the selected classifier [9]. Wrapper methods have been suggested as the superior approach when regarding predictive accuracy compared to filter methods [10]. Finally, embedded methods are classifiers that have the ability to discard input features for a more discriminative feature subset.

In microarray gene expression studies, wrapper methods tend to be dominated by evolutionary algorithms such as particle swarm optimization (PSO) and genetic algorithms (GA) as the search algorithm with a selection of different classifiers. Both PSO and GA are similar in that they are metaheuristics, evolutionary algorithms, and attempt to identify optimal solutions through iteratively improving a population of candidate solutions [11]. Various alternatives of PSO have been produced along its own evolution. This includes the likes of multi-objective PSO [12] and discrete or binary PSO (BPSO) [13]. In gene signature selection studies, BPSO is amongst the most popular of PSO alternatives. In BPSO, dimensions represent features and positions are binary bits that flip 0 or 1 to indicate a non-selection or selection of a feature, respectively. In gene signature selection, BPSO has also been used as feature selection for wrapper methods with classifiers such as k-NN [14], SVM [15], Naïve-Bayes [16], and decision trees [17].

Variants of BPSO have been proposed to help overcome some of the challenges that conventional BPSO faces. This includes enhanced BPSO (EBPSO) which was purpose built to select for small subsets of informative genes without losing accuracy performance for cancer classification [18]. EBPSO focuses on improving the limitations that are seen in conventional BPSO's sigmoid velocity and individual gene selection functions. Further investigation of these equations in traditional BPSO reveal some limitations in selecting for features to create an appropriate feature signature. Mohamad et al. demonstrated that the probabilities of a feature being selected, and it not being selected are the same at 0.5, or a 50/50 chance. Thus, conventional BPSO is primed to select for 50% of the input feature length which restricts its ability to produce succinct signatures out of the input features. This in turn could potentially have an effect on accuracy performance also. Conventional BPSO could be selecting for gene signatures that are accurate as a combination of informative features, redundant features, and non-informative features. EBPSO aims to overcome these limitations by introducing the scalar quantity called particle speed and modifying the sigmoid and updated position functions. These implementations were put in place to increase the probability of a feature to not be selected, and thus decrease the probability of the feature to be selected for. This would result in smaller feature signature lengths, whilst only selecting for the most informative features towards signature performance.

Major issues around popular methods for gene signature selection relate to identifying gene signatures that are not well validated on other clinical cohorts. This could be due to identifying gene signatures of interest that are specific to the characteristics of a given dataset, not of a given disease type. Determining the underlying biology driving the specific selection of genes to make up the signature is also hard to achieve [19]. Pathway analysis on gene signatures has been proposed to help overcome these issues by identifying genes within a signature related to a key biology. Additionally, gene signature discovery tends to revolve around identifying a single signature. This has been replicated previously with a single breast cancer dataset [20]. Various highly predictive signatures were produced from the same analysis explained through properties of the data. These included suggesting that many genes were correlated with the classification task, the differences between these correlations being small, and these correlations changing dramatically over different training data subsets. A possible reason for this has been suggested as a patient or sample having the potential to contain unique variations and heterogeneity, which in turn could affect markers for outcome [20].

This study assumes that the successful development of EBPSO will allow for the generation of gene signatures on transcriptomics cohorts that are both accurate and succinct in their number of associated features. Additionally, the presentation of top ranked signatures on their fitness values will allow for greater efforts in the selection of gene signatures for validation on external clinical cohorts and for their transition towards clinical use. Thus, this study aims to develop EBPSO on a PSO Python research module which employs conventional BPSO and validate the potential of unique candidate gene signatures with different underlying biology from EBPSO on selected transcriptomics cohorts.

## 2. Materials & methods

### 2.1. Simulated and publicly available clinical transcriptomics cohorts

It is critical that the performance of both EBPSO and BPSO in selecting for gene signatures within transcriptomics cohorts is evaluated appropriately. To adhere to this, simulated datasets were generated, and publicly available transcriptomics cohorts were selected for testing the two PSO algorithms. Simulated datasets allow for greater control over knowing what features are important towards gene signature classification and whether the PSO methods were able to accurately select for these. Additionally, simulated datasets also show what features are not important towards signature classification and whether the PSO techniques are able to not select for these.

Simulated datasets have been produced by using the scikit-learn Python library [21]. The make_classification utility within the scikit-learn datasets module allows for the generation of randomised classification datasets. Artificial datasets are created with control over the number of samples, features, informative features, and class separation. Class separation is the value that separated the informative features between the classes. Higher values thus resulted in more informative features between the classes and should provide greater accuracy for the given classification task. Two artificial datasets were generated, consisting of a binary and a multi-class gene expression simulated cohort. Both simulated datasets had 200 samples, a class separation value of five, and a balanced number of classification labels. Additionally, both simulated datasets had 500 features, and of these features 20 were informative features towards their associated classification labels.

Regarding the selection of publicly available clinical gene expression cohorts, the first of these was a diffuse large B-cell lymphoma (DLBCL) patient gene expression cohort, consisting of 7129 genes and 77 samples [22]. Of these 77 samples, 58 (75%) were DLBCL samples and the remaining 19 (25%) were follicular lymphoma samples (FL). Thus, the PSO algorithms will be identifying genetic signatures that have the ability to distinguish between the two lymphomas. The next of these gene expression cohorts included breast cancer patients consisting of different molecular subtypes including triple negative (TN) and human epidermal growth factor receptor 2 (HER2) positive [23]. TN breast cancer is defined as estrogen receptor (ER), progesterone receptor (PgR), and HER2 negative in its molecular subtype. This study selected only the triple negative and HER2-positive patients. This filtered cohort thus contained 31 samples and 54,675 genes, 17 of which were triple negative samples (55%) and the remaining 14 being HER2-positive samples (45%). The final gene expression cohort was of 248 patients with locally advanced prostate cancer commencing radical radiotherapy with androgen deprivation therapy

(ADT) [24]. Biochemical failure status was selected for classification due to its higher number of events in comparison with other phenotype class labels within this cohort. Biochemical failure is defined when PSA levels begin to rise again in prostate cancer patients following treatment. Other genetic signatures have been proposed for predicting biochemical failure [25,26]. These studies have highlighted the need for an effective method to predict biochemical risk to help therapeutic strategy decisions for prostate cancer patients. The cohort consisted of 19,453 genes, with 64 (26%) patient samples with biochemical failure status, and the remaining 184 (74%) did not. A summary of both the simulated and transcriptomics cohorts can be seen (Table 1).

### 2.2. Enhancement of PySwarms Python library

To perform the conventional version of BPSO, this study utilised the *PySwarms* Python module [27]. The *PySwarms* module was also adapted to develop a Python module for EBPSO. The pandas module version 1.2.4 was used for reading and manipulating the publicly available transcriptomics cohorts. The numpy version 1.17.2 module was used for manipulating data throughout the script also [28].

EBPSO uses information gain ratio filtering as a feature deduction and pre-processing technique to select the top 500 ranked features. This study however introduces an avenue of improvement for EBPSO's feature reduction, and of relevance towards a gene selection study, by replacing information gain ratio filtering with a differential expression p-value filtering technique to select the top 250 differentially expressed features. The number of ranked features from the filtering technique is reduced from 500 to 250 to better identify the important features within the candidate signatures, and to produce more succinct candidate signatures overall. These features were thus selected to be used as inputs for the two PSO methods. The differential expression filtering was performed with the *limma* R package version 3.48.3, using R version 4.1.2, as a well-known algorithm specifically built for differential expression analysis [29]. A high-level schematic showing how the data flow from the full transcriptomics cohorts to the visualization of the top candidate gene signatures can be seen (Fig. 1).

The scikit-learn library version 0.24.1 was also utilised to perform classification using SVM, leave-one-out (LOO) cross validation (CV), and to make predictions with trained models on the input dataset. SVMs are defined as supervised learning models in machine learning that analyse data towards classification or regression [30]. CV is a model validation technique for evaluating how statistical analysis results are generalized towards independent testing data [31]. Essentially, it evaluates how accurate a predictive model is. LOOCV splits the input dataset by using one sample as the testing data, and the remaining samples as the training data. This is repeated until every sample in the input dataset has been used as testing data. The final evaluation value for these predictive models is the average evaluation value seen across all the predictive models produced. In this study, this would be the average accuracy performance seen across all the predictive models produced by LOOCV. Runtime parameters for both PSO methods and SVM matched those in the original EBPSO implementation.

Results generated for each of the two PSO methods on each of the data sets were acquired using the same runtime specifications and on the same computing system. The computing system was on an Ubuntu 18.04.5 LTS, with an Intel® Xeon® processor, and a GPU card as a Quadro P2000. The processor includes 16 CPUs with a processing power of 3.70 GHz with 64 GB of memory. A single CPU was used on the DLBCL and breast cancer datasets to run each of the two PSO methods. Due to the greater number of samples within the simulated datasets with 200 samples and the prostate cancer cohort for biochemical failure prediction, additional CPUs were needed for appropriate overall runtimes. Thus, four CPUs were used to run each of the two PSO methods on these datasets. The runtimes for these datasets were represented as a single CPU for consistent comparisons of all the datasets used. These single CPU runtimes were represented simply as four times the magnitude of the recorded runtimes on four CPU's.

Ten runs of each of the PSO methods were run on each of the selected datasets in order to better present the performance and algorithm time complexity of each of the methods by using averages of their resulting runtime metrics. These runtime metrics used to evaluate the performance and time complexity of the PSO methods included their classification accuracy, the number of features selected, and the time taken to complete a single PSO run. The most accurate and sufficient gene signature selected over the shortest amount of time would have the greatest evaluated performance. The completion of a single PSO run is defined by 500 iterations of the updated particle positions. Additionally, the candidate signature produced from each single PSO run was used to generate hierarchical clustering heatmaps with their respective input datasets. This was achieved by using the Matplotlib (version 3.1.1) [32] and seaborn (version 0.10.1) Python modules. The plot_cost_history utility was adapted and improved to allow for two different cost history arrays from the two PSO algorithms to be directly compared on the same image. In this scenario, the average value at each iteration was measured over the ten runs of each PSO method to demonstrate their performance over the 500 iterations.

Regarding the analysis of the selected gene expression cohorts, EBPSO was run separately as an additional run following the previous 10 runs. In this separate additional run, the top three gene signatures based on their associated cost values are retained. This allowed for the comparison of the features that made up these selected signatures with each other and a previously defined gene signature for each clinical cohort. This was performed to demonstrate whether or not EBPSO was able to identify unique and biologically relevant candidate gene signatures. An eight probeset ID signature had been previously identified in distinguishing between the two lymphomas for the DLBCL cohort. Due to the lack of cohort specific gene signatures in the literature for the breast and prostate cancer cohorts, the *limma* R package was used to define a ten gene signature. This was comprised of the top ten most differentially expressed genes based on their associated p-values for HER2 against TN in the breast cancer cohort and biochemical failure prediction for the prostate cancer cohort.

**Table 1**
Summary of the simulated and selected clinical transcriptomics cohorts to validate EBPSO and BPSO.

| Dataset name | Samples | Features | Classes | Class balance | Class separation | Informative features |
|---|---|---|---|---|---|---|
| Binary class simulated | 200 | 500 | 2 | 100 (50%) / 100 (50%) | 5 | 20 |
| Multi-class simulated | 200 | 500 | 3 | 67 (33.5%)/ 67 (33.5%) / 66 (33%) | 5 | 20 |
| DLBCL | 77 | 7129 | 2 | 77 (75%) / 19 (25%) | – | – |
| Breast | 31 | 54,675 | 2 | 17 (55%) / 14 (45%) | – | – |
| FASTMAN | 248 | 19,453 | 2 | 64 (26%) / 184 (74%) | – | – |

Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization; DLBCL, Diffuse Large B-Cell Lymphoma.

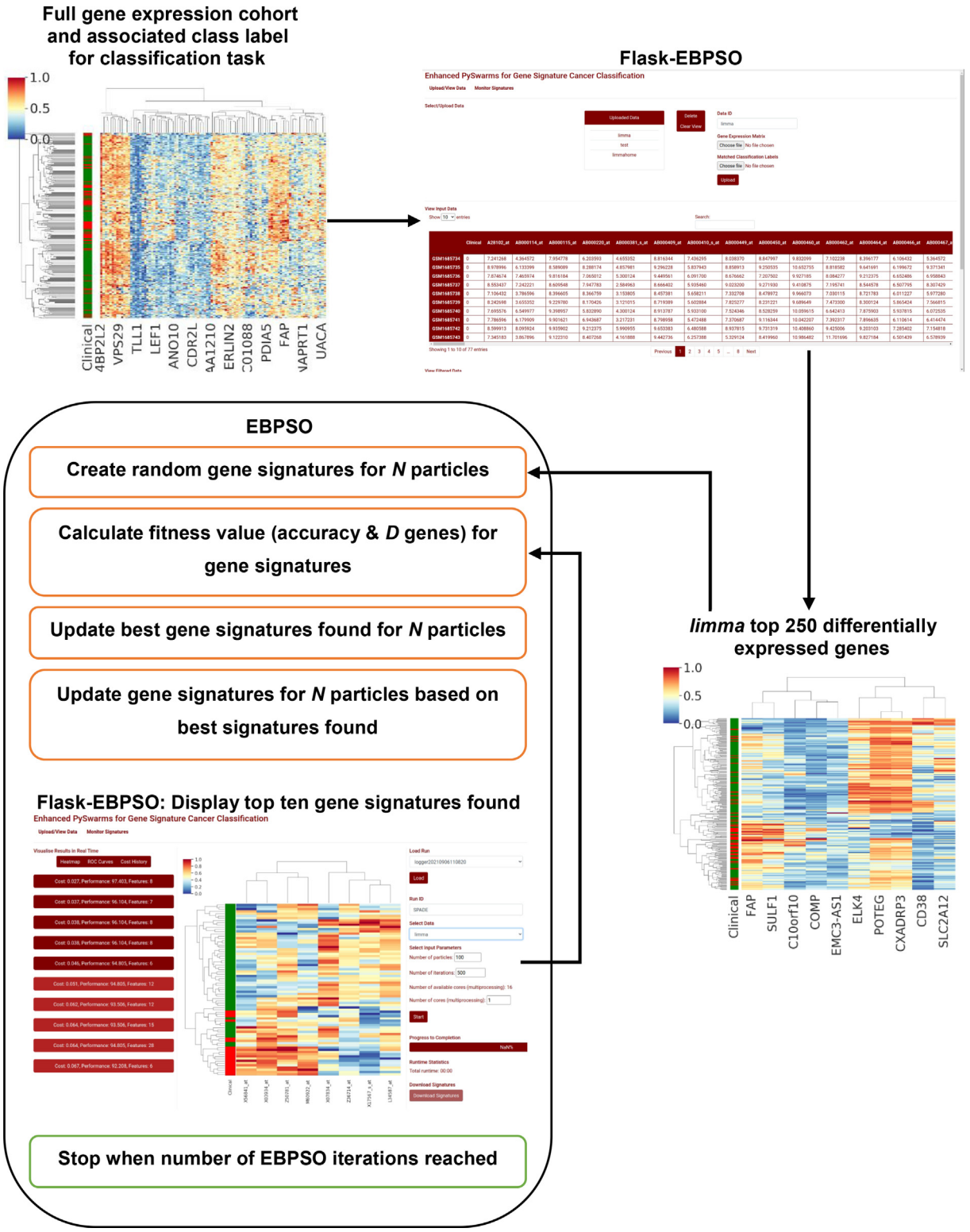


**Fig. 1.** High level schematic showing the data flow from the full transcriptomics cohort to Flask-EBPSO. The *limma* R package selects the top 250 differentially expressed genes to be used in EBPSO. After each iteration of EBPSO, Flask-EBPSO visualizes the top candidate gene signatures through hierarchical clustering heatmaps and ROC curves. This is repeated until the number of EBSPO iterations has been reached. Abbreviations: EBPSO, enhanced binary particle swarm optimization; ROC, receiver operating characteristic.

### *2.3. Biochemical failure survival analysis*

In addition to biochemical failure status being recorded for the patients associated with the prostate cancer cohort, their respective time to biochemical failure survival data was also available. Using the genes contained within selected gene signatures, hierarchical clustering was performed with Ward's linkage method and Euclidean distance to define two patient subgroups (k = 2). Associ-

ations to biochemical failure survival were determined through Cox proportional hazards survival analysis of the patient subgroups.

### 2.4. Web-based micro-framework EBPSO application

A web-based micro-framework was developed using the Flask Python module. To deal with the new functionality that Flask provides for EBPSO, the developed EBPSO module was further updated and improved to adhere to these. These changes include retaining the ranked history of unique signatures produced by EBPSO. New visualization functions have been generated also. The top ten unique candidate signatures produced from EBPSO are visualised through hierarchical clustering heatmaps and receiver operating characteristic (ROC) curves in Flask EBPSO. ROC curves are generated by plotting the true positive rate (TPR) or sensitivity against the false positive rate (FPR) or specificity. The TPR relates to the number of correctly identified real positive cases in a dataset against the total number of real positive cases. The FPR relates to the number of correctly identified real negative cases against the total number of real negative cases. Thus, ROC curves are illustrations demonstrating the diagnostic ability of binary classification. The scikit-learn library was used for performing CV, train SVM classifiers for class label prediction probabilities, and compute the ROC curve values. The StratifiedKFold utility allowed for CV to be performed, splitting the input dataset with five splits. This resulted in five different training and testing datasets to be evaluated. This k-fold CV method was used instead of the previously defined LOOCV as this would provide a quick assessment of how well a trained model performs for visualisation purposes.

Interested collaborators are invited to contact the authors to access the program code developed in this study.

## 3. Results

### 3.1. Comparative analysis between EBPSO and PySwarms BPSO

Regarding performance on the simulated datasets, EBPSO demonstrated the ability to perform as accurately as BPSO whilst consistently identifying candidate signatures with substantially less associated features and within a significantly faster timeframe (Table 2). Following this, performance on the gene expression cohorts shows similar results. Only the FASTMAN prostate cancer cohort for biochemical failure prediction showed increased accuracy for BPSO regarding the best signature produced. In this case, *PySwarms* BPSO had superior accuracy for its best signature with 90% performance, whilst EBPSO still a similar accuracy with 85.1%. The overview of candidate gene signatures from EBPSO and BPSO over ten runs can be seen for all the simulated and clinical gene expression data sets in Appendix A.

### 3.2. Best gene signatures from EBPSO and BPSO on simulated datasets

Two simulated datasets were generated to represent the high dimensionality gene expression cohorts that EBPSO should be used towards for gene signature selection. To represent simpler binary classification tasks, one of these simulated datasets had two different classification labels, whilst the other simulated dataset had three different classes to represent more complex multi-classification tasks.

Focusing on the binary class simulated dataset, hierarchical clustering of the 20 informative features shows each features relationship towards these two classes (Fig. 2A). EBPSO produced a candidate gene signature of five features that demonstrated almost perfect accuracy of 99.5% towards classification. The candidate gene signature from EBPSO included two informative features as *193* and *241*, as well as another three non-informative features (Fig. 2B). BPSO produced a similarly accurate candidate signature, but with 154 features it failed to generate a signature that was succinct. The candidate signature from BPSO included seven informative features, but also included an additional 147 non-informative genes (Fig. 2C). EBPSO performed more favourably in comparison to *PySwarms* BPSO regarding the average cost history over ten runs of each method (Fig. 2D).

Using the more complex multi-class simulated dataset, hierarchical clustering for the 20 informative features shows the closely correlated relationship between these features and the three classes (Fig. 3A). EBPSO's candidate signature included 77 features, seven of which were informative features (Fig. 3B). BPSO produced a candidate signature of 155 features in length (Fig. 3C). Similar performance from EBPSO was observed in comparison to the 200 sample simulated dataset with two classes (Fig. 3D).

### 3.3. Best gene signatures from EBPSO and BPSO on clinical cohorts

Having evaluated EBPSO in simulated gene expression cohorts, the algorithm was evaluated on publicly available gene expression cohorts. The first of these was the DLBCL cohort, with PSO algorithms classifying between the DLBCL and FL, two different lymphomas [22]. The previously identified eight probeset signature

**Table 2**
Comparing the best candidate gene signatures produced from EBPSO and BPSO on simulated and real patient gene expression data sets. Note that runtimes for the simulated datasets and the GSE116918 FASTMAN dataset was represented as single CPU runs, but was run on four CPU's.

| Dataset | Statistics | EBPSO | | | BPSO | | |
|---|---|---|---|---|---|---|---|
| | | Best | Average | S.D. | Best | Average | S.D. |
| Binary class simulated | Accuracy (%) | 99.5 | 99.5 | 0 | 99.5 | 99.5 | 0 |
| | Genes | 5 | 9.7 | 2.8 | 154 | 157.1 | 2 |
| | Time (min) | 1129 | 1185 | 37.7 | 7363 | 7426 | 44.5 |
| Multi-class simulated | Accuracy (%) | 99 | 96.5 | 1.9 | 99 | 99 | 0 |
| | Genes | 77 | 56.7 | 59.8 | 155 | 159.6 | 3.5 |
| | Time (min) | 1447 | 1465 | 49 | 10202.1 | 10264.7 | 48.1 |
| DLBCL | Accuracy (%) | 100 | 100 | 0 | 100 | 99.2 | 0.7 |
| | Genes | 5 | 5.4 | 0.8 | 69 | 69.2 | 5.2 |
| | Time (min) | 68.9 | 70.5 | 1.1 | 207 | 204.8 | 5.8 |
| GSE43358 | Accuracy (%) | 100 | 100 | 0 | 100 | 100 | 0 |
| | Genes | 2 | 2.3 | 0.5 | 57 | 63.6 | 3.3 |
| | Time (min) | 17.8 | 18.2 | 0.3 | 29.2 | 29.4 | 0.3 |
| GSE116918 FASTMAN | Accuracy (%) | 85.1 | 83.8 | 0.9 | 90 | 88.8 | 0.8 |
| | Genes | 15 | 12.3 | 7 | 98 | 92.8 | 6.1 |
| | Time (min) | 1464 | 1515 | 58 | 5522 | 5451 | 87.8 |

Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization; S.D., Standard Deviation.

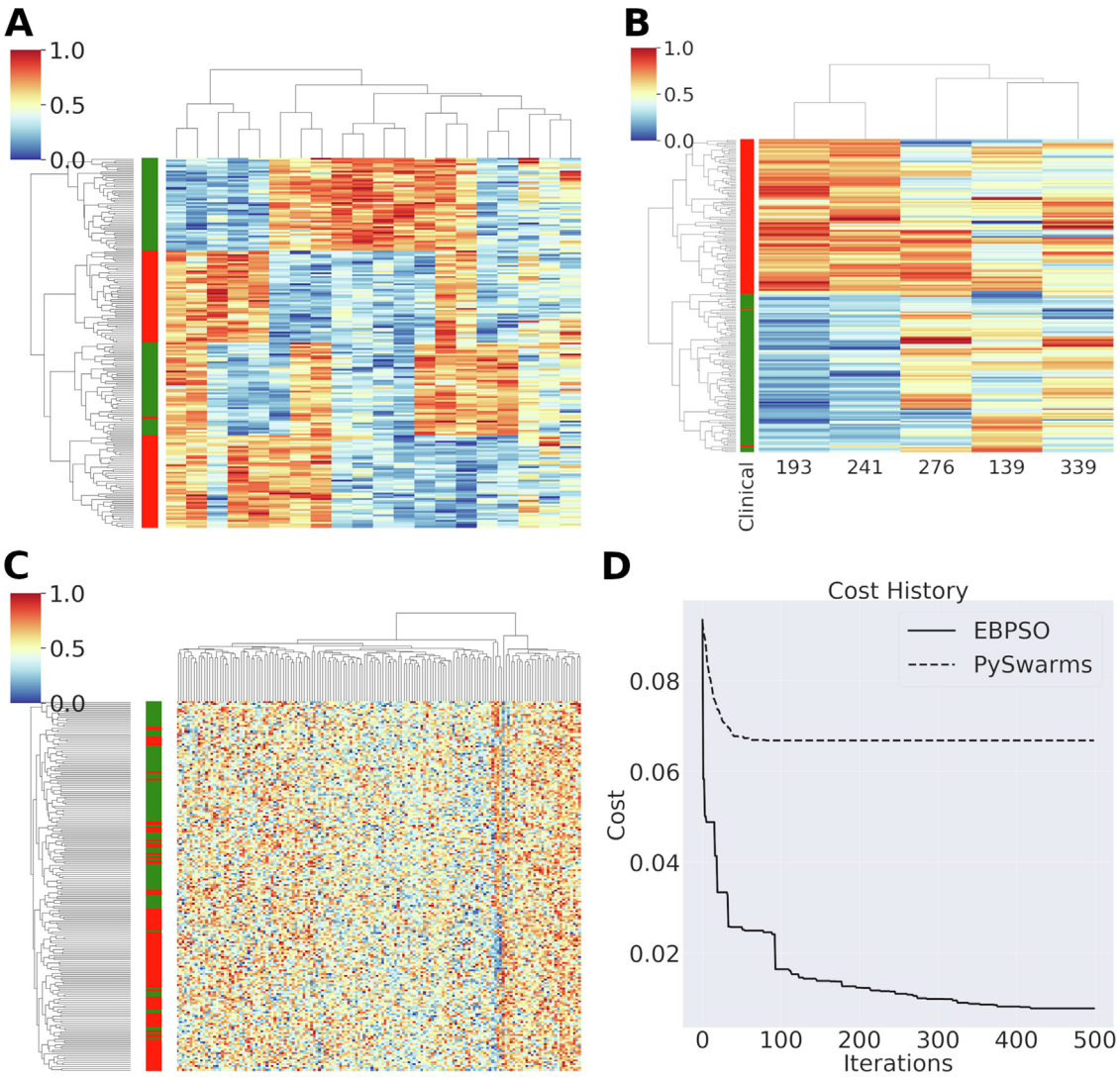


**Fig. 2.** EBPSO and *PySwarms* BPSO on a simulated dataset of 200 samples, 500 features, two classes, and 20 informative features with a class separation of five. A. Hierarchical clustering of the 20 informative features towards the two classes. B. Hierarchical clustering of the candidate signature selected by EBPSO. C. Hierarchical clustering of the candidate signature selected by *PySwarms* BPSO. D. Cost history over 500 iterations for EBPSO (solid) and *PySwarms* BPSO (dashed). Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization.

for distinguishing between these two lymphomas has been visualized with hierarchical clustering (Fig. 4A). EBPSO produced a candidate signature with five probeset IDs out of the 250 lowest differential expression p-value filtered input features to the algorithm with 100% accuracy (Fig. 4B). BPSO however generated a much larger candidate signature of 69 probesets with 100% accuracy also (Fig. 4C). Regarding the average cost performance over ten runs of each PSO algorithm, EBPSO performed more favourably over *PySwarms* BPSO (Fig. 4D). Between the top three selected candidate signatures and the known eight probeset signature, six probesets are shared amongst them out of a combined total of 22 probesets between them (Fig. 4E). Five probesets are shared between the first and second selected signatures from EBPSO as L17131_rna1_at (HMGA1), U46006_s_at (CSPR2), X54941_at (CKS1B), X78992_at (ZFP36L2) and M83751_at (MANF). Additionally, one was shared with the first and second selected signature and the known signature as D87119_at (TRIB2). Each of the probesets selected by the top three selected candidate signatures were shown to be statistically significant for differential expression (Table 3).

The next of these gene expression cohorts looked to distinguish between TN and HER2-positive breast cancer patients. This study used the *limma* R package to identify the top ten most DEGs based on p-value between the two breast cancer molecular subtypes (Fig. 5A) [29]. This allowed for the creation of a known genetic signature that could be compared with the results of the two PSO algorithms. EBPSO produced a candidate signature consisting of only two probeset IDs (Fig. 5B). BPSO however generated a candidate signature of 57 features (Fig. 5C). Much like the previous simulated and clinical datasets, EBPSO was able to outperform *PySwarms* BPSO regarding their average cost value comparison (Fig. 5D). The comparison of the features selected from the top three candidate signatures produced from a single run of EBPSO with the top ten DEGs identified by *limma* can be seen within its associated Venn diagram (Fig. 5E). Each of the features selected by the top three selected candidate signatures were shown to be statistically significant for differential expression (Table 4). No features were shared amongst the four signatures, demonstrating the selection of unique signatures which all performed favourably with

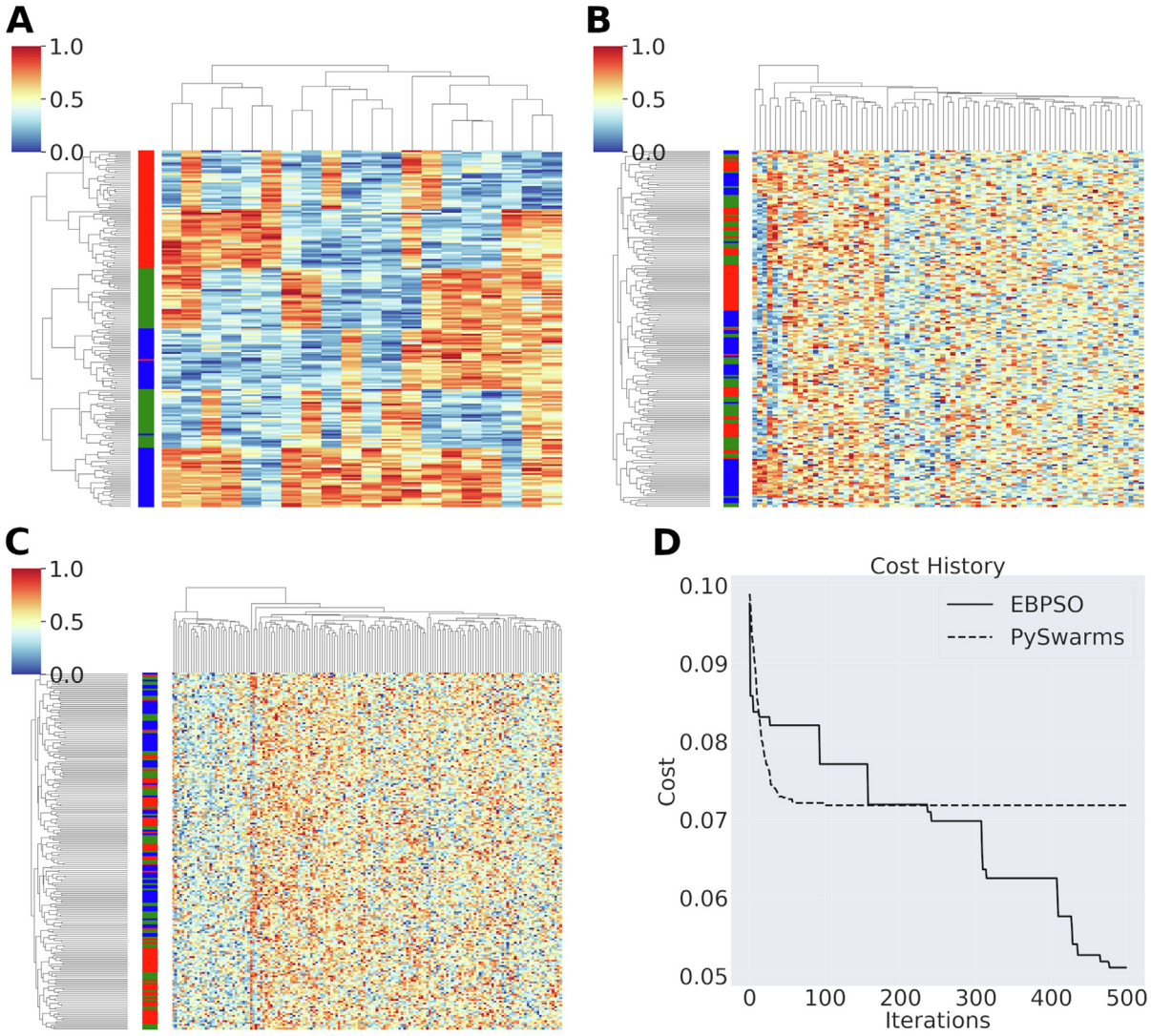


**Fig. 3.** EBPSO and *PySwarms* BPSO on a simulated dataset of 200 samples, 500 features, three classes, and 20 informative features with a class separation of five. A. Hierarchical clustering of the 20 informative features towards the three classes. B. Hierarchical clustering of the candidate signature selected by EBPSO. C. Hierarchical clustering of the candidate signature selected by *PySwarms* BPSO. D. Cost history over 500 iterations for EBPSO (solid) and *PySwarms* BPSO (dashed). Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization.

100% accuracy for classification between HER2 against TN breast cancer samples.

The final clinical gene expression cohort was of locally advanced prostate cancer patients commencing radical radiotherapy with ADT for biochemical failure prediction. The *limma* R package identified the top ten most statistically significant DEGs for biochemical failure status in this cohort for the generation of a reference gene signature (Fig. 6A). EBPSO identified a 15 gene candidate signature, but this failed to produce a highly accurate distinction for biochemical failure status (Fig. 6B). BPSO in comparison produced a candidate signature of 98 genes, but with an increase in accuracy for biochemical failure prediction in comparison to EBPSO (Fig. 6C). EBPSO outperformed *PySwarms* BPSO regarding their average cost value and the number of genes within their candidate signature, however (Fig. 6D). One gene was shared between the top selected signature from EBPSO and the previously defined *limma* DEG signature for biochemical failure status, as fibroblast activation protein alpha (FAP; Fig. 6E). Additionally, one gene was shared between the first and third selected signatures from EBPSO, being elongation factor for RNA polymerase II 2 (ELL2). Finally, one gene was shared between the second and third selected signatures from EBPSO, as prostate cancer associated transcript 4 (PCAT4). FAP was the only gene that was statistically significant following FDR correction as a differentially expressed gene selected by the top three candidate signatures, and this gene was only selected by the top candidate signature (Table 5). All the genes selected by the top three candidate signatures were however statistically significant with an unadjusted p-value.

As this clinical gene expression cohort also contained associated patient survival data, the selected gene signatures were evaluated with survival analysis to investigate their prognostic ability on time to biochemical failure. Hierarchical clustering on the selected genes within the selected signatures were used to define two subgroups for these patients. Cox proportional hazard survival modelling was then applied to the subgroups to determine their prognostic ability. The hierarchical clustering subgroups for the *limma* DEG for biochemical failure status signature produced a statistically significant poor prognosis *Subgroup2* of 78/248 (31.4%) patients [HR = 3.56 (2.25 – 5.64); $p < 0.001$] (Fig. 7A). The subgroups produced for the EBPSO 15 gene candidate signature similarly identified a poor prognosis *Subgroup2* of 109/248 (44%) patients, but with a greater risk of a biochemical failure event

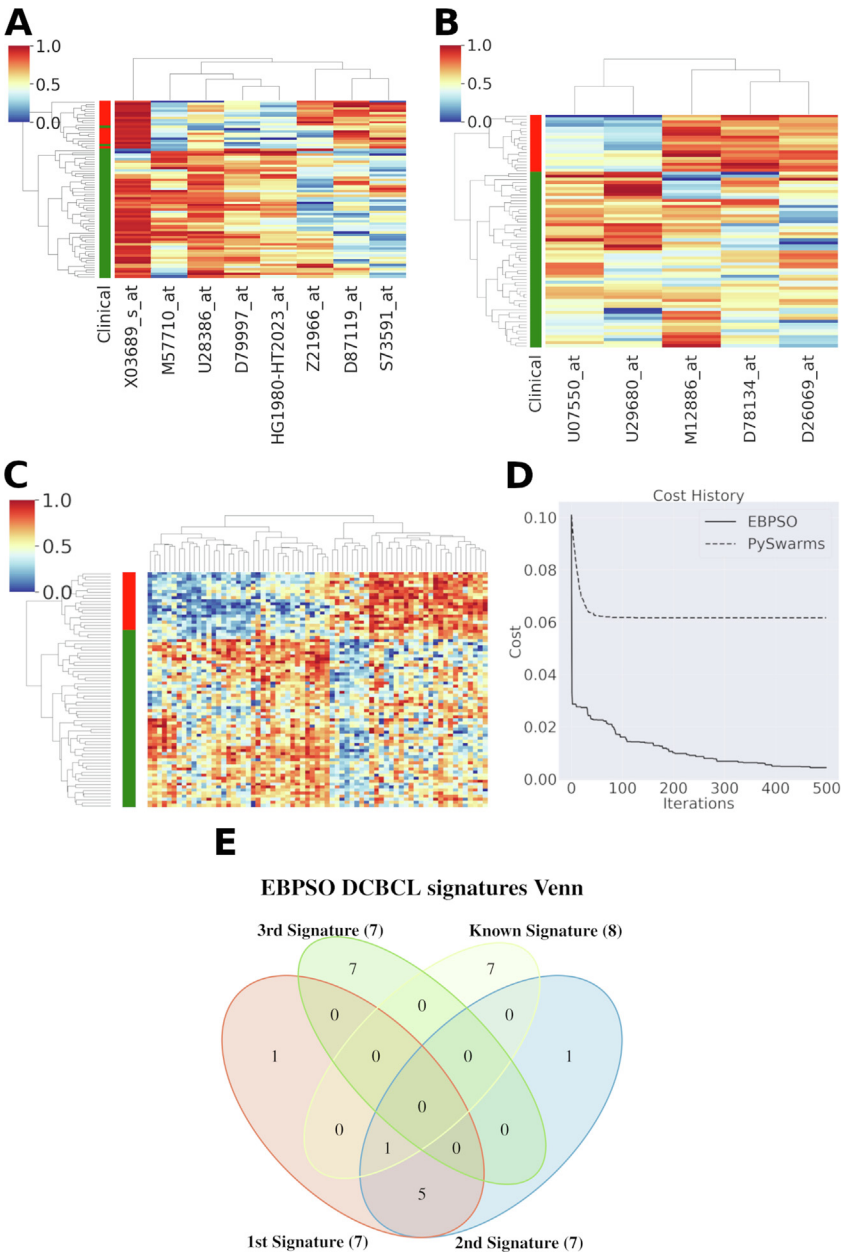


**Fig. 4.** EBPSO and *PySwarms* BPSO on the DLBCL data set. A. Hierarchical clustering of eight previously identified features for best class separation towards the two classes. B. Hierarchical clustering of the candidate signature selected by the EBPSO. C. Hierarchical clustering of the candidate signature selected by *PySwarms* BPSO. D. Cost history over 500 iterations for EBPSO (solid) and *PySwarms* BPSO (dashed). E. Venn diagram comparing the top three candidate signatures from a single run of EBPSO and the eight previously identified features in A. Abbreviations: EBPSO, Enhanced Binary Particle Swarm; BPSO, Binary Particle Swarm Optimization; DLBCL, Diffuse Large B-Cell Lymphoma.

[HR = 4.1 (2.47 – 6.81); p < 0.001] (Fig. 7B). The subgroups for the BPSO 98 gene candidate signature had a much larger poor prognosis *Subgroup2* of 178/248 (71.8%) patients, and had a decreased risk of a biochemical failure event in comparison with EBPSO [HR = 2.97 (1.53 – 5.79); p < 0.001] (Fig. 7C). The subgroups produced for the top two ranked candidate signatures from the separate single run of EBPSO both identified a statistically significant poor prognosis *Subgroup2* (Fig. 7D). The subgroups produced for the third ranked candidate signature from EBPSO identified a worse prognosis *Subgroup1* of 195/248 (78.6%) patients, which was not statistically significant for time to biochemical failure [HR = 0.52 (0.27 – 1.01); p = 0.054].

### *3.4. HTML GUI with Flask micro web framework*

In addition to the creation of the adapted EBPSO *PySwarms* Python module, a web-based micro-framework analytical application was also developed for the EBPSO module (Video 1; Fig. 8). This web-based application allowed for real time analytical visualisations and runtime statistics from the algorithm. More importantly, it would allow for the top ten candidate gene signatures selected from EBPSO to be viewed. This study hypothesizes that some of the candidate signatures will be unique regarding its selected features.

The Flask EBPSO micro-framework has two main pages, the Upload/View Data page, and the Monitor Signatures page. The

**Table 3**
Summary statistics from *limma* for the gene features selected by the top three ranked gene signatures selected by EBPSO on the DLBCL dataset for DLBCL vs FL. Feature names in bold relate to features that have appeared in more than one of the top three ranked gene signatures from EBPSO.

| Signature rank | Feature name | Gene | LogFC | p-value | Adj(p-value) |
|---|---|---|---|---|---|
| 1st ranked | **D87119_at** | **TRIB2** | −1.557 | 3.436e-11 | 6.123e-08 |
| | **L17131_rna1_at** | **HMGA1** | 1.321 | 1.791e-09 | 8.51e-07 |
| | **U46006_s_at** | **CSRP2** | −1.724 | 6.651e-09 | 1.6e-06 |
| | **X54941_at** | **CKS1B** | 2.037 | 3.634e-07 | 3.365e-05 |
| | D26069_at | ACAP2 | −0.722 | 2.233e-06 | 1.373e-04 |
| | **X78992_at** | **ZFP36L2** | −1.324 | 9.046e-06 | 4.243e-04 |
| | **M83751_at** | **MANF** | 0.778 | 1.143e-05 | 4.97e-04 |
| 2nd ranked | **D87119_at** | **TRIB2** | −1.557 | 3.436e-11 | 6.123e-08 |
| | **L17131_rna1_at** | **HMGA1** | 1.321 | 1.791e-09 | 8.51e-07 |
| | **U46006_s_at** | **CSPR2** | −1.724 | 6.651e-09 | 1.6e-06 |
| | **X54941_at** | **CKSB1B** | 2.037 | 3.634e-07 | 3.365e-05 |
| | M22960_at | CTSA | 0.75 | 1.01e-06 | 7.12e-05 |
| | **X78992_at** | **ZFP36L2** | −1.324 | 9.046e-06 | 4.243e-04 |
| | **M83751_at** | **MANF** | 0.778 | 1.143e-05 | 4.97e-04 |
| 3rd ranked | M74093_at | CCNE1 | 1.87 | 1.448e-11 | 3.44e-08 |
| | M14328_s_at | ENO1 | 0.899 | 1.191e-09 | 6.531e-07 |
| | M23323_s_at | CD3E | −0.924 | 2.203e-09 | 9.551e-07 |
| | L19437_at | TALDO1 | 0.749 | 3.311e-07 | 3.147e-05 |
| | Z35227_at | RHOH | −0.978 | 6.402e-07 | 5.015e-05 |
| | X66867_cds1_at | MAX | −1.133 | 1.451e-05 | 5.868e-04 |
| | HG4258-HT4528_at | — | −0.967 | 2.878e-05 | 9.634e-04 |

Abbreviations: LogFC, Log Fold Change; Adj(p-value), Adjusted p-value; TRIP2, Tribbles Pseudokinase 2; HGMA1, High Mobility group AT-hook 1; CSRP2, Cysteine and glycine Rich Protein 2; CKS1B, CDC28 protein Kinase regulatory Subunit 1B; ACAP2, ArfGAP with Coiled-coil, Ankyrin repeat and PH domains 2; ZFP36L2, ZFP36 ring finger protein Like 2; MANF, Mesencephalic Astrocyte derived Neurotrophic Factor; CTSA, Cathepsin A; CCNE1, Cyclin E1; ENO1, Enolase 1; CD3E, CD3e molecule; TALDO1, Transaldolase 1; RHOH, Ras Homolog family member H; MAX, MYC Associated factor X.

Upload/View Data page allows for the end user to load the transcriptomics data and its associated clinical information class labels for classification as CSV files. This page also allows for the previously uploaded datasets to be viewed for review. The Monitor Signatures page runs EBPSO and shows real time visualizations. Additionally, the final visualizations and candidate signature statistics are saved when EBPSO has completed. These completed sessions can be reloaded at another time point and visualized on the web-based application for greater interactivity of the produced results.

## 4. Discussion

The results presented with these simulated and real patient gene expression cohorts demonstrate that EBPSO consistently outperforms BPSO. Additionally, EBPSO was shown to outperform a supervised feature selection method with *limma* differential expression for biochemical failure prediction for prostate cancer patients. These results also reflect the improved performance of PSO as a feature selection technique in comparison with other evolutionary algorithms and unsupervised methods in combination with SVM [33].

Using the simulated cohorts, it can be demonstrated that both of the different PSO methods failed to distinguish the highly informative features from non-informative features due to the inclusion of non-information features in candidate gene signatures. When these highly informative features are included within a gene signature, their contribution towards classification may be highly significant. In this scenario, the inclusion of non-informative features may not affect accuracy performance, and thus these PSO algorithms may fail to separate these non-informative features from their candidate signatures.

It is worth noting the deceased accuracy, in comparison to the other patient gene expression cohorts, from the best performing candidate gene signatures towards biochemical failure prediction for prostate cancer patients from both PSO methods investigated. Based on the top ten DEGs for biochemical failure in this cohort seen in Fig. 6, it is clear that the failure to create a distinction between biochemical failure status represents a more complex classification task in comparison to the other gene expression cohorts. Additionally, the classification tasks associated with the other gene expression cohorts are related to making predictions for well-defined cancer subtypes in DLBCL against FL, and TN against HER2-postive breast cancers. Biochemical failure however is an indicator of prostate cancer disease progression and its prediction for definitive radiotherapy treated prostate cancer patients is a less studied and more complex classification task. Despite one prostate cancer gene signature being shown to be prognostic in predicting for biochemical failure in this cohort [34], biomarker discovery in this cohort has yet to be investigated for comparison.

### 4.1. EBPSO identifies signatures with informative features and few non-informative features

The original EBPSO algorithm focused heavily on classification accuracy, the number of genes within a signature, and the time taken to complete a single run of EBPSO. Whilst these are important in the definition of the algorithm, there is no mention of what these selected genes are, or more importantly their importance towards classification accuracy. We used simulated datasets in order to identify these important or informative genes to determine if they had been selected by candidate signatures produced from EBPSO.

In the simulated datasets, the informative features were identified by both EBPSO and *PySwarms* BPSO. Other non-informative features were also included in these candidate signatures, however. Some of these non-informative features were also not contributing towards classification either and were not needed to be included in the candidate signature. As previously discussed, possible reasons for this include the potential for informative genes with large class separation values increasing overall accuracy for a candidate gene signature and thus the inclusion of some non-informative genes may not affect is ability in classification accuracy. These results show that the candidate signatures identified from EBPSO may not contain important or informative features only, and could be trapped in closely related local minima and

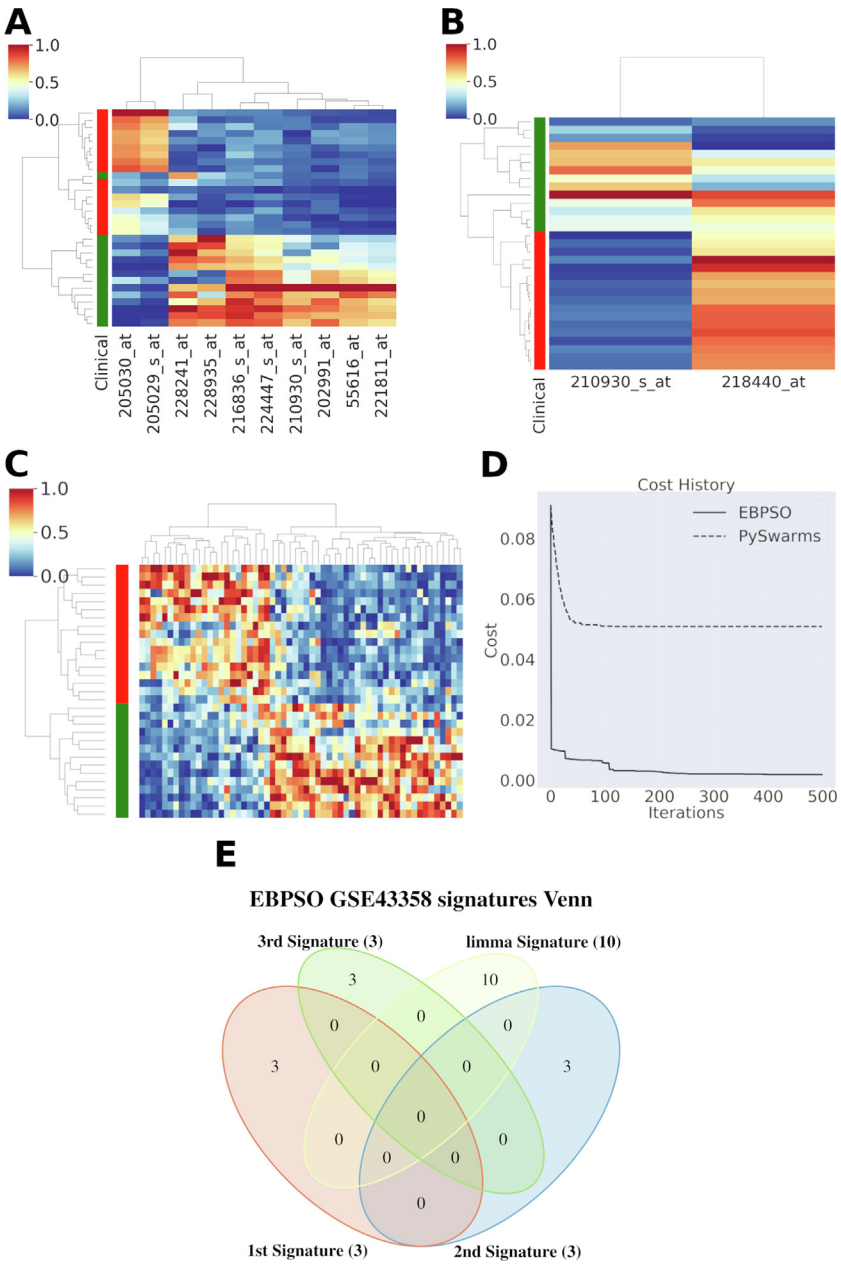


**Fig. 5.** EBPSO and *PySwarms* BPSO on GSE43358. A. Hierarchical clustering of the top ten features selected on p-value from *limma*. B. Hierarchical clustering of the candidate signature selected by EBPSO. C. Hierarchical clustering of the candidate signature selected by *PySwarms* BPSO. D. Cost history over 500 iterations for EBPSO (solid) and *PySwarms* BPSO (dashed). E. Venn diagram comparing the top three candidate signatures from a single run of EBPSO and the top ten features selected from *limma* in A. Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization.

**Table 4**
Summary statistics from *limma* for the gene features selected by the top three ranked gene signatures selected by EBPSO on GSE43358 for TNBC vs HER2 breast cancer subtypes.

| Signature rank | Feature name | Gene | LogFC | p-value | Adj(p-value) |
|---|---|---|---|---|---|
| 1st ranked | 211026_s_at | MGLL | −0.994 | 2.308e-05 | 0.007 |
| | 218440_at | MCC1 | 0.827 | 1.791e-09 | 0.007 |
| | 201728_s_at | KIAA0100 | −0.921 | 3.253e-05 | 0.008 |
| 2nd ranked | 219344_at | SLC29A3 | −0.625 | 5.835e-06 | 0.003 |
| | 227279_at | TCEAL3 | −0.981 | 6.056e-06 | 0.003 |
| | 224809_x_at | TINF2 | −0.373 | 8.197e-06 | 0.004 |
| 3rd ranked | 221732_at | CANT1 | −0.94 | 1.914e-06 | 0.002 |
| | 222400_s_at | ADI1 | 0.58 | 4.098e-05 | 0.009 |
| | 223344_s_at | MS4A7 | −0.858 | 4.175e-05 | 0.009 |

Abbreviations: LogFC, Log Fold Change; Adj(p-value), Adjusted p-value; MGLL, Monoglyceride Lipase; MCC1, Methylcrotonyl-CoA Carboxylase subunit 1; SLC29A3, solute Carrier family 29 member 3; TCEAL3, Transcription Elongation factor A Like 3; TINF2, TERF1 Interacting Nuclear Factor 2; CANT1, Calcium Activated Nucleotidase 1; ADI1, Acireductone Dioxygenase 1; MS4A7, Membrane Spanning 4-domains A7.

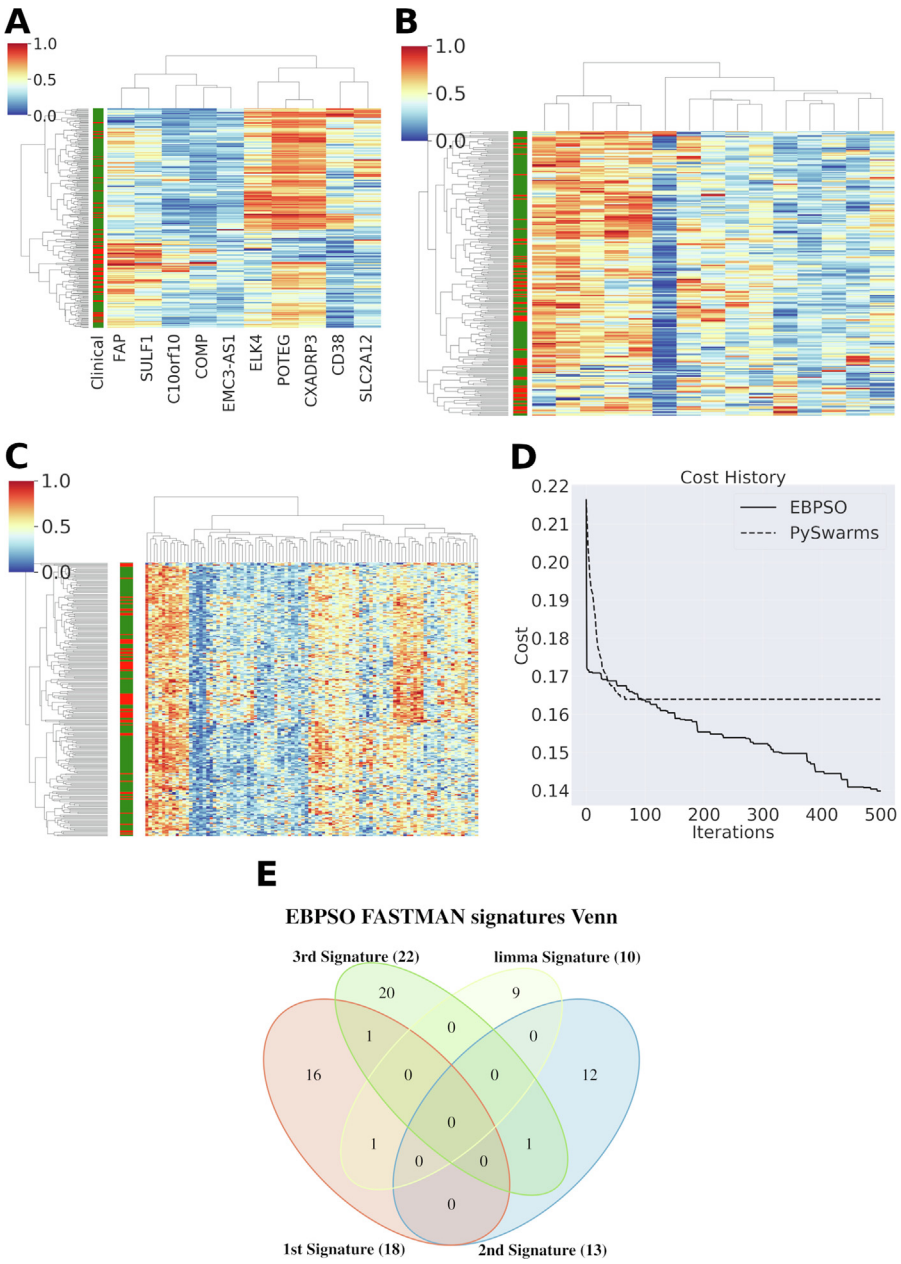


**Fig. 6.** EBPSO and *PySwarms* BPSO on GSE116918 FASTMAN. A. Hierarchical clustering of the top ten features selected based on p-value from *limma*. B. Hierarchical clustering of the candidate signature selected by EBPSO. C. Hierarchical clustering of the candidate signature selected by *PySwarms* BPSO. D. Cost history over 500 iterations for EBPSO (solid) and *PySwarms* BPSO (dashed). E. Venn diagram comparing the top three candidate signatures from a single run of EBPSO and the top ten features selected from *limma* in A. Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization.

not the best solution global minima. Regardless, 99% accuracy and above was demonstrated with these candidate signatures on these simulated datasets, but the inclusion of non-informative features has larger implications when identifying genetic signatures on clinical cohorts.

### *4.2. EBPSO identifies unique signatures with different underlying biology*

Regarding the clinical data sets, there is the potential to further validate the signature selection from EBPSO by exploring the underlying biology driving their selection. The identification of the most informative genetic signatures amongst these highly dimensional transcriptomic cohorts will allow for greater efforts in biomarker discovery with greater diagnostic, predictive, and prognostic potential. Prognostic gene signatures for identifying which early breast cancer patients would have more aggressive disease in the future have already been translated into clinical practice, highlighting the clinical potential of gene signatures [35]. The EBPSO signatures have been selected based on both their classification accuracy and how succinct they are in terms of the number of genes involved within the signature. The more succinct a signature can become whilst retaining its classification accuracy is most beneficial towards its clinical translation. This is reflected in the reduction of financial cost associated with using a succinct signature within clinical practice.

Amongst the transcriptional datasets, no single gene or feature was shared amongst the top three candidate signatures produced

**Table 5**
Summary statistics from *limma* for the gene features selected by the top three ranked gene signatures selected by EBPSO on GSE116918 FASTMAN for biochemical failure status. Feature names in bold relate to features that have appeared in more than one of the top three ranked gene signatures from EBPSO. Abbreviations for the official gene symbols can be seen in Appendix B.

| Signature rank | Feature name | LogFC | p-value | Adj(p-value) |
|---|---|---|---|---|
| 1st ranked | FAP | 0.599 | 4.069e-07 | 0.008 |
| | MX1 | 0.331 | 1.424e-05 | 0.046 |
| | RAB27B | −0.507 | 2.844e-05 | 0.061 |
| | RGS16 | 0.341 | 5.305e-05 | 0.069 |
| | UACA | 0.225 | 0.0003 | 0.137 |
| | ADC | 0.188 | 0.0003 | 0.137 |
| | IGFBP3 | 0.247 | 0.0004 | 0.158 |
| | AKAP7 | −0.226 | 0.0012 | 0.247 |
| | FCER1G | 0.208 | 0.0024 | 0.302 |
| | ANO10 | 0.214 | 0.0038 | 0.362 |
| | COL1A2 | 0.323 | 0.004 | 0.364 |
| | THBS1 | 0.164 | 0.0047 | 0.394 |
| | GLIPR1 | 0.193 | 0.0049 | 0.402 |
| | **ELL2** | −0.199 | 0.0064 | 0.419 |
| | MFSD4 | −0.209 | 0.007 | 0.432 |
| | APPBP2 | −0.149 | 0.0131 | 0.498 |
| | PDIA5 | −0.2 | 0.0138 | 0.505 |
| | OR51D1 | 0.203 | 0.0198 | 0.579 |
| 2nd ranked | RTCA | −0.274 | 0.0002 | 0.107 |
| | TLL1 | 0.218 | 0.0002 | 0.112 |
| | SIAH1 | −0.318 | 0.002 | 0.271 |
| | ZNF382 | 0.229 | 0.002 | 0.281 |
| | PPAPDC1B | −0.344 | 0.002 | 0.294 |
| | ASPN | 0.382 | 0.003 | 0.345 |
| | CHRNA2 | −0.325 | 0.003 | 0.357 |
| | LYZ | 0.455 | 0.004 | 0.362 |
| | PCED1A | 0.151 | 0.005 | 0.394 |
| | RBP7 | −0.223 | 0.008 | 0.442 |
| | DESI2 | 0.139 | 0.013 | 0.501 |
| | **PCAT4** | −0.52 | 0.014 | 0.511 |
| | SYNPO | 0.153 | 0.019 | 0.575 |
| 3rd ranked | ORL51L1 | −0.257 | 0.0007 | 0.197 |
| | FNDC1 | 0.602 | 0.0008 | 0.217 |
| | IFI44L | 0.343 | 0.0009 | 0.217 |
| | TMEM138 | 0.139 | 0.001 | 0.236 |
| | TRPM8 | −0.482 | 0.001 | 0.236 |
| | ZNF702P | −0.376 | 0.001 | 0.245 |
| | PDCD1LG2 | 0.135 | 0.001 | 0.245 |
| | HOX19 | −0.352 | 0.002 | 0.271 |
| | CTSD | 0.323 | 0.002 | 0.274 |
| | MRPL17 | −0.292 | 0.003 | 0.319 |
| | SAMD3 | 0.332 | 0.003 | 0.341 |
| | TMSB10 | 0.283 | 0.004 | 0.362 |
| | SUSD4 | −0.24 | 0.004 | 0.362 |
| | MPEG1 | 0.256 | 0.004 | 0.371 |
| | SLFN5 | 0.207 | 0.006 | 0.412 |
| | **ELL2** | −0.199 | 0.006 | 0.419 |
| | PRSS27 | 0.208 | 0.008 | 0.441 |
| | **PCAT4** | −0.52 | 0.014 | 0.511 |
| | CRIP1 | 0.136 | 0.017 | 0.558 |
| | ZNF613 | −0.216 | 0.018 | 0.567 |
| | AIFM1 | −0.139 | 0.02 | 0.581 |
| | GFM2 | −0.173 | 0.025 | 0.611 |

Abbreviations: LogFC, Log Fold Change; Adj(p-value), Adjusted p-value.

from EBPSO and a previously defined signature for classification. In the DLBCL dataset for DLBCL against FL, five features were shared between the first and second placed candidate signatures produced from EBPSO, being L17131_rna1_at, U46006_s_at, X54941_at, X78992_at and M83751_at. L17131_rna1_at is a probeset ID which is annotated towards the high mobility group AT-hook 1 (HMGA1) gene. HGMA1 has been identified as a master regulator for contributions towards disease progression of FL patients for transformation towards the aggressive DLBCL [36–38]. To complement this, HGMA1 was downregulated in FL samples and upregulated in DLBCL samples in this dataset. X54941_at is annotated as the CDC28 protein kinase regulatory subunit 1B (CKS1B). CKS1B has been associated with greater survival outcomes following chemotherapy on FL patients [39].

For the candidate signatures produced by EBPSO for the breast cancer cohort, none of their features were seen to be shared. In the prostate cancer cohort for biochemical failure prediction, the FAP gene was shared between the top candidate signature produced from EBPSO and the *limma* ten gene signature. FAP has been associated with metastatic disease in prostate cancer [40], and in turn was also upregulated in the prostate cancer cohort. Additionally, it has been seen to be expressed in human prostate cancer stroma [41]. The ELL2 gene was shared between the first and third candidate signatures from EBPSO. ELL2 is androgen responsive gene that is largely seen as a tumour suppressor in the prostate [42] and has been shown to be downregulated in advanced prostate cancer [43]. This was also reflected in the prostate cancer cohort as being downregulated for biochemical failure status.

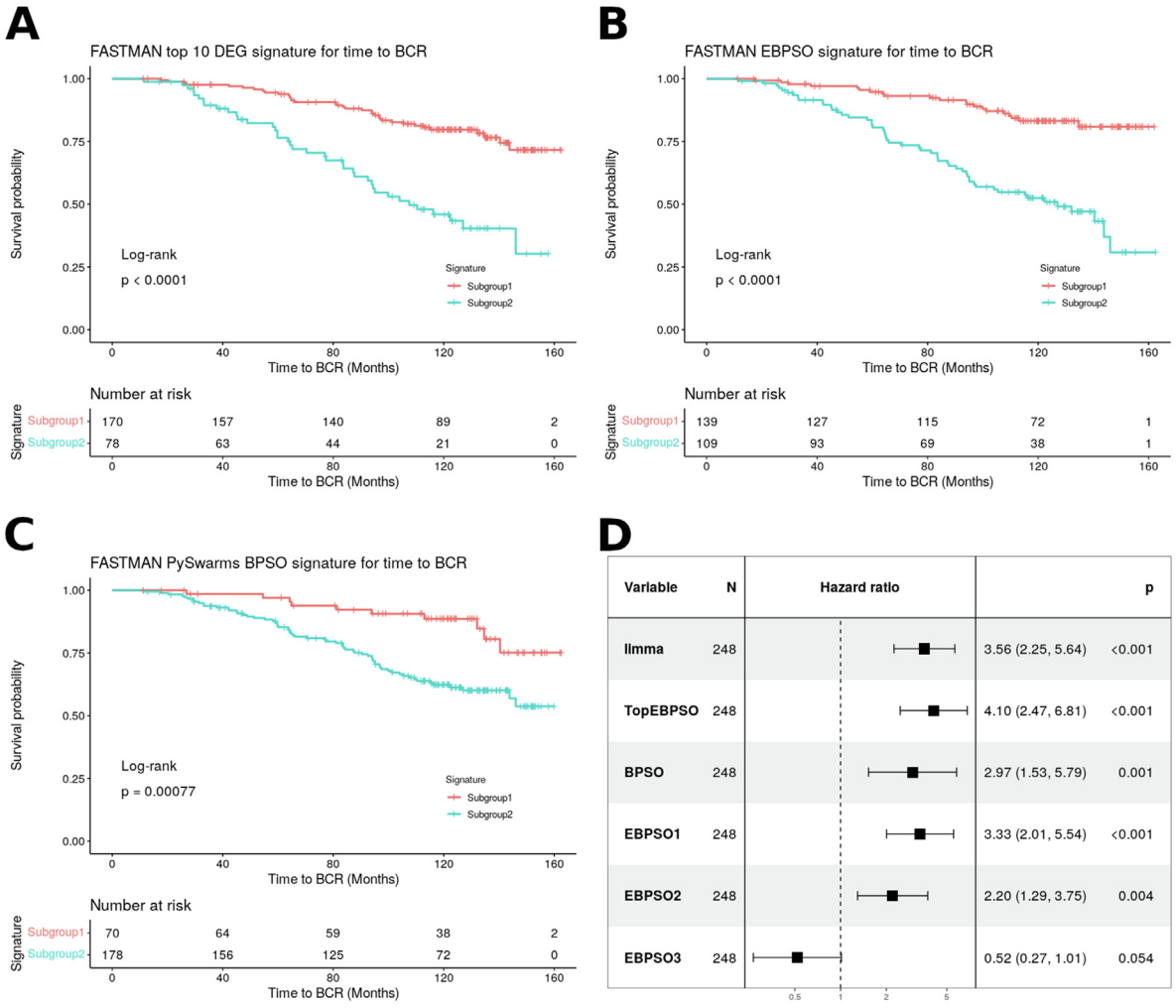


**Fig. 7.** Survival analysis of hierarchical clustering patient subgroups with selected gene signatures from EBPSO and *PySwarms* BPSO on GSE116918 FASTMAN. A. Kaplan-Meier plot of the subgroups for the top ten features selected based on p-value from *limma*. B. Kaplan-Meier plot of the subgroups for the candidate signature selected by EBPSO. C. Kaplan-Meier plot of the subgroups for the candidate signature selected by *PySwarms* BPSO. D. Forest plot of the signatures from *limma*, EBPSO (TopEBPSO), and *PySwarms* BPSO, and the top three ranked candidate signatures from a single run of EBPSO (EBPSO1-3; see Fig. 6E). Abbreviations: DEG, Differentially Expressed Gene; BCR, Biochemical Recurrence; EBPSO, Enhanced Binary Particle Swarm Optimization; BPSO, Binary Particle Swarm Optimization.

However, it has been reported that ELL2 is overexpressed in AR-negative neuroendocrine prostate cancer as an oncogene [44].

### 4.3. EBPSO for multiple signature selection on big data cohorts

EBPSO could be used for identifying a range of unique gene signatures in highly dimensional cohorts to help mine these ever increasing datasets. High throughput technologies allow for the generation of huge datasets at efficient costs, leading to what is being defined as the 'big data' era in bioinformatics [3]. As these datasets increase in information volume over time, traditional bioinformatics tools and software will struggle to fully landscape different key drivers in their biology amongst different predictors for clinical outcome and potential therapeutic benefit. Thus, new bioinformatics tools and software will need to be able to appropriately mine these vast datasets with accuracy, the ability to handle the amount of information, and produce a sufficient runtime for analysis [45]. This study utilises PSO as an evolutionary algorithm to provide unique candidate solutions for potential gene signature selection towards diagnostic, predictive, and prognostic biomarkers. The implementation of EBPSO furthers this by also demonstrating high accuracy, succinct signatures, and with improved runtimes. Other types of evolutionary algorithms have been utilised in this field also. Examples of this include the Atlas Correlation Explorer (ACE) [46]. ACE can identify patterns within TCGA between different matched -omics data types and clinical information. It has also demonstrated its ability to identify novel cancer biomarkers in big data cohorts with fast runtimes.

It has been suggested that gene signature discovery is often specific to its training data, and thus the inclusion or exclusion of some samples would produce a different gene signature [1]. As previously discussed, gene signature selecting is dominated by defining a single signature. This study has demonstrated how multiple unique candidate gene signatures of interest could potentially exist within a single transcriptomic cohort based on the same classification task. These unique signatures of interest would be selected by different end-users through characteristics such as its associated cost value, performance accuracy, number of genes, and the biology involved in the selected gene.

### 4.4. Study limitations

EBPSO has demonstrated its strong performance in identifying accurate gene signatures with limited number of features on both

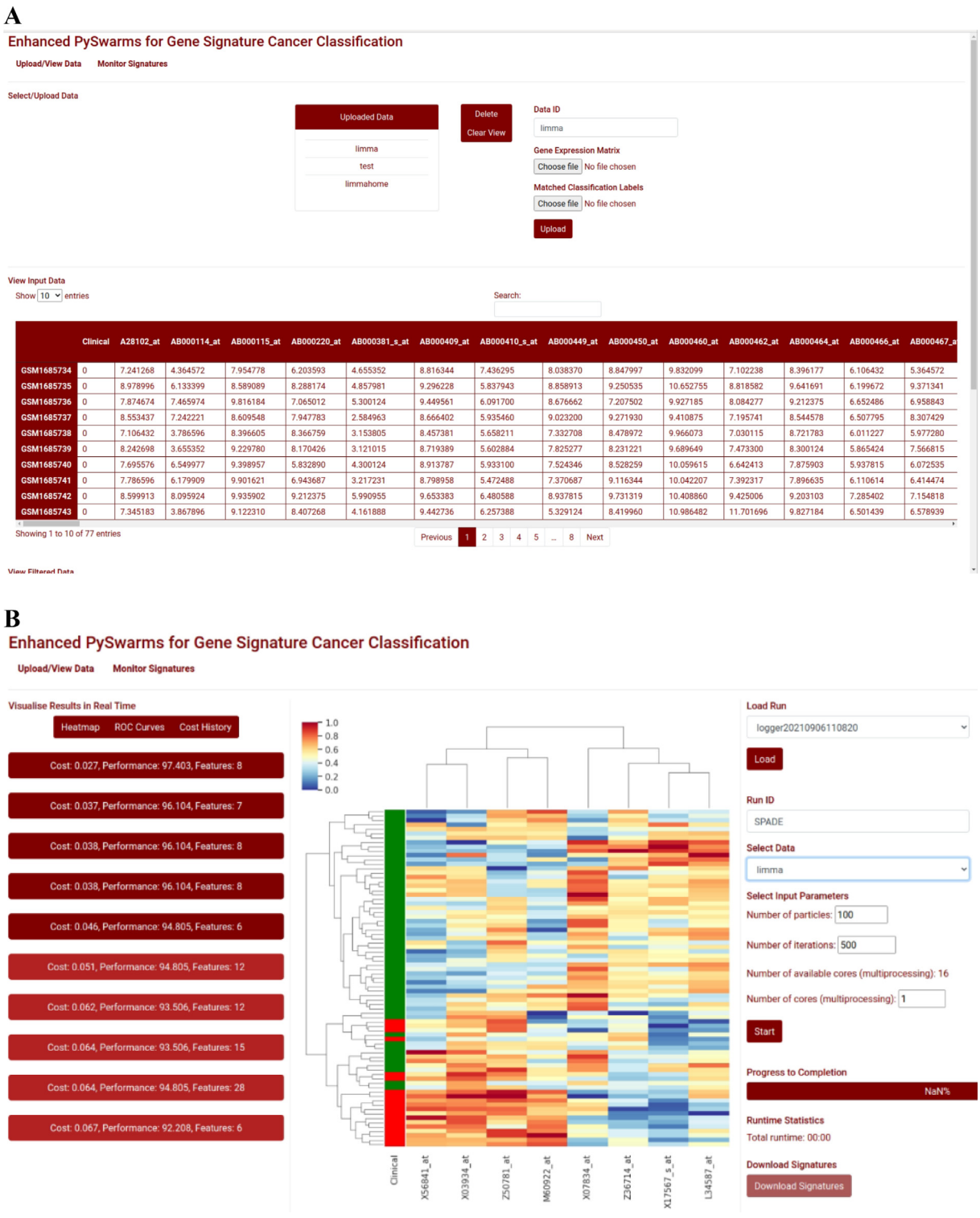


**Fig. 8.** EBPSO web-based analytical application with Flask. A. Upload/View Data homepage for loaded data sets to be previewed or deleted and new data sets to be loaded into the application's file system. B. Monitor Signatures page displaying: visualization options and an interactive gene signature performance cost leaderboard (left panel); real time visualizations (middle panel); and user input parameters including loading historical application runs and button for downloading the signatures from a completed application run (right panel). Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization.

simulated and clinical transcriptomics datasets. However, its performance on gene signature selection has only been applied to microarray transcriptomics. RNA-seq is a next generation sequencing (NGS) technique to capture the transcriptome [47]. RNA-seq has been adopted due to microarrays pitfalls which include cross-hybridization artifacts and poor quantification of lowly and highly expressed genes. Additionally, it has the ability to detect novel transcripts, allele-specific expression and splice junctions [48]. RNA-seq and microarray technology have been compared to demonstrate RNA-seq's ability to detect more DEGs with higher fold-changes [49]. EBPSO adapted *PySwarms* was developed to aid gene signature selection on highly dimensional datasets in pre-

**Table A1**
Candidate gene signature selection results over ten runs of EBPSO on simulated datasets. These simulated datasets were created based on binary and multi-class classification tasks.

| Run | Binary class | | Multi-class | |
|---|---|---|---|---|
| | Accuracy (%) | Genes | Accuracy (%) | Genes |
| 1 | 99.5 | 8 | 97.5 | 12 |
| 2 | 99.5 | 11 | 96.5 | 57 |
| 3 | 99.5 | 16 | 84 | 50 |
| 4 | 99.5 | 8 | 93.5 | 40 |
| 5 | 99.5 | 9 | 96.5 | 25 |
| 6 | 99.5 | 5 | 99 | 218 |
| 7 | 99.5 | 10 | 99 | 77 |
| 8 | 99.5 | 11 | 95 | 29 |
| 9 | 99.5 | 10 | 97.5 | 20 |
| 10 | 99.5 | 9 | 96 | 19 |
| Average ± S.D. | 99.5 ± 0 | 9.7 ± 2.8 | 96.5 ± 1.9 | 56.7 ± 59.8 |

Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization.

**Table A2**
Candidate gene signature selection results over ten runs of EBPSO on real patient gene expression cohorts DLBCL, GSE43358, and GSE116918 FASTMAN.

| Run | DLBCL | | GSE43358 | | GSE116918 FASTMAN | |
|---|---|---|---|---|---|---|
| | Accuracy (%) | Genes | Accuracy (%) | Genes | Accuracy (%) | Genes |
| 1 | 100 | 5 | 100 | 2 | 83.5 | 4 |
| 2 | 100 | 5 | 100 | 2 | 83.5 | 11 |
| 3 | 100 | 5 | 100 | 3 | 84.7 | 22 |
| 4 | 100 | 7 | 100 | 2 | 92.7 | 10 |
| 5 | 100 | 5 | 100 | 2 | 85.1 | 26 |
| 6 | 100 | 7 | 100 | 2 | 83.5 | 11 |
| 7 | 100 | 5 | 100 | 2 | 83.9 | 6 |
| 8 | 100 | 5 | 100 | 2 | 85.1 | 15 |
| 9 | 100 | 5 | 100 | 3 | 83.1 | 11 |
| 10 | 100 | 5 | 100 | 3 | 82.7 | 7 |
| Average ± S.D. | 100 ± 0 | 5.4 ± 0.8 | 100 ± 0 | 2.3 ± 0.5 | 83.8 ± 0.9 | 12.3 ± 7 |

Abbreviations: EBPSO, Enhanced Binary Particle Swarm Optimization; DLBCL, Diffuse Large B-Cell Lymphoma.

sent studies and for the future big data eta within bioinformatics. It is assumed that future studies will look to utilise RNA-seq for transcriptome capturing due to its advantages over microarray technologies [50]. Thus, it is crucial that it is properly validated on RNA-seq datasets to ensure its ability to identify accurate and succinct gene signatures on the transcriptome capturing technology.

### 4.5. Future perspectives

Further improvements could look at repurposing the original EBPSO implementation from a feature selection algorithm and towards a classification model development framework for additional functionality of the method. The features ranking method utilised in the original EBPSO implementation uses the information gain ratio technique to pre-select the top 500 features towards the classification labels on the full original dataset to be used as input for EBPSO. However, feature selection on the full original data when EBPSO uses the full dataset to evaluate PSO's feature selection ability introduces performance bias into these estimates. To overcome this and to better evaluate the classification models produced by EBPSO (as opposed to a feature selection technique), separate training and unseen testing datasets would need to be generated on the whole dataset before the feature ranking method. This would allow pre-selected features for input towards EBPSO to be independent from the final testing dataset.

## 5. Conclusions

This study has successfully adapted a PSO Python module in *PySwarms* towards EBPSO for greater efforts in predictive gene signature selection. EBPSO has been demonstrated to perform favourably for gene signature generation in comparison to conventional BPSO. This has been evaluated as having similar predictive accuracy performance, significantly smaller gene signature lengths, and dramatic increases in runtime to completion. Using real cancer transcriptomics cohorts, EBPSO has demonstrated the ability to identify accurate, succinct, and prognostically significant gene signatures that are unique from one another. This has been proposed as a promising alternative to overcome the issues regarding traditional single gene signature generation. Interpretation of key genes within the signatures provided biological insights into the candidate signatures associated functions that were well correlated to their cancer type. Within the DLBCL dataset for DLBCL against FL classification, HGMA1 was identified within the top two candidate signatures from EBPSO. HGMA1 has been associated as a master regulator of disease progression of FL patients towards the more aggressive DLBCL, and was upregulated for DLBCL samples in the DLBCL dataset. Despite the moderate predictive accuracy, but statistically significant prognostic ability, on the FASTMAN prostate cancer dataset for biochemical failure classification, the top candidate signature identified the FAP gene in its gene signature. FAP has been associated with metastatic disease in prostate cancer is seen to be expressed in prostate cancer stroma, and was upregulated for biochemical failure status in the prostate cancer dataset. This study proposes improving EBPSO by functioning as an integrated feature selection and classification model parameter selection technique for increased predictive performance of the candidate gene signature classification models. Additionally, development of downstream analysis for a selected candidate signature would be beneficial to better translate these gene signatures towards validation cohorts and the potential use within the clinic.

**Table B1**
Abbreviations of the official gene symbols from Table 5.

| Signature rank | Feature name | Full gene name |
|---|---|---|
| 1st ranked | FAP | Fibroblast Activation Protein alpha |
| | MX1 | MX dynamin like GTPase 1 |
| | RAB27B | RAB27B, member RAS oncogene family |
| | RGS16 | Regulator of G protein Signaling 16 |
| | UACA | Uveal Autoantigen with Coiled-coil domains and Ankyrin repeats |
| | ADC | — |
| | IGFBP3 | Insulin like Growth Factor Binding Protein 3 |
| | AKAP7 | A-Kinase Anchoring Protein 7 |
| | FCER1G | FC Epsilon Receptor IG |
| | ANO10 | Anoctamin 10 |
| | COL1A2 | Collagen type I Alpha 2 chain |
| | THBS1 | Thrombospondin 1 |
| | GLIPR1 | GLI Pathogenesis Related 1 |
| | **ELL2** | **Elongation factor for RNA polymerase II 2** |
| | MFSD4 | Major Facilitator Superfamily Domain containing 4 |
| | APPBP2 | Amyloid beta Precursor Protein Binding Protein 2 |
| | PDIA5 | Protein Disulfide Isomerase family A member 5 |
| | OR51D1 | Olfactory Receptor family 51 subfamily D member 1 |
| 2nd ranked | RTCA | RNA 3′-Terminal phosphate Cyclase |
| | TLL1 | Tolloid Like 1 |
| | SIAH1 | SIAH E3 ubiquitin protein ligase 1 |
| | ZNF382 | Zinc Finger protein 382 |
| | PPAPDC1B | Phosphatidic Acid Phosphatase type 2 Domain Containing 1B |
| | ASPN | Asporin |
| | CHRNA2 | Cholinergic Receptor Nicotinic Alpha 2 subunit |
| | LYZ | Lysozyme |
| | PCED1A | PC-Esterase Domain containing 1A |
| | RBP7 | Retinol Binding Protein 7 |
| | DESI2 | Desumoylating Isopeptidase 2 |
| | **PCAT4** | **Prostate Cancer Associated Transcript 4** |
| | SYNPO | Synaptopodin |
| 3rd ranked | ORL51L1 | — |
| | FNDC1 | Fibronectin type III Domain Containing 1 |
| | IFI44L | Interferon Induced protein 44 Like |
| | TMEM138 | Transmembrane protein 138 |
| | TRPM8 | Transient Receptor Potential cation channel subfamily M member 8 |
| | ZNF702P | Zinc Finger protein 702, Pseudogene |
| | PDCD1LG2 | Programmed Cell Death 1 Ligand 2 |
| | HOX19 | Homeobox-leucine zipper protein HOX19 |
| | CTSD | Cathepsin D |
| | MRPL17 | Mitochondrial Ribosomal Protein L17 |
| | SAMD3 | Sterile Alpha Motif Domain containing 3 |
| | TMSB10 | Thymosin Beta 10 |
| | SUSD4 | Sushi Domain containing 4 |
| | MPEG1 | Macrophage Expressed 1 |
| | SLFN5 | Schlafen Family member 5 |
| | **ELL2** | **Elongation factor for RNA polymerase II 2** |
| | PRSS27 | Serine Protease 27 |
| | **PCAT4** | **Prostate Cancer Associated Transcript 4** |
| | CRIP1 | Cysteine Rich Protein 1 |
| | ZNF613 | Zinc Finger protein 613 |
| | AIFM1 | Apoptosis Inducing Factor Mitochondria associated 1 |
| | GFM2 | GTP dependent ribosome recycling Factor Mitochondrial 2 |

## CRediT authorship contribution statement

**Ross G. Murphy:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Alan Gilmore:** Conceptualization, Writing – review & editing. **Seedevi Senevirathne:** Conceptualization, Methodology. **Paul G. O'Reilly:** Writing – original draft, Writing – review & editing. **Melissa LaBonte Wilson:** Supervision, Resources, Funding acquisition. **Suneil Jain:** Supervision, Resources, Funding acquisition. **Darragh G McArt:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Overview of candidate gene signatures from EBPSO over ten runs

In simulated data sets, EBPSO demonstrated excellent average accuracy performance of at least 96.5 %, whilst still producing succinct signatures with an average number of features selected at 56.7 at its maximum (Table A.1). In the binary classification simulated dataset, highly accurate signatures were selected with feature numbers as small as five. The multi-class simulated dataset produced similar performances in comparison to the binary classification dataset. Much larger feature numbers in the selected signatures in the multi-class data set were produced. The average number of features in a signature was 56.7. The standard deviation for this was also very high at 59.8, higher than the average number of features in a signature. This is showcased through a mix of differently sized signatures, both significantly small and large, across the ten runs of EBPSO.

In the selected clinical data sets, EBPSO also demonstrates highly accurate and compact selected signatures much like the simulated datasets. 100 % accuracy was achieved in all of the ten runs of EBPSO for the DLBCL and the GSE43358 HER2 against TN breast cancer data set (Table A.2). Poorer signature performances were demonstrated on the GSE116918 FASTMAN prostate cancer data set for classification on biochemical failure status. The greatest accuracy for a signature for this data set was at 85.1 %.

## Appendix B. Abbreviations of the official gene symbols from Table 5.

Table B1

## Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.033.

## References

[1] Cun Y, Fröhlich H. Biomarker gene signature discovery integrating network knowledge. Biology 2012;1(1):5–17. https://doi.org/10.3390/biology1010005.
[2] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–6. https://doi.org/10.1038/415530a.
[3] Greene CS, Tan J, Ung M, Moore JH, Cheng C. Big data bioinformatics. J Cell Physiol 2014;229(12):1896–900. https://doi.org/10.1002/jcp.24662.
[4] Clough E, Barrett T. The gene expression omnibus database. Methods Mol. Biol. (Clifton NJ) 2016;1418:93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
[5] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31(1):68–71. https://doi.org/10.1093/nar/gkg091.
[6] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764.
[7] Yu JX, Sieuwerts AM, Zhang Y, Martens JW, Smid M, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer 2007;7:182. https://doi.org/10.1186/1471-2407-7-182.
[8] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics (Oxford, England) 2007;23(19):2507–17. https://doi.org/10.1093/bioinformatics/btm344.

[9] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97 (1–2):273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.
[10] Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 2004;31(2):91–103. https://doi.org/10.1016/j.artmed.2004.01.007.
[11] Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia 1995;4:1942–8.
[12] Mason K, Duggan J, Howley E. Multi-objective dynamic economic emission dispatch using particle swarm optimisation variants. Neurocomputing 2017;270:188–97. https://doi.org/10.1016/j.neucom.2017.03.086.
[13] El-Maleh AH, Sheikh AT, Sait SM. Binary particle swarm optimization (BPSO) based state assignment for area minimization of sequential circuits. Appl Soft Comput 2013;13(12):4832–40. https://doi.org/10.1016/j.asoc.2013.08.004.
[14] Dara, S. & Banka, H. (2014). A binary PSO feature selection algorithm for gene expression data. 2014 International Conference on Advances in Communication and Computing Technologies (ICACACT 2014), Mumbai; 1-6. 10.1109/EIC.2015.7230734.
[15] Xi M, Sun J, Liu L, Fan F, Wu X. Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. Comput Math Methods Med 2016;2016:3572705. https://doi.org/10.1155/2016/3572705.
[16] Jain I, Jain VK, Jain R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Appl Soft Comput 2018;62:203–15. https://doi.org/10.1016/j.asoc.2017.09.038.
[17] Chen KH, Wang KJ, Tsai ML, Wang KM, Adrian AM, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinf 2014;15:49. https://doi.org/10.1186/1471-2105-15-49.
[18] Mohamad MS, Omatu S, Deris S, Yoshioka M, Abdullah A, Ibrahim Z. An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes. Algorith Mol Biol: AMB 2013;8(1):15. https://doi.org/10.1186/1748-7188-8-15.
[19] Gönen M. Statistical aspects of gene signatures and molecular targets. Gastrointestinal Cancer Res: GCR 2009;3(2 Suppl):S19–21.
[20] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics (Oxford, England) 2005;21 (2):171–8. https://doi.org/10.1093/bioinformatics/bth469.
[21] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. Scikit-learn: machine learning in python. J Machine Learn 2011;12:2825–30. https://doi.org/10.5555/1953048.2078195.
[22] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8(1):68–74. https://doi.org/10.1038/nm0102-68.
[23] Fumagalli D, Blanchet-Cohen A, Brown D, Desmedt C, Gacquer D, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. BMC Genomics 2014;15(1):1008. https://doi.org/10.1186/1471-2164-15-1008.
[24] Jain S, Lyons CA, Walker SM, McQuaid S, Hynes SO, et al. Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. Ann Oncol 2018;29 (1):215–22. https://doi.org/10.1093/annonc/mdx637.
[25] Wu X, Lv D, Lei M, Cai C, Zhao Z, et al. A 10-gene signature as a predictor of biochemical recurrence after radical prostatectomy in patients with prostate cancer and a Gleason score ≥7. Oncology letters 2020;20(3):2906–18. https://doi.org/10.3892/ol.2020.11830.
[26] Shi R, Bao X, Weischenfeldt J, Schaefer C, Rogowski P, et al. A Novel Gene Signature-Based Model Predicts Biochemical Recurrence-Free Survival in Prostate Cancer Patients after Radical Prostatectomy. Cancers 2019;12(1):1. https://doi.org/10.3390/cancers12010001.
[27] Miranda. PySwarms: a research toolkit for Particle Swarm Optimization in Python. J Open Source Software 2018;3(21):433. https://doi.org/10.21105/joss.00433.
[28] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, et al. Array programming with NumPy. Nature 2020;585:357–62. https://doi.org/10.1038/s41586-020-2649-2.
[29] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43(7):e47.
[30] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1007/BF00994018.
[31] Allen DM. The relationship between variable selection and data agumentation and a method for prediction. Technometrics 1974;16(1):125–7. https://doi.org/10.1080/00401706.1974.10489157.
[32] Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng 2007;9 (3):90–5. https://doi.org/10.1109/MCSE.2007.55.
[33] Kristiyanti DA, Wahyudi M. Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review. In: 2017 5th International Conference on Cyber and IT Service Management (CITSM). p. 1–6. https://doi.org/10.1109/CITSM.2017.8089278.
[34] Yang L, Roberts D, Takhar M, Erho N, Bibby B, et al. Development and validation of a 28-gene hypoxia-related prognostic signature for localized prostate cancer. EBioMedicine 2018;31:182–9. https://doi.org/10.1016/j.ebiom.2018.04.019.
[35] Varnier R, Sajous C, de Talhouet S, Smentek C, Péron J, et al. Using breast cancer gene expression signatures in clinical practice: unsolved issues. Ongoing Trials Future Perspect Cancers 2021;13(19):4840. https://doi.org/10.3390/cancers13194840.
[36] Bisikirska B, Bansal M, Shen Y, Teruya-Feldstein J, Chaganti R, Califano A. Elucidation and pharmacological targeting of novel molecular drivers of follicular lymphoma progression. Cancer Res 2016;76(3):664–74. https://doi.org/10.1158/0008-5472.CAN-15-0828.
[37] Wang R, Shen J, Wang Q, Zhang M. Bortezomib inhibited the progression of diffuse large B-cell lymphoma via targeting miR-198. Biomed Pharmacother 2018;108:43–9. https://doi.org/10.1016/j.biopha.2018.08.151.
[38] Glas AM, Kersten MJ, Delahaye LJ, Witteveen AT, Kibbelaar RE, Velds A, et al. Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. Blood 2005;105 (1):301–7. https://doi.org/10.1182/blood-2004-06-2298.
[39] Björck E, Ek S, Landgren O, Jerkeman M, Ehinger M, et al. High expression of cyclin B1 predicts a favorable outcome in patients with follicular lymphoma. Blood 2005;105(7):2908–15. https://doi.org/10.1182/blood-2004-07-2721.
[40] Hintz HM, Gallant JP, Vander Griend DJ, Coleman IM, Nelson PS, LeBeau AM. Imaging fibroblast activation protein alpha improves diagnosis of metastatic prostate cancer with positron emission tomography. Clin Cancer Res 2020;26 (18):4882–91. https://doi.org/10.1158/1078-0432.CCR-20-1358.
[41] Tuxhorn JA, Ayala GE, Smith MJ, Smith VC, Dang TD, Rowley DR. Reactive stroma in human prostate cancer: induction of myofibroblast phenotype and extracellular matrix remodeling. Clin Cancer Res 2002;8(9):2912–23.
[42] Yang T, Jing Y, Dong J, Yu X, Zhong M, et al. Regulation of ELL2 stability and polyubiquitination by EAF2 in prostate cancer cells. Prostate 2018;78 (15):1201–12. https://doi.org/10.1002/pros.23695.
[43] Zhong M, Pascal LE, Cheng E, Masoodi KZ, Chen W, et al. Concurrent EAF2 and ELL2 loss phenocopies individual EAF2 or ELL2 loss in prostate cancer cells and murine prostate. Am J Clin Experiment Urol 2018;6(6):234–44.
[44] Wang Z, Pascal LE, Chandran UR, Chaparala S, Lv S, et al. ELL2 is required for the growth and survival of AR-negative prostate cancer cells. Cancer Manage Res 2020;12:4411–27. https://doi.org/10.2147/CMAR.S248854.
[45] Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods (San Diego, Calif) 2016;111:21–31. https://doi.org/10.1016/j.ymeth.2016.08.014.
[46] Gilmore AR, Alderdice M, Savage KI, O'Reilly PG, Roddy AC, et al. ACE: A workbench using evolutionary genetic algorithms for analyzing association in TCGA. Cancer Res 2019;79(8):2072–5. https://doi.org/10.1158/0008-5472.CAN-18-1976.
[47] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10(1):57–63. https://doi.org/10.1038/nrg2484.
[48] Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Zhang W, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol 2015;16(1):133. https://doi.org/10.1186/s13059-015-0694-1.
[49] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS one 2014;9(1):e78644.
[50] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf 2013;14:91. https://doi.org/10.1186/1471-2105-14-91.