Gene selection to predict the cancer type from genome-scale CRISPR–Cas9 screens

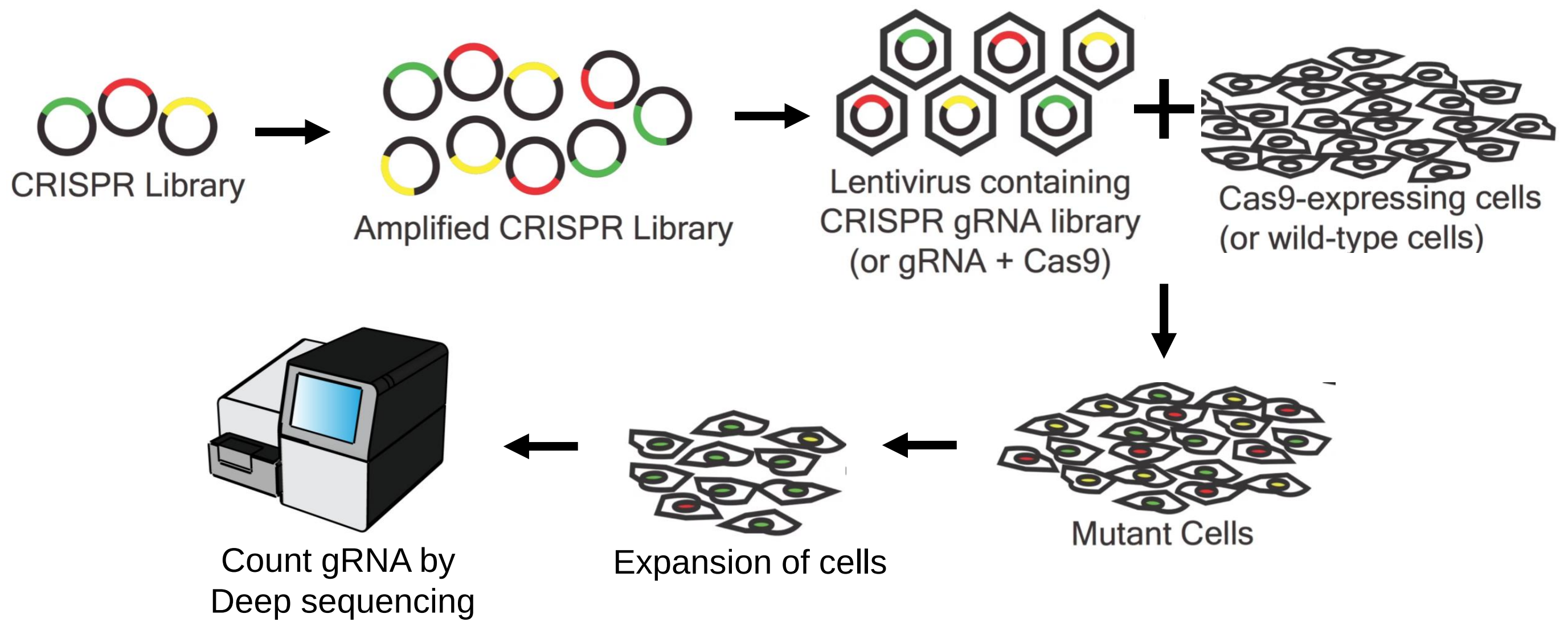
HYUNG JUN YANG [1] and HONG-PYO LEE[2]
[1]Department of Energy Resources Enigneering, Stanford University
[2]Department of Mechanical Engineering, Stanford University

BACKGROUND AND SIGNIFICANCE

Cancer is a complex disease derived from genetic and epigenetic mutations that develop uncontrollable cell proliferation. With sequencing and genetic techniques remarkably advancing, numerous mutations has been founded across diverse cancer types. DNA micro-arrays has been widely applied to observe thousands of mutational genes simultaneously in cancers and determine whether those genes are active, hyperactive or silent in normal or cancerous tissue. However, it is not clear whether such mutations and genes in cancers are functional cancer drivers. Therefore, a central challenge to the development of new cancer therapies is to systematically investigate the role of these mutational genes to derive cancer uncontrollable growth.

As the CRISPR-Cas9 system has emerged as a powerful tool for genome editing and transcriptional regulation of specific genes in a variety of cancer types, this technique has been remarkably improved the accuracy of testing functional roles of genes to derive cancer growth. Extraordinary efforts have characterized cancer growth dependencies of all genes using genome-scale in vitro CRISPR screens in hundreds of cancer cell lines. Since this new technique generates huge amounts of raw data, new analytical methods must be developed to sort out whether each cancer type have distinctive signatures of gene function on cancer growth over other types of cancer cells.

In this work, we applied machine learning algorithms to find the optimal set to classify cancer type with raw data with hypothesis that each cancer type has a distinct set of genes which represent whole growth phenotype of the specific cancer. Our primary goal is to identify a minimal feature (gene) set that still achieves reasonable classification of cancer type, so that feature selection is the key problem in this project.

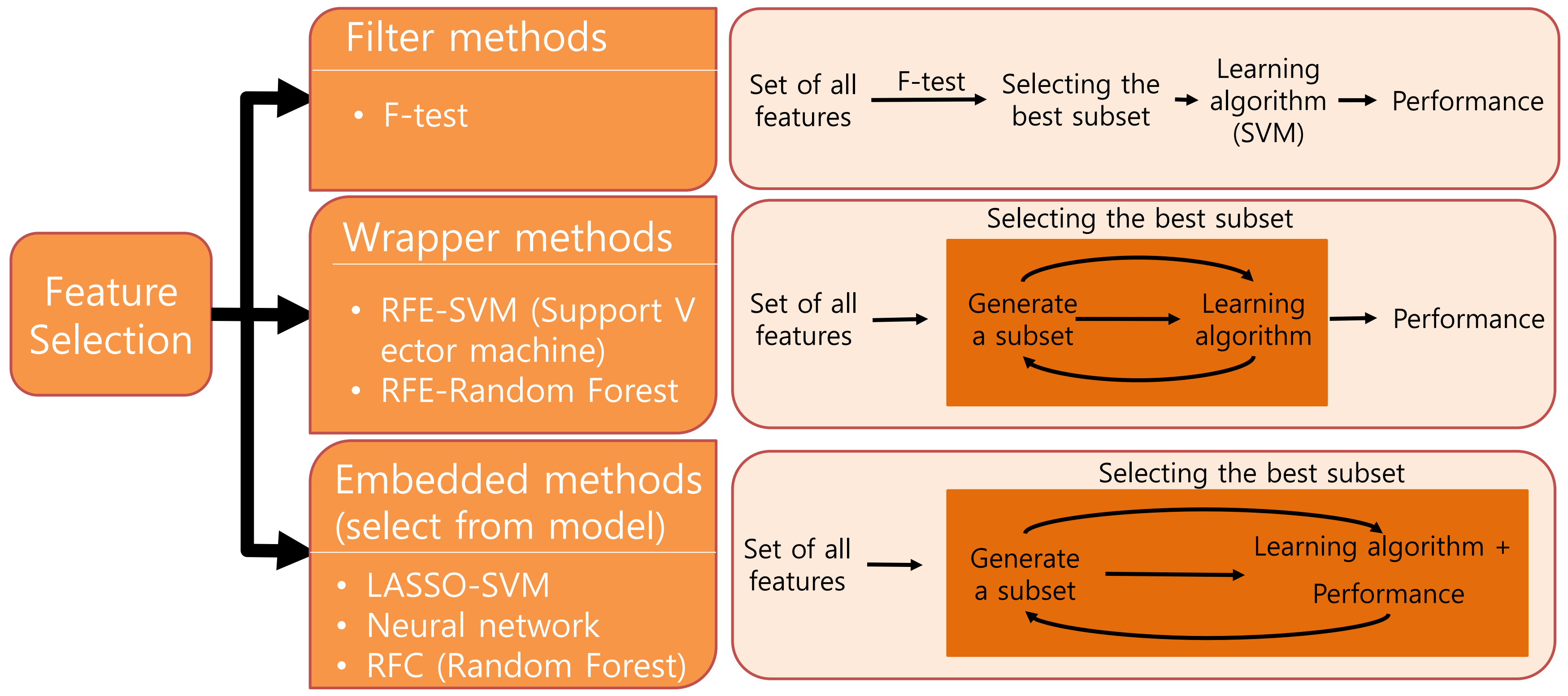

CRISPR Library
Amplified CRISPR Library
Lentivirus containing CRISPR gRNA library (or gRNA + Cas9)
Cas9-expressing cells (or wild-type cells)
Mutant Cells
Expansion of cells
Count gRNA by Deep sequencing

DATASET AND FEATURES

Two publicly available datasets were used for this work. Both set of data were obtained through the Dep Map project which have characterized the role of genes on cancer growth phenotype using genome-scale in vitro CRISPR screens. The first set of data characterized from 517 cancer cell lines was used for training and validation purposes. The second set of data characterized from 325 cancer cell lines was used for testing purposes. Total number of features is 16,183, and feature selection is performed among these features.

In both data sets, the raw data was in the form of a matrix containing positive or negative numbers indicating the impact (dependency) of a particular gene on cancer cell growth. If the positive number or negative number were represented in a particular gene (column) and a cell line (row), that indicated the role of the gene was respectively to inhibit (positive number) or promote (negative number) growth of the cancer cells.

APPLIED METHODS FOR GENE SELECTION



Feature Selection

Filter methods
- F-test

Set of all features → F-test → Selecting the best subset → Learning algorithm (SVM) → Performance

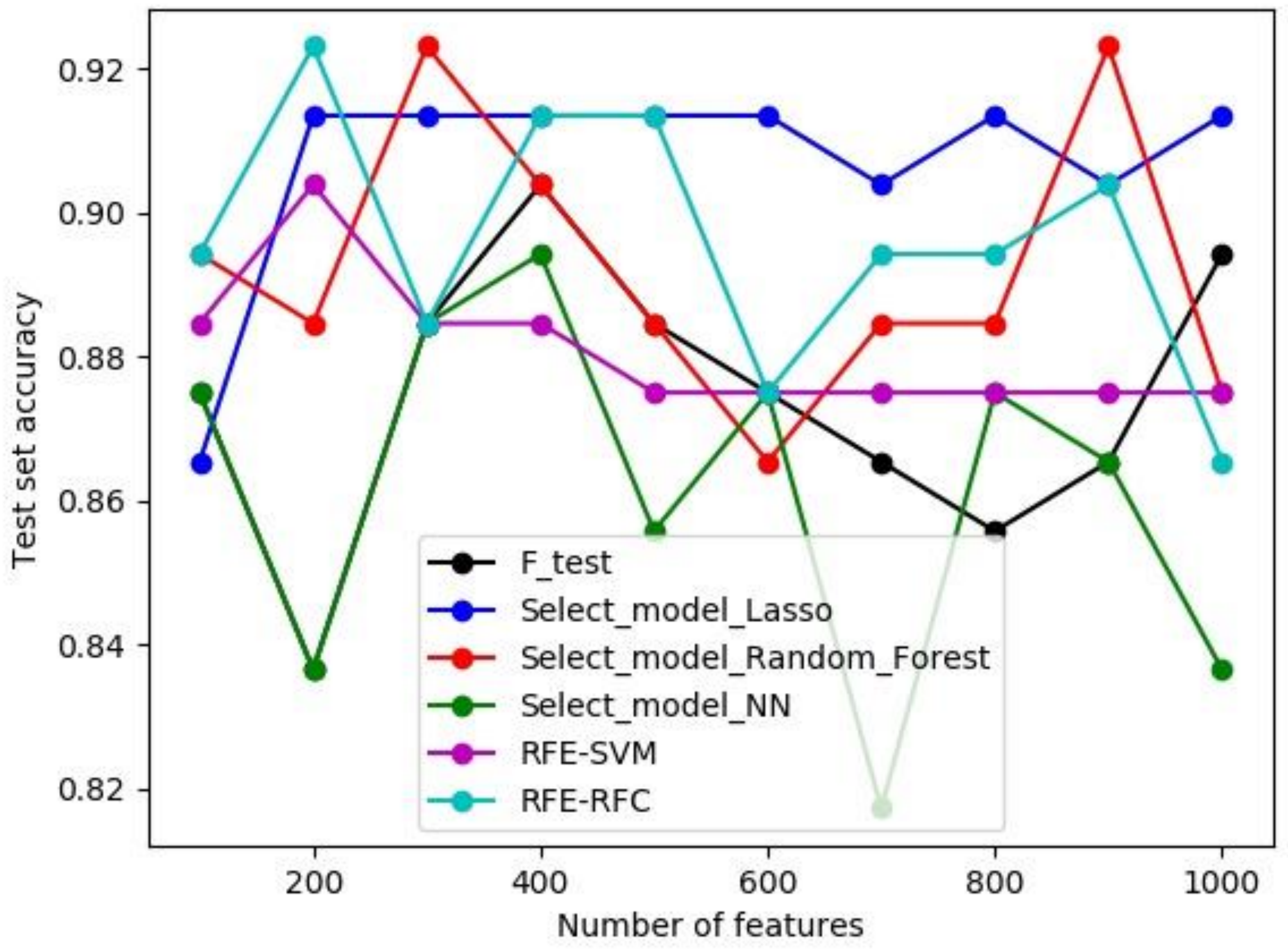
Wrapper methods
- RFE-SVM (Support Vector machine)
- RFE-Random Forest

Selecting the best subset
Set of all features → Generate a subset → Learning algorithm → Performance

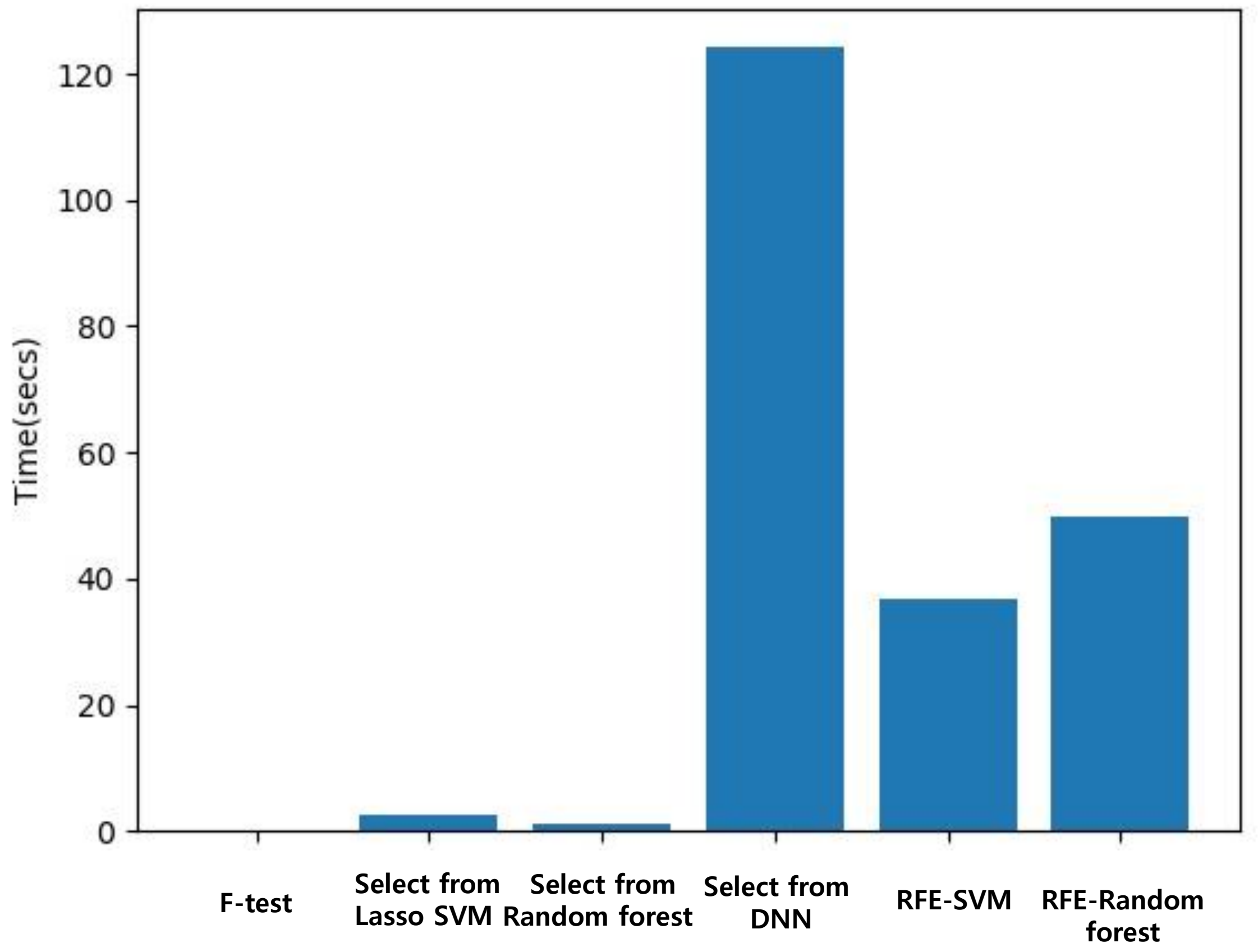
Embedded methods (select from model)
- LASSO-SVM
- Neural network
- RFC (Random Forest)

Selecting the best subset
Set of all features → Generate a subset → Learning algorithm + Performance

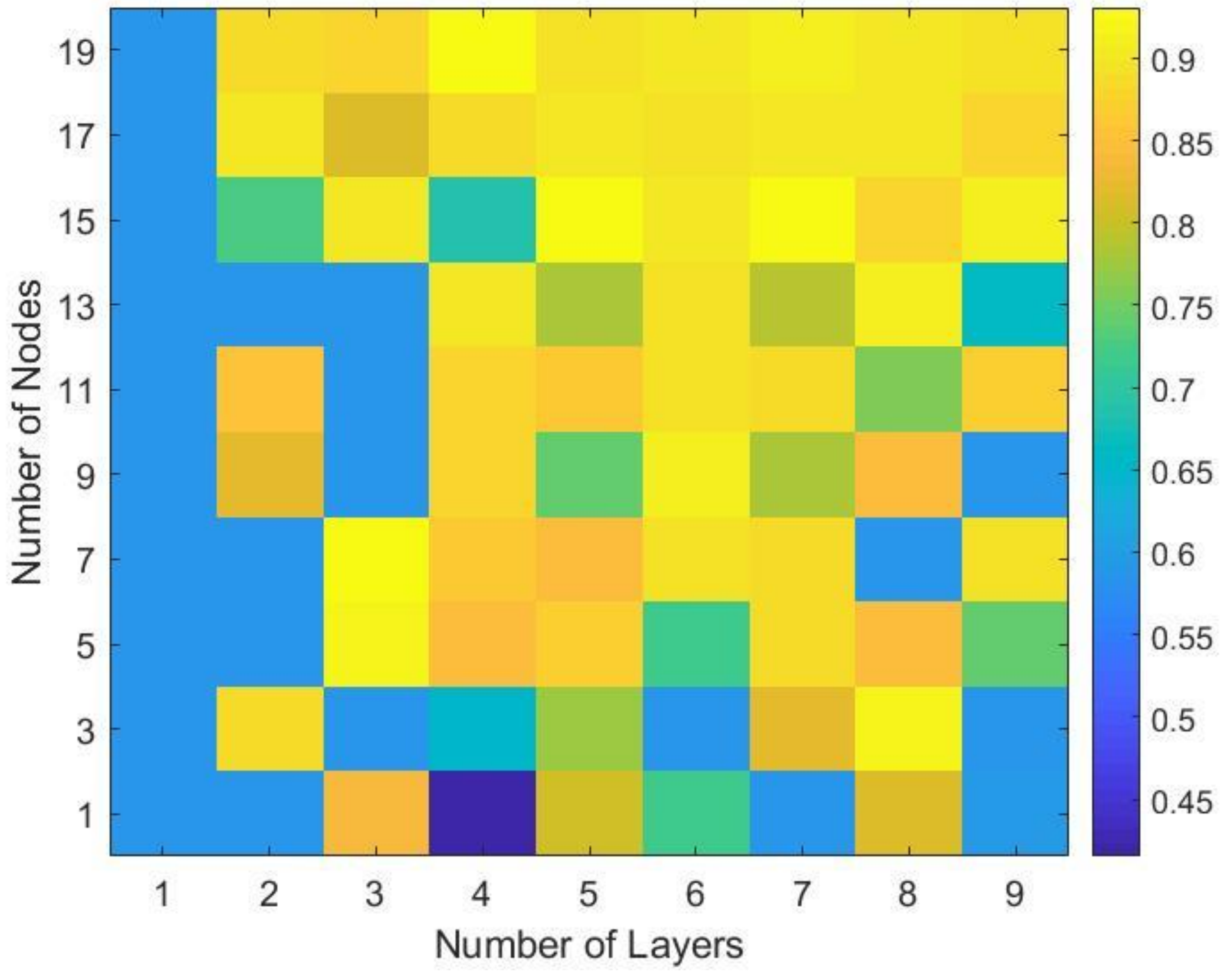
COMPARING PREDICTION ACCURACY WITH RESPECT TO METHODS AND THE NUMBER OF FEATURES



COMPARING COMPUTATIONAL TIME WITH RESPECT TO METHODS AND THE NUMBER OF FEATURES



PREDICTION ACCURACY OF DEEP LEARNING BASED FEATURE SELECTION WITH RESPECT TO THE NUMBER OF LAYERS AND NODES



CONCLUSIONS AND FUTURE WORK

Conclusions

- Among diverse methods, RFE-Random forest and select from Lasso SVM show the best average performance.

Methods	F test	Select from Lasso SVM	Select from Random forest	Select from DNN	RFE-SVM	RFE-Random forest
Average Accuracy	0.874	0.907	0.892	0.862	0.881	0.896

- The optimal number of feature for cancer type classification is 200, and its prediction accuracy is 0.923.
- The performance of deep learning-based feature selection is highly sensitive to the number of nodes and layers.
- Hyper-parameter optimization is necessary for efficient deep learning-based feature selection.

Future work

- Explore more specified types of cancers (4 types of cancers such as blood cancer, carcinoma, sarcoma, and others)
- Examine the biological functions represented from the selected genes to find the new target for specific cancer treatment

REFERENCES

1. Meyers, Robin M., et al. "Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells." Nature genetics 49.12 (2017): 1779.
2. Behan, Fiona M., et al. "Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens." Nature 568.7753 (2019): 511.

Contact:
HYUNG JUN YANG - hjyang3@stanford.edu, Hongpyo Lee – hongpyo@stanford.edu