Author Response

We thank three reviewers for their valuable feedback. We address the comments and the concerns as follows.

To Reviewer #2

a) Feature independence. Our model does not work on the feature independence assumption but in a performance-driven manner. If a subset of combined features yields the optimal performance, the feature subset will be selected. Our experiments suggest that the importance of a feature correlated with others in the selected subset is ranked higher than that of an independent feature less relevant to the target. It is hence not biased towards linear solutions.

b) Related work. (1) REINFORCE and LTPA were proposed for instancewise FIR while ours tackles populationwise FIR problems. In fact, instancewise FIR for local explanations is quite distinct from populationwise FIR for global explanations and converting instancewise to populationwise FIR requires non-trivial mechanisms (see https://arxiv.org/pdf/1907.03039.pdf). Thus, we cannot directly compare ours to such instancewise methods. (2) Technically, LTPA does not work on input features and hence cannot conduct feature selection. REINFORCE works only for visual input while ours works for different input types including visual input. REINFORCE was implemented with reinforcement learning due to its non-differentiable nature while our model is carried out with supervised learning.

c) Connection to evolutionary computation (EC). We really appreciate your insight by understanding our work from an EC perspective. This insight may allow us to highlight our contributions from another angle: for feature selection, (1) ours uses a single learning model (enabling different feature masks to be used simultaneously during learning) to carry out the functionality of a population of learning models in EC; (2) instead of purely stochastic operations on population in EC, ours uses a more efficient gradient-guided local stochastic search strategy. We will add this insight in revision.

d) Computational cost. Indeed, computational cost of our dual-net model is high due to the use of two DNNs and the alternate learning routine. We will re-run all experiments on the same environment and report detailed results.

e) Performance and motivation. (1) Our model works well for a large feature set when there are enough training examples required by deep learning. For demonstration, we have just conducted an experiment on the UCI gene expression cancer RNA-Seq data set (https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq), where there are 801 instances and 20,531 features. With 4-fold cross-validation (450, 150 and 201 instances used for training, validation and test, respectively), our model yields $99.38 \pm 0.00\%$ testing accuracy on 49 selected features ($s = 49$). In literature, to the best of our knowledge, the best performance in the same settings on this data set is 98.81% (see Highlight and Table 7, https://doi.org/10.1016/j.ygeno.2019.11.004). For your information, the code provided by the authors of CCM [14] (the strongest method in our comparative study) does not work on our server as it requires 158GB+ memory for this data set. (2) The motivation of our training procedure is generally described in paragraph 1 of Sect. 3.3 and the input-gradient guided local search idea was motivated by the work presented in [12], as stated in Phase II-A.

To Reviewer #3

1. Convergence analysis. We agree to this point. We will summarize empirical observations in the revised main text and make a formal convergence analysis of our alternate learning algorithm in our ongoing research.

2. Connection to latest work. Thanks for pointing out two papers related to our work. Our work distinguishes those from the following aspects: a) those methods yield populationwise FIR by aggregating or integrating instancewise FIR in a sub-optimal manner, while our models directly learns populationwise FIR in an optimal way; b) for feature selection, those methods work as filtering so another learner has to be re-trained on the selected subset for a target task, while ours works as embedding by accomplishing feature selection and a target task together in an end-to-end manner. We will make a connection to two papers and report comparative results between ours and those methods in revision.

3. Instancewise vs. populationwise. We entirely agree to this point. We will be extending our model to instancewise FIR and study a connection between population and instancewise FIR in our ongoing work.

4. Computational cost and scalability. a) Regarding the computational overhead issue, see our response **d)** to reviewer #2; b) We will investigate scalability issues although any deep learning models effectively working for a large dataset can be used as an operator in our dual-net model.

To Reviewer #4

Number of important features. Our problem formulation for a fixed s is a common setting in feature selection, e.g., CCM [14]. When the ground-truth is unknown in real applications, results yielded by different s values (sub-solutions to feature selection) are still useful to reveal some meaningful relationship between selected features and the target. To find out the optimal value of s, the common setting allows a model like ours to work on different s values in parallel.

Other questions. a) in all the experiments, $2d$ is the input dimension of the operator in ours while d is the input dimension of all other methods; b) In Eq.(2b), dividing 2 in the MSE loss is default to facilitate the gradient computation.