A Novel Filter-Wrapper Hybrid Greedy Ensemble Approach Optimized Using the Genetic Algorithm to Reduce the Dimensionality of High-Dimensional Biomedical Datasets

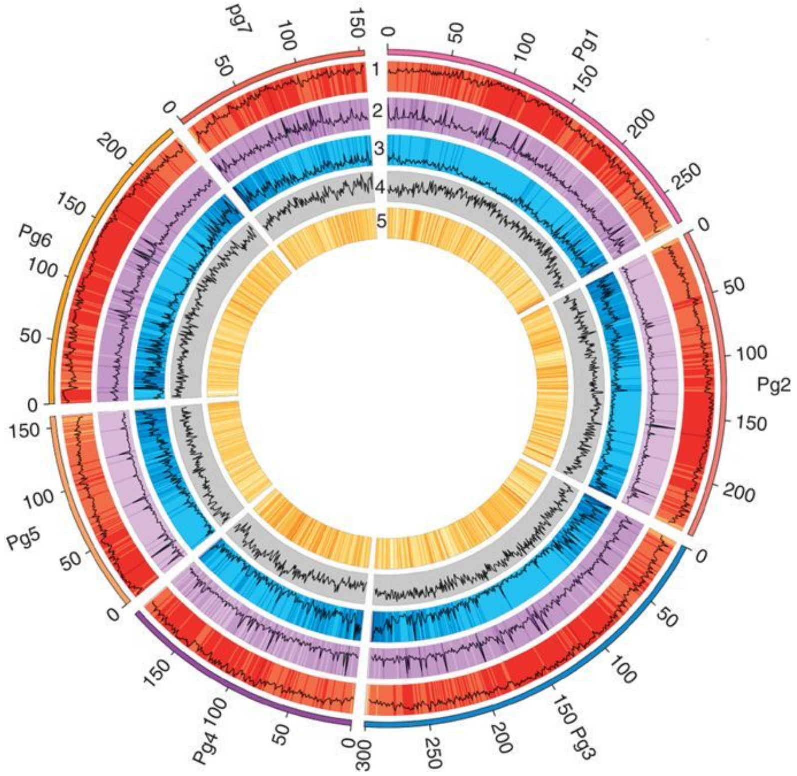
Tushaar Gangavarapu
Nagamma Patil

Department of Information Technology

Please cite this article as: T. Gangavarapu and N. Patil, “A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,” Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

Introduction

- Whether more information leads to more informed decision making?
  - Irrelevancy: low learnability
  - Redundancy: high training time
  - Noise: classification errors
  - Computation cost: expensive
- How to choose a feature selection approach that matches the problem structure and captures the inherent patterns in the data?
  - Intuition-based: unreliable
  - Brute force: infeasible
  - Determine heuristically: convergence?



Large number of features can be extracted using genome sequences*

* Varshney et al. "Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments." Nature (2017).

Literature Survey

Work(s)	Methodology	Highlights	Remarks
Tu et al. 2019 [1]	Multi-strategy grey wolf optimization with three search strategies and parameter optimization	Effective feature Selection with disperse foraging	Heuristic search with convergence problem
Dong et al. 2018 [2]	Hybrid genetic algorithm with feature granulation	Faster feature selection due to bottom-up search	Heuristic search not guided by correlation
Masood et al. 2017 [3]	Hybrid and wrapper algorithms based on an incremental search using an ELM classifier	RIG and ELM based feature selection with good performance	Extremely data specific
Min et al. 2014 [4]	Backtracking and heuristic search algorithm to search for optimal feature subspaces	Similar performance of backtracking and heuristic algorithms	Heuristic search with convergence problem
Ekwevugbe et al. 2013 [5]	Filter approaches to determine the optimal feature subspace	Correlation-based approaches with faster training	Classifier independence limits accuracy

Outcome of Literature Survey

- Which feature selection?: multiple filter, wrapper, embedded, hybrid, and heuristic approaches; which one accurately matches the problem structure? always an issue
- Filter-based approaches: faster computation, but heavily dependence on correlation and classifier independence limits their accuracy
- Wrapper-based, embedded, and hybrid approaches: domain adaptability and high computational cost of training, but reliable performance
- Metaheuristic search approaches: population-based mechanism guides the search, but convergence problem and correlation-unguided search can be a bottleneck!
- Need for an ensemble?: set of predetermined feature selection approaches
  - Voting-based ensemble?: simply a brute force ensemble inefficient
  - Greedy ensemble?: penalize bad-performing selection methods and their features
- Time and accuracy tradeoff: use a hybrid of filter and wrapper approaches

Key Contributions

Aim: Generate an optimal and instructive feature subspace with a lower computational complexity

- Design of a filter-wrapper hybrid ensemble selection approach that kindles an optimal feature subspace by greedily combining the subspaces generated by various predetermined feature selection techniques based on specific performance dependant penalty parameters
- Leveraging heuristic search strategies such as greedy parameter-wise optimization and GA to determine the optimal values of the penalty factors which affect how different feature subspaces are ensembled to engender an optimal feature subspace
- We present detailed benchmarking results of our hybrid greedy ensemble feature selection approach on three distinct high-dimensional biomedical datasets. Our experimental results indicate the efficiency and robustness of the proposed approach over the base feature selection methods, and other prolific filter and wrapper methods

Biomedical Datasets

Dataset	#Samples	#Dimensions	#Classes (#samples per class)	Balanced?
TIS[6]	13,375	927	2 (3,312; 10,063)	No
Skin Cancer[7]	10,015	2,352	7 (327; 514; 1,099; 115; 6,705; 142; 1,113)	No
Seizure[8]	11,500	179	5 (2,300; 2,300; 2,300; 2,300; 2,300)	Yes

- TIS: extracted from genomic sequences of a selected set of vertebrates (from GenBank), involving the process of finding sites where translation of mRNA to proteins initiates
- Skin Cancer: extracted from the pixel information of 28×28 RGB dermatoscopic images of the Skin Cancer MNIST: HAM10000 dataset
- Epileptic Seizure Recognition: consists of five sets of single channel 23.6 seconds long electroencephalogram segments that are weakly stationary and selected after a visual inspection of artifacts

Please cite this article as: T. Gangavarapu and N. Patil, "A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets," Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

Proposed Methodology

Feature selection using the selected methods

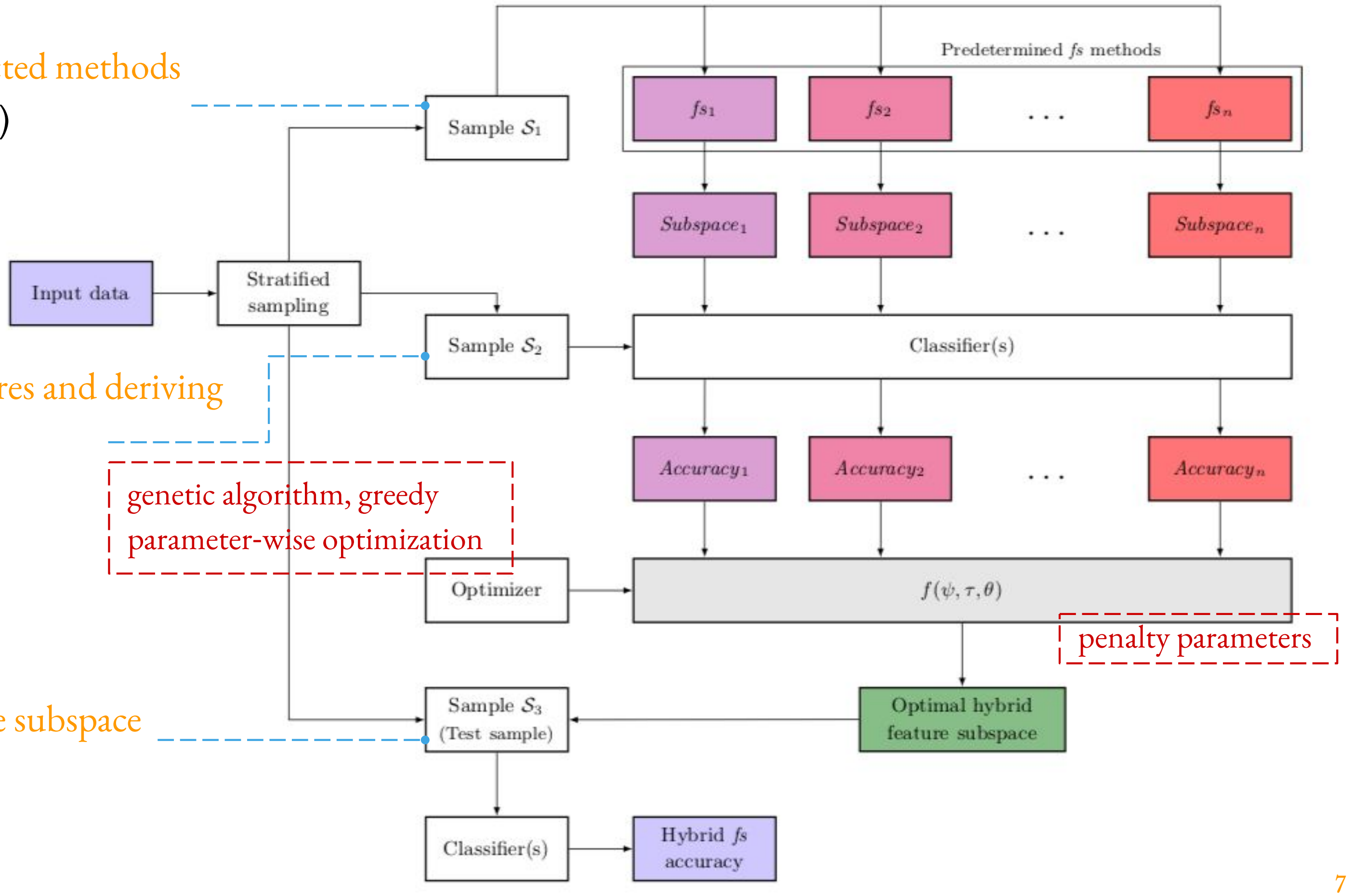
Feature space: #features(dataset)

Evaluation of the selected features and deriving the hybrid feature subspace

Feature space: #features(S_1)

Evaluation of the hybrid feature subspace

Feature space: hybrid

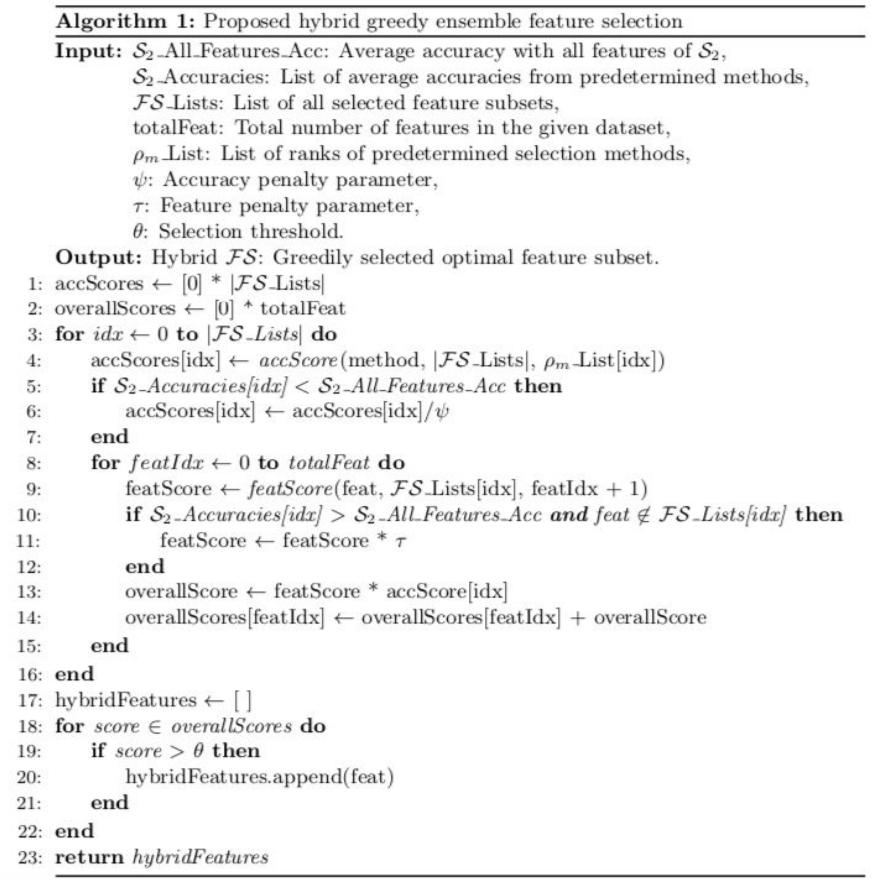


genetic algorithm, greedy parameter-wise optimization

penalty parameters

Please cite this article as: T. Gangavarapu and N. Patil, “A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,” Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

Filter-Wrapper Hybrid Greedy Ensemble



```
Algorithm 1: Proposed hybrid greedy ensemble feature selection
Input: $\mathcal{S}_2$_All_Features_Acc: Average accuracy with all features of $\mathcal{S}_2$,
       $\mathcal{S}_2$_Accuracies: List of average accuracies from predetermined methods,
       $\mathcal{FS}$_Lists: List of all selected feature subsets,
       totalFeat: Total number of features in the given dataset,
       $\rho_m$_List: List of ranks of predetermined selection methods,
       $\psi$: Accuracy penalty parameter,
       $\tau$: Feature penalty parameter,
       $\theta$: Selection threshold.
Output: Hybrid $\mathcal{FS}$: Greedily selected optimal feature subset.
accScores ← [0] * |$\mathcal{FS}$_Lists|
overallScores ← [0] * totalFeat
for idx ← 0 to |$\mathcal{FS}$_Lists| do
    accScores[idx] ← accScore(method, |$\mathcal{FS}$_Lists|, $\rho_m$_List[idx])
    if $\mathcal{S}_2$_Accuracies[idx] < $\mathcal{S}_2$_All_Features_Acc then
        accScores[idx] ← accScores[idx]/$\psi$
    end
    for featIdx ← 0 to totalFeat do
        featScore ← featScore(feat, $\mathcal{FS}$_Lists[idx], featIdx + 1)
        if $\mathcal{S}_2$_Accuracies[idx] > $\mathcal{S}_2$_All_Features_Acc and feat $\notin$ $\mathcal{FS}$_Lists[idx] then
            featScore ← featScore * $\tau$
        end
        overallScore ← featScore * accScore[idx]
        overallScores[featIdx] ← overallScores[featIdx] + overallScore
    end
end
hybridFeatures ← [ ]
for score $\in$ overallScores do
    if score > $\theta$ then
        hybridFeatures.append(feat)
    end
end
return hybridFeatures
```

- Scoring of features (featScore) and selection methods (accScore):

$$featScore(f, \mathcal{FS}, \rho_f) = \begin{cases} \frac{|\mathcal{FS}|-\rho_f+1}{|\mathcal{FS}|}, & f \in \text{ranked } \mathcal{FS} \\ \frac{1}{|\mathcal{FS}|}, & f \in \text{unranked } \mathcal{FS} \\ \frac{-1}{|\mathcal{FS}|}, & f \notin \mathcal{FS} \end{cases}$$

$$accScore(m, \mathcal{M}, \rho_m) = \frac{|\mathcal{M}| - \rho_m + 1}{|\mathcal{M}|}$$

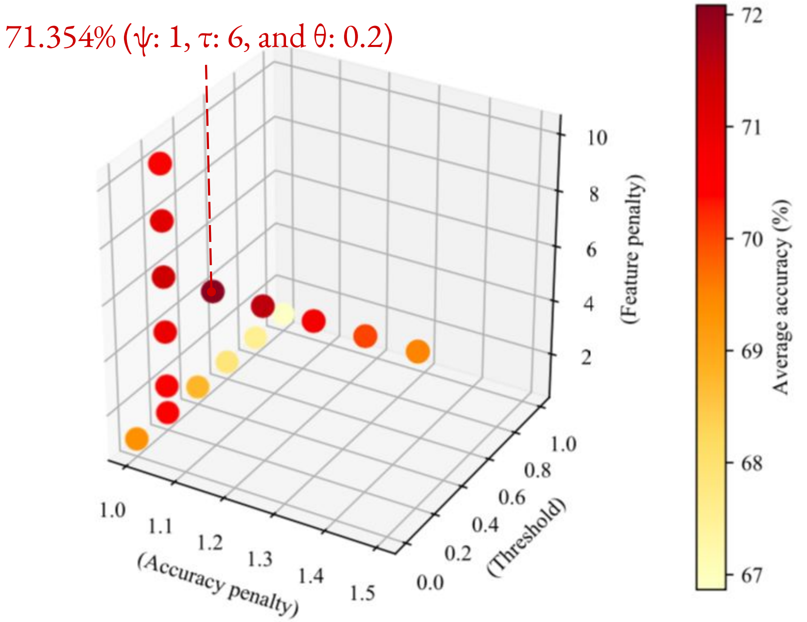
- Penalty parameters for greedy ensembling of base feature subspaces
    - Accuracy penalty: accScore/ψ
    - Feature penalty: featScore×τ
- Overall feature scoring and hybrid feature selection (θ):

$$overallScore(f, \mathcal{M}) = \sum_{m}^{\mathcal{M}} featScore(f) \times accScore(m)$$

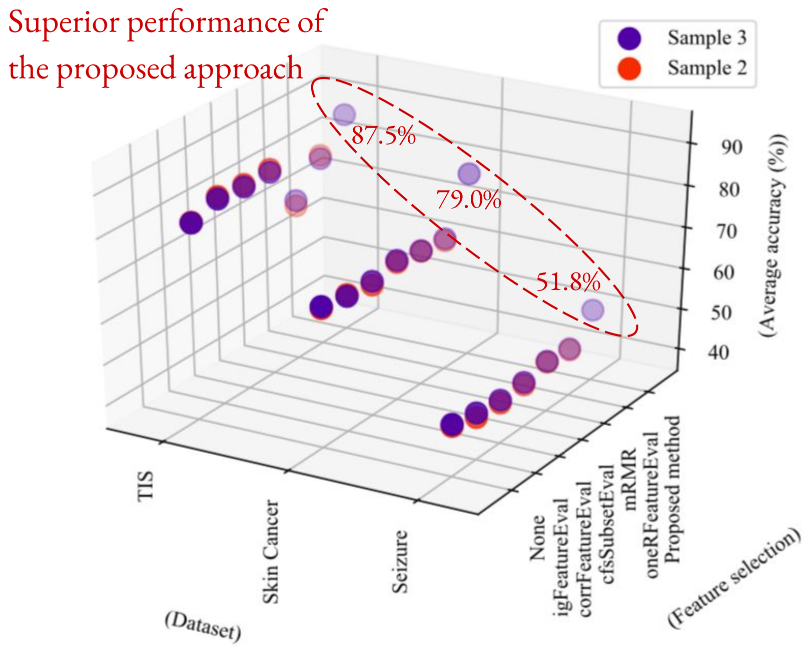
- Optimization of penalty parameters

Please cite this article as: T. Gangavarapu and N. Patil, “A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,” Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

Results with Optimization of Penalty Parameters


71.354% (ψ: 1, τ: 6, and θ: 0.2)

The effect of ψ, τ, and θ on Skin Cancer dataset
(greedy parameter-wise optimization)


Superior performance of the proposed approach

The effect of ψ, τ, and θ on Skin Cancer dataset
(genetic algorithm (N = 50, p_c = 0.6, p_m = 0.1))

Please cite this article as: T. Gangavarapu and N. Patil, “A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,” Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

Results with Genetic Optimization: Comparison

Dataset	Chromosome			Sample (S_3) Average Accuracy (%)	
	ψ	τ	θ	Base Selection Method (Highest)	**Proposed Ensemble**
TIS[6]	6.08	20.83	0.78	IgFeatureEval: 83.948	87.449
Skin Cancer[7]	1.07	7.72	0.14	CfsSubsetEval: 68.534	78.912
Seizure[8]	1.39	2.58	0.01	None: 47.131	51.811

Dataset	Sample (S_3) Average Accuracy (%)				
	Filter Approaches		Wrapper Approaches		Proposed Greedy Ensemble
	RF Feature Importance	Chi-Square Test	RFE + SVM (Linear)	RFE + SVM (RBF)	
TIS[6]	84.182	84.220	79.968	83.290	87.449
Skin Cancer[7]	67.356	67.794	67.307	67.375	78.912
Seizure[8]	46.601	46.299	46.818	50.450	51.811

Conclusions and Future Work

- Proposed a penalty based filter-wrapper hybrid greedy ensemble approach to facilitate optimal feature selection
- Ensemble greedily selects the features from the subspaces obtained from the predetermined base selection methods
- Specific performance dependent penalty parameters were used to penalize the base feature subspaces essential to achieve the optimal ensembling of those subspaces
- At any point in time, only a stratified sample and not the entire dataset is not used for computation; the computational complexity is significantly reduced
- We leverage effective heuristic search strategies including the greedy parameter-wise optimization and the GA to obtain optimal values of the penalty parameters
- The proposed method introduces additional (penalty) parameters which require prior training to obtain the optimal setting in advance

Please cite this article as: T. Gangavarapu and N. Patil, "A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets," Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.

References

[1] Tu, Qiang, Xuechen Chen, and Xingcheng Liu. "Multi-strategy ensemble grey wolf optimizer and its application to feature selection." Applied Soft Computing 76 (2019): 16-30. Accessible: sciencedirect/science/article/pii/S1568494618306793.

[2] Dong, Hongbin, et al. "A novel hybrid genetic algorithm with granular information for feature selection and optimization." Applied Soft Computing 65 (2018): 33-46. Accessible: sciencedirect/science/article/pii/S1568494618300048.

[3] Masood, Mustafa K., Yeng Chai Soh, and Chaoyang Jiang. "Occupancy estimation from environmental parameters using wrapper and hybrid feature selection." Applied Soft Computing 60 (2017): 482-494.

[4] Min, Fan, Qinghua Hu, and William Zhu. "Feature selection with test cost constraint." International Journal of Approximate Reasoning 55.1 (2014): 167-179.

[5] Ekwevugbe, Tobore, Neil Brown, and Vijay Pakka. "Real-time building occupancy sensing for supporting demand driven hvac operations." (2013). Accessible: oaktrust.library.tamu.edu/handle/1969.1/151431.

[6] Pedersen, Anders Gorm, and Henrik Nielsen. "Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis." Ismb. Vol. 5. 1997. Accessible: nus/~wongls/courses/cs2220/2008/TIS-data.html.

[7] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." Scientific data 5 (2018): 180161.

[8] Andrzejak, Ralph G., et al. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state." Physical Review E 64.6 (2001): 061907.

Please cite this article as: T. Gangavarapu and N. Patil, “A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets,” Applied Soft Computing Journal (2019) 105538, https://doi.org/10.1016/j.asoc.2019.105538.