Scaling tree-based automated machine learning to biomedical big data with a dataset selector

This manuscript (permalink) was automatically generated from trang1618/tpot-ds-ms@7dac4e6 on November 13, 2018.

Authors

- **Trang T. Le**[☯]
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104
- **Weixuan Fu**[☯]
  0000-0002-6434-5468 · weixuanfu · weixuanfu
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104
- **Jason H. Moore**[†]
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104 · Funded by National Institute of Health Grant Nos. LM010098, LM012601

[☯] — These authors contributed equally to this work.

[†] — Direct correspondence to jhmoore@upenn.edu.

Abstract

[Background, gap in field] We introduce two new features implemented in TPOT that helps increase the system's scalability: dataset selector (DS) and Template. Dataset selector provides the option to specify subsets of the features, reducing the computational expense of TPOT at the beginning of each pipeline to only evaluate on a smaller subset of data rather than the entire dataset. Consequently, DS makes TPOT applicable on large data sets by slicing the data into smaller sets of features and allowing genetic algorithm to select the best subset in the final pipeline. Template enforces type constraints with strongly typed genetic programming. We show that DS and Template help reduce TPOT computation time and potentially provide more interpretable results. Independent of a previous study that identified significant association with depressions severity of the enrichment scores of two modules, we find with TPOT-DS that one of the modules is largely predictive of the clinical diagnosis of each individual.

Introduction

For many bioinformatics problems of classifying individuals into clinical categories from high-dimensional biological data, choosing a classifier is merely one step of the arduous process that leads to predictions. To detect patterns among features (*e.g.*, clinical variables) and their associations with the outcome (*e.g.*, clinical diagnosis), a data scientist typically has to design and test different complex machine learning frameworks that consist of data exploration, feature engineering, model selection and prediction. Automated machine learning (AutoML) systems were developed to automate this challenging and time-consuming process. These intelligent systems increase the accessibility and scalability of various machine learning applications by efficiently solving an optimization problem to discover pipelines that yield satisfactory outcomes, such as prediction accuracy. Consequently, AutoML allows data scientists to focus their effort in applying their expertise in other important research components such as developing meaningful hypotheses or communicating the results.

[other AutoML systems]

Tree-based Pipeline Optimization Tool (TPOT) is a genetic programming-based AutoML system that automates the laborious process of designing a machine learning pipeline to solve a supervised learning problem. At its core, TPOT uses genetic programming (GP) [1] to optimize a series of feature preprocessors and machine learning models with the objective of maximizing classification accuracy. While most AutoML systems primarily focus on model selection and hyperparameter optimization, TPOT also pays attention to feature selection and feature engineering in building a complete pipeline. Applying GP with the NSGA-II Pareto optimization [2], TPOT optimizes the accuracy achieved by the pipeline while accounting for its complexity.

Specifically, to automatically generate and optimize these machine learning pipelines, TPOT utilizes the Python package DEAP [3] to implement the GP algorithm.

Given no a priori knowledge about the problem, TPOT has been showed to frequently outperform standard machine learning analyses [4,5]. Effort has been made to specialize TPOT for human genetics research, which results in a useful extended version of TPOT, TPOT-MDR, that features Multifactor Dimensionality Reduction and an Expert Knowledge Filter [6]. However, at the current stage, TPOT still requires great computational expense to analyze large datasets such as in genome-wide association studies or gene expression analyses. Consequently, application of TPOT on real-world datasets has been limited to small sets of features [7].

In this work, we introduce two new features implemented in TPOT that helps increase the system's scalability. First, the Dataset Selector (DS) allows the users to pass specific subsets of the features, reducing the computational expense of TPOT at the beginning of each pipeline to only evaluate on a smaller subset of data rather than the entire dataset. Consequently, DS makes TPOT applicable on large data sets by slicing the data into smaller sets of features (*e.g.* genes) and allowing genetic algorithm to select the best subset in the final pipeline. Second, Template enables the option for strongly typed GP, a method to enforce type constraints in genetic programming. By letting users specify a desired structure of the resulting machine learning pipeline, Template helps reduce TPOT computation time and potentially provide more interpretable results.

Methods

We begin with description of the two novel additiona to TPOT, Dataset Selector and Template. Then, we provide detail of a real-world RNA-Seq expression dataset and describe a simulation approach to generate data comparable with the expression data. Finally, we discuss other methods and performance metrics for comparison. The `R` and `Python` scripts for simulation and analysis are publicly available on the GitHub repository https://github.com/lelaboratoire/tpot-ds.

Dataset Selector

TPOT's current operators include sets of feature pre-processors, feature transformers, feature selection techniques, and supervised classifiers and regressions. In this study, we introduce a new operator called Dataset Selector (DS) that enables biologically guided group-level feature selection. Specifically, taking place at the very first stage of the pipeline, DS passes only a specific subset of the features onwards. Hence, with DS, users can specify subsets of features of interest to reduce the feature space's dimension at pipeline initialization. From predefined subsets of features, the DS operator allows TPOT to select the best subset that maximize average accuracy in k-fold cross validation (5-fold by default).

For example, in a gene expression analysis of major depressive disorder, a neuroscientist can specify collections of genes in pathways of interest and identify the important collection that helps predict the depression severity. Similarly, in a genome-wide association study of breast cancer, an analyst may assign variants in the data to different subsets of potentially related variants and detect the subset associated with the breast cancer diagnosis. In general, the DS operator allows for compartmentalization the feature space to smaller subsets based on *a priori* expert knowledge about the biomedical dataset. From here, TPOT selects the most relevant group of features, which can be utilized to motivate further analysis on that small group of features in biomedical research.

Template

Parallel with the establishment of the Dataset Selector operator, we now offer TPOT users the option to define a Template that provides a way to specify a desired structure for the resulting machine learning pipeline, which will reduce TPOT computation time and potentially provide more interpretable results.

Current implementation of Template supports linear pipelines, or path graphs, which are trees with two nodes (operators) of vertex degree 1, and the other $n - 2$ nodes of vertex degree 2. Further, Template takes advantage of the strongly typed genetic programming framework that enforces data-type constraints [8] and imposes type-based restrictions on which element (*i.e.*, operator) type can be chosen at each node. In strongly typed genetic programming, while the fitness function and parameters remain the same, the initialization procedure and genetic operators (*e.g.*, mutation, crossover) must respect the enhanced legality constraints [8]. With a Template defined, each node in the tree pipeline is assigned one of the five major operator types: dataset selector, feature selection, feature transform, classifier or regressor. Moreover, besides the major operator types, each node can also be assigned more specifically as a method of an operator, such as decision trees for classifier. An example Template is Dataset selector → Feature transform → Decision trees.

Datasets

We apply TPOT with the new DS operator on both simulated datasets and a real world RNA-Seq gene expression dataset. With both real-world and simulated data, we hope to acquire a comprehensive view of the strengths and limitations of TPOT in the next generation sequencing domain.

Real-world RNA-Seq expression data

We employed TPOT-DS on an RNA-Seq expression dataset of 78 individuals with major depressive disorder (MDD) and 79 healthy controls (HC) from Ref. [9]. Gene expression levels were quantified from reads of 19,968 annotated protein-coding genes and underwent a series of preprocessing steps including low read-count and outlier removal, technical and batch effect

adjustment, and coefficient of variation filtering. Consequently, whole blood RNA-Seq measurements of 5,912 genes were obtained and are now used in the current study to test for association with MDD status. We use the 23 subsets of interconnected genes called depression gene modules (DGMs) identified from the RNA-Seq gene network module analysis [9] as input for the DS operator.

Simulation methods

The simulated datasets were generated using the `R` package `privateEC`, which was designed to simulate realistic effects to be expected in gene expression or resting-state fMRI data. In the current study, to be consistent with the real expression dataset, we simulate interaction effect data with *m* = 200 individuals (100 cases and 100 controls) and *p*= 5,000 real-valued features with 4% functional (true positive association with outcome) for each training and testing set. Full details of the simulation approach can be found in Refs. [10,9]. Briefly, the privateEC simulation induces a differential co-expression network random normal expression levels and permute the values of targeted features within the cases to generate interactions. Further, by imposing a large number of background features (no association with outcome), we seek to assess TPOT-DS's performance in accommodating large numbers of non-predictive features.

To closely resemble the module size distribution in the RNA-Seq data, we first fit a Γ distribution to the observed module sizes then sample from this distribution values for the simulated subset size, before the total number of features reaches 4,800 (number of background features). Then, the background features were randomly placed in each subset corresponding to its size. Also, for each subset $S_i, i = 1, \ldots, n$, a functional feature s_j belongs to the subset with the probability

$$P(s_j \in S_i) \sim 1.618^{-i} \tag{1}$$

where 1.618 is an approximation of the golden ratio and yields a reasonable distribution of the functional features: they are more likely to be included in the earlier subsets (subset 1 and 2) than the later ones.

Performance assessment

For each simulated and real-world dataset, after randomly splitting the entire data in two balanced smaller sets (75% training and 25% holdout), we trained TPOT-DS with the Template `Dataset Selector-Transformer-Classifier` on training data to predict class (e.g., diagnostic phenotype in real-world data) in the holdout set. We assess the performance of TPOT-DS by quantifying its ability to correctly select the most important subset (containing most functional features) in 100 replicates of TPOT runs on simulated data with known underlying truth. We also compare the out-of-sample accuracy of TPOT-DS's exported pipeline on the holdout set with that of standard TPOT (with `Transformer-Classifier` Template, no DS operator) and XGBoost [11], a fast and an efficient implementation of the gradient tree boosting method that has shown much

utility in many winning Kaggle solutions [12] and been successfully incorporated in several neural network architectures [13,14]. In the family of gradient boosted decision trees, XGBoost accounts for complex non-linear interaction structure among features and leverages gradient descents and boosting (sequential ensemble of weak classifiers) to effectively produce a strong prediction model. To obtain the optimal performance for this baseline model, we tune XGBoost hyperparameters using the `R` package `caret` [15] version 6.0-80 with the repeated cross-validation algorithm and random search method.

Results

Our main goal is to test the performance of methods to identify features that discriminate between groups and optimize the classification accuracy.

Simulated data

We compare the accuracy of each method for r = 100 replicate simulated data sets with moderate interaction effect. These values of the effect size in the simulations generate adequately challenging data sets so that the methods' accuracies stay moderate and do not cluster around 0.5 or 1. Each replicate data set is split into training and holdout. The TPOT-DS, standard TPOT and XGBoost models are built from the training dataset, then the trained model is applied to the independent holdout data to obtain the generalization accuracy.

Our simulation design produces a reasonable distribution of the functional features in all subsets, of which proportions are shown in Table [S1]. According to Eq. 1, the earlier the subset, the more functional features it has. Therefore, our first aim is to determine how well TPOT-DS can identify the first subset 1 that contains the largest number of informative features. With the specified template `Dataset Selector-Transformer-Classifier`, in 100 replications, TPOT-DS correctly selects subset 1 in the resulting pipeline 75 times (Fig. 1), with an average cross-validated accuracy on the training set of 0.73 and out-of-sample accuracy of 0.69.

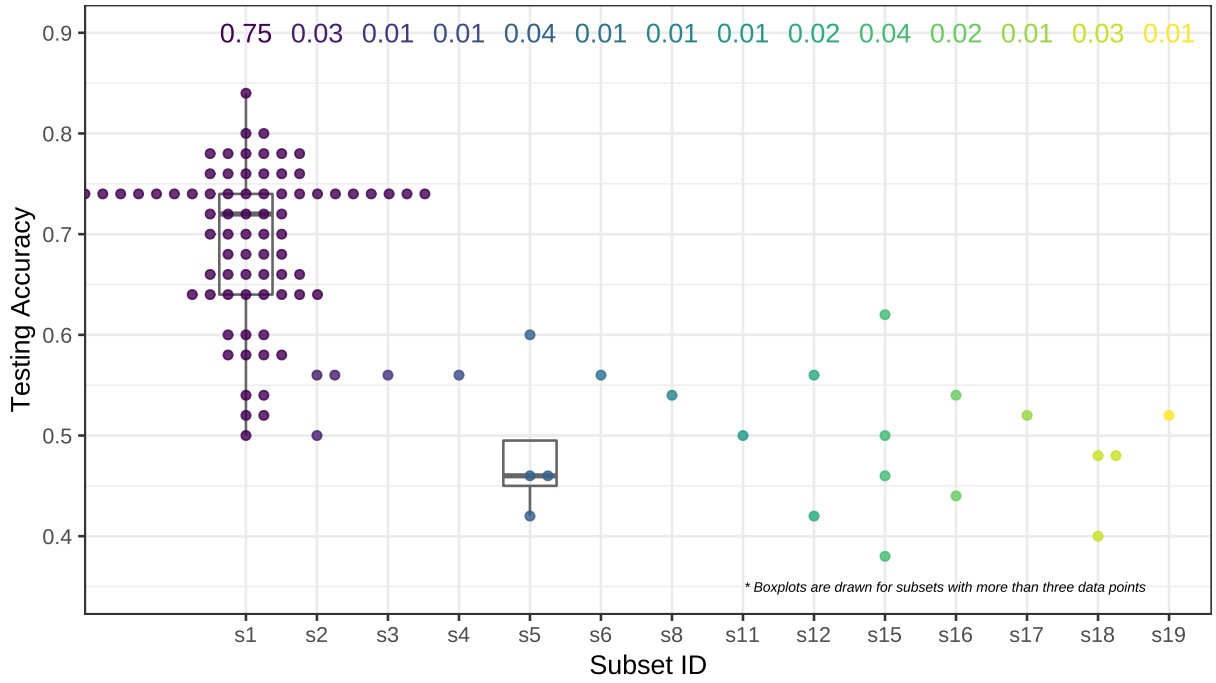

Figure 1: TPOT-DS's out-of-sample accuracy in simulated data with selected subset

Without DS, the standard TPOT and tuned XGBoost models respectively report a cross-validated accuracy of [0.661] and 0.533, and out-of-sample accuracy of [0.565] and 0.575.

RNA-Seq expression data

We apply standard TPOT, TPOT-DS and XGBoost to the RNA-Seq study of 78 major depressive disorder (MDD) subjects and 79 healthy controls (HC) described in [9]. The dataset contains 5,912 genes after preprocessing and filtering (see Methods for more detail). We excluded 277 genes that did not belong to 23 subsets of interconnected genes (DGMs) so that the dataset remains the same across the three methods. As with simulated data, all models are built from the training dataset (61 HC, 56 MDD), then the trained model is applied to the independent holdout data (18 HC, 22 MDD) to obtain the generalization accuracy.

In 100 replications, TPOT-DS selects DGM-5 (291 g enes) 64 times to be the subset most predictive of the diagnosis status (Fig. 2), with an average cross-validated accuracy on the training set of 0.715 and out-of-sample accuracy of 0.636. In the previous study with a modular network approach, we showed that DGM-5 has statistically significant associations with depression severity measured by the Montgomery-Åsberg Depression Scale (MADRS). Further, with 82% overlap of DGM-5's genes in a separate dataset from the RNA-Seq study by Mostafavi et al. [16], this gene collection's enrichment score was also shown to be significantly associated with the diagnosis status in this independent dataset.

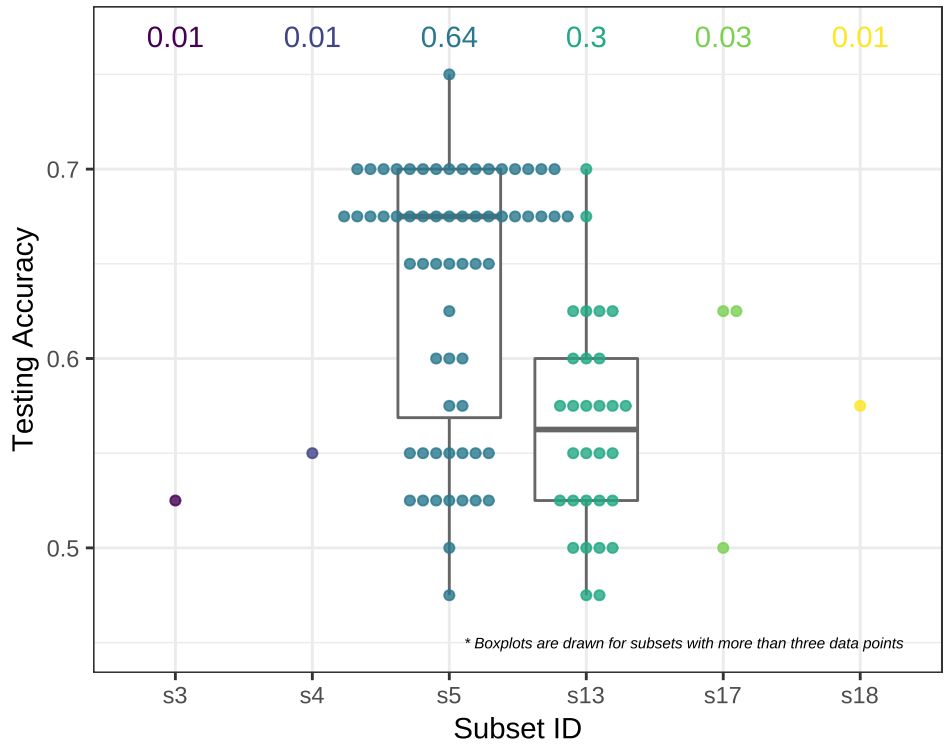

Figure 2: TPOT-DS's out-of-sample accuracy in RNA-Seq expression data with selected subset

After DGM-5, DGM-13 (134 genes) was selected by TPOT-DS 30 times (Fig. 2), with an average cross-validated accuracy on the training set of 0.717 and out-of-sample accuracy of 0.563. Previously, this module's enrichment score did not show statistically significant association with the MADRS.

Without DS, the standard TPOT and tuned XGBoost models respectively report a cross-validated accuracy of [] and 0.543, and out-of-sample accuracy of [] and 0.525.

Discussion

To our knowledge, TPOT-DS is the first AutoML tool to offer the option of feature selection at the group level. Previously, it was computationally expensive for any AutoML program to process biomedical big data. TPOT-DS is able to identify the most meaningful group of features to include in the prediction pipeline. We assessed TPOT-DS's out-of-sample prediction accuracy compared to standard TPOT and XGBoost, another state-of-the-art machine learning method. We applied TPOT-DS to real-world expression data to demonstrate the identification of biologically relevant groups of genes.

Implemented with a strongly typed GP, Template allows users to pre-specify a particular pipeline structure, which speeds up the automation computation time and provides potentially more interpretable results. Hence, Template enables the comparison between the two TPOT implementations, with and without DS.

We simulated data of the similar scale and chalenging enough for the models to have similar predictive power as in the real-world RNA-Seq data. TPOT-DS correctly selects the first subset (containing the most important features) 75% of the time with high holdout accuracy (0.69). When another subset is chosen in the final pipeline, this method still produces holdout accuracy comparable to that of standard TPOT and XGBoost (0.565 - 0.575).

Interestingly enough, TPOT-DS repeatedly selects DGM-5 to include in the final pipeline. In a previous study, we showed DGM-5 and DGM-17 enrichment scores were significantly associated with depression severity [9]. We also remarked that DGM-5 contains many genes that are biologically relevant or previously associated with mood disorders [9] and its enriched pathways such as apoptosis indicates a genetic signature of MDD pertaining shrinkage of brain region-specific volume due to cell loss [17,18].

TPOT-DS also select DGM-13 as a potentially predictive group of features with smaller out-of-sample accuracy compared to DGM-5 ($0.563 < 0.636$). []

It is important to discuss the complexity - interpretability trade-off in the context of AutoML. While arbitrarily-shaped pipelines may yield predictions competitive to human-level performance, these pipelines are often too complex to be interpretable. Vice versa, a simpler pipeline with defined steps of operators may be easier to interpret but not yield the optimal accuracy. Finding the optimal pipeline complexity that yields reasonable model interpretation and generalization remains a challenging task for AutoML application in biomedical big data.

Another limitation of this analysis is that subsets have to be predefined prior to executing TPOT-DS. While this option is desirable when *a prior* knowledge on the biological data is available, it might pose as a challenge when this is not the case, such as when analyzing data of a brand-new disease. Nevertheless, one can perform a clustering method such as *k*-means to group features prior to performing TPOT-DS on the data.

Extensions of TPOT-DS will involve overlapping subsets, which will require pipeline complexity reformulation beyond the total number of operators included in a pipeline. Also, a future design to support tree structures for Template will enable TPOT-DS to identify more than one subset that have high predictive power of the outcome.

References

1. **Genetic Programming: An Introduction: on the Automatic Evolution of Computer Programs and Its Applications**
Wolfgang Banzhaf, Frank D. Francone, Robert E. Keller, Nordin Peter
ACM Digital Library (1998) https://dl.acm.org/citation.cfm?id=280485

2. **A fast and elitist multiobjective genetic algorithm: NSGA-II**
K. Deb, A. Pratap, S. Agarwal, T. Meyarivan
IEEE Transactions on Evolutionary Computation (2002-04) https://doi.org/bnw2vv
DOI: 10.1109/4235.996017

3. **DEAP: Evolutionary Algorithms Made Easy**
Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, Christian Gagné
Journal of Machine Learning Research (2012) http://www.jmlr.org/papers/v13/fortin12a.html

4. **Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science**
Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, Jason H. Moore
Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16 (2016) https://doi.org/gfgqv2
DOI: 10.1145/2908812.2908918

5. **Identifying and Harnessing the Building Blocks of Machine Learning Pipelines for Sensible Initialization of a Data Science Automation Tool**
Randal S. Olson, Jason H. Moore
arXiv (2016-07-29) https://arxiv.org/abs/1607.08878v1

6. **Toward the automated analysis of complex diseases in genome-wide association studies using genetic programming**
Andrew Sohn, Randal S. Olson, Jason H. Moore
Proceedings of the Genetic and Evolutionary Computation Conference on - GECCO '17 (2017) https://doi.org/gfgqv3
DOI: 10.1145/3071178.3071212

7. **Integrated machine learning pipeline for aberrant biomarker enrichment (i-mAB): characterizing clusters of differentiation within a compendium of systemic lupus erythematosus patients**
Trang T. Le, Nigel O. Blackwood, Jaclyn N. Taroni, Weixuan Fu, Matthew K. Breitenstein
arXiv (2018-03-08) https://arxiv.org/abs/1803.04487v1

8. **Strongly Typed Genetic Programming**
David J. Montana
Evolutionary Computation (1995-06) https://doi.org/ct3mnb
DOI: 10.1162/evco.1995.3.2.199

9. **Identification and replication of RNA-Seq gene network modules associated with depression severity**
Trang T. Le, Jonathan Savitz, Hideo Suzuki, Masaya Misaki, T. Kent Teague, Bill C. White, Julie H. Marino, Graham Wiley, Patrick M. Gaffney, Wayne C. Drevets, … Jerzy Bodurka
Translational Psychiatry (2018-09-05) https://doi.org/gd7jx7
DOI: 10.1038/s41398-018-0234-3 · PMID: 30185774 · PMCID: PMC6125582

10. **Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure**
Caleb A Lareau, Bill C White, Ann L Oberg, Brett A McKinney
BioData Mining (2015-02-03) https://doi.org/gb5fpr
DOI: 10.1186/s13040-015-0040-x · PMID: 25685197 · PMCID: PMC4326454

11. **XGBoost**
Tianqi Chen, Carlos Guestrin
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16 (2016) https://doi.org/gdp84q
DOI: 10.1145/2939672.2939785

12. **Kaggle: Your Home for Data Science**https://www.kaggle.com/

13. **Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation**
Huiting Zheng, Jiabin Yuan, Long Chen
Energies (2017-08-08) https://doi.org/gbtqbr
DOI: 10.3390/en10081168

14. **A Novel Image Classification Method with CNN-XGBoost Model**
Xudie Ren, Haonan Guo, Shenghong Li, Shilin Wang, Jianhua Li
Digital Forensics and Watermarking (2017) https://doi.org/gfgvf3
DOI: 10.1007/978-3-319-64185-0_28

15. **Building Predictive Models inRUsing thecaretPackage**
Max Kuhn
Journal of Statistical Software (2008) https://doi.org/gdgzwf
DOI: 10.18637/jss.v028.i05

16. **Type I interferon signaling genes in recurrent major depression: increased expression detected by whole-blood RNA sequencing**
S Mostafavi, A Battle, X Zhu, JB Potash, MM Weissman, J Shi, K Beckman, C Haudenschild, C McCormick, R Mei, … DF Levinson
Molecular Psychiatry (2013-12-03) https://doi.org/f6qdpr
DOI: 10.1038/mp.2013.161 · PMID: 24296977 · PMCID: PMC5404932

17. **A meta-analysis examining clinical predictors of hippocampal volume in patients with major depressive disorder.**
Margaret C McKinnon, Kaan Yucel, Anthony Nazarov, Glenda M MacQueen
Journal of psychiatry & neuroscience : JPN (2009-01) https://www.ncbi.nlm.nih.gov/pubmed/19125212
PMID: 19125212 · PMCID: PMC2612082

18. **Increased apoptosis in patients with major depression: A preliminary study.**
E Eilat, S Mendlovic, A Doron, V Zakuth, Z Spirer
Journal of immunology (Baltimore, Md. : 1950) (1999-07-01) https://www.ncbi.nlm.nih.gov/pubmed/10384158
PMID: 10384158