Comprehensive Study of Semi-Supervised Learning for DNA Methylation-based Supervised Classification of Central Nervous System Tumors

Quynh T. Tran

St. Jude Children's Research Hospital

Md Zahangir Alom

St. Jude Children's Research Hospital

Brent A. Orr (✉ brent.orr@stjude.org)

St. Jude Children's Research Hospital

Research Article

Keywords: Semi-supervised learning, neural network, artificial intelligence, supervised classifiers, DNA-methylation, central nervous system tumor, machine learning, random forest

Posted Date: February 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1284799/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Optimal treatment for cancer patients relies on accurate pathological diagnosis. The world health organization has defined almost 100 known central nervous system (CNS) tumors, making the histopathological diagnostic process particularly challenging. To help reduce variability, validation costs, and standardize the histopathological diagnostic process, supervised machine learning techniques using epigenetic data have been developed. These methods require large labeled training data sets to obtain clinically acceptable classification accuracy. However, labeling pathology data for machine learning models is time-consuming and resource-intensive, especially for rare tumor types. On the other hand, there is now abundance of unlabeled epigenetic data across multiple databases. To maximize the utility of both labeled and unlabeled data, semi-supervised learning (SSL) approach has been proposed and shown to be effective in genomic field. However, it has not yet been explored with epigenetic data nor demonstrated beneficial to CNS tumor classification.

Results: This paper explores the application of semi-supervised machine learning on methylation data to improve accuracy of supervised learning models in classifying CNS tumors. We comprehensively evaluated 11 SSL methods and developed a novel combination approach including a self-training with editing using support vector machine (SETRED-SVM) as the base learner model with an L2-penalized, multinomial logistic regression model to obtain high confidence labels from a small labeled instances. Results across 8 differently trained random forest and neural net models show that SETRED-SVM followed with multinomial logistic regression approach is able to produce high confidence training data set, leading to a statistically significant increase in prediction accuracy of 82 CNS tumors.

Conclusions: The proposed combination of semi-supervised technique and multinomial logistic regression holds the potential to leverage the abundant publically available unlabeled methylation data effectively. Such an approach is highly beneficial in providing additional training examples, especially for scarce tumor types, to boost the prediction accuracy of supervised models.

Background

Artificial intelligent (AI) technologies have been widely adopted in the diagnostic process of various biomedical disciplines [1–4]. Furthermore, with the advent of high-throughput technologies such as microarrays and nucleic acid sequencers, the use of machine learning and deep learning has also become increasingly indispensable in the field of cancer genomics [5–7]. The introduction of these advanced computational methods has provided many opportunities to improve health care and increase the precision of oncologic diagnosis.

A key challenge in medical science is the precise classification of diseases and the development of optimal therapies. This is particularly more challenging in classifying brain tumors due to the developmental complexity of the brain. The World Health Organization has defined 82 central nervous system (CNS) tumor classes, encompassing a broad spectrum from benign neoplasms, which can be

treated by surgery alone, to malignant tumors that respond poorly even with aggression adjuvant therapy. With the advancement in AI and the abundance of genomic and epigenomic data, methylation-based classification of human tumors has emerged as an essential diagnostic tool in the clinical laboratory. Supervised models have been implemented to assist in diagnosing CNS tumors and sarcomas [8, 9].

These initially deployed models are clinically useful but have inherent limitations. Constructing optimal supervised models for methylation-based classification in the clinical environment is dependent on having a comprehensive set of labeled “gold standard” data for training and validation. The current reference sets are not entirely complete, yielding a significant proportion of unclassifiable tumors [8]. Furthermore, the reference cohorts suffer from a considerable class imbalance due to the lack of sufficient examples of rare tumor types to train supervised classification models; thus, degrading model performance.

To fully leverage methylation profiling and machine learning for tumor classification, models should be improved over time by augmenting the training cohorts with additional labeled reference examples of rare tumors and relabeling samples after additional molecular substructures have been identified within known tumor types. In addition, with the vast publically available methylation profiling data, model updates would benefit from combining well-characterized data with relevant tumor profiles acquired from large public repositories.

Obtaining additional labeled training data for improving CNS tumor classifiers can be challenging. Current “gold standard” approaches to sample labeling for methylation cohorts include a histomorphologic assessment by expert pathologists, orthogonal molecular testing, and unsupervised methods such as dimensionality reduction or cluster analysis. However, establishing a ground truth methylation class is difficult for a subset of tumors because they lack defining gene abnormalities or copy number changes. Additionally, closely related molecular subgroups within tumor types can be challenging to distinguish unbiasedly. Therefore, the cost in time, effort, and additional testing make this degree of rigor in labeling infeasible, particularly when applied to large cohorts.

Semi-supervised learning (SSL), an approach to machine learning, is often used when labeling data is not feasible or requires substantial resources [10, 11]. SSL has been applied to a diverse set of problems in biomedical science and has shown to be effective in predicting unlabeled data using a very small number of labeled instances. Example applications in genomic medicine include gene finding, miRNA discovery, predicting gene regulatory networks, and survival modeling [12–16]. Although SSL has shown to be effective in the genomic field, it has not been explored with epigenetic data from human tumors. Rare applications to brain tumor classification have been limited to deriving tumor classes from radiologic images [17].

In this work, we present the first application of using semi-supervised methods to assign 82 CNS tumor subclasses using DNA-methylation data and demonstrate the benefits of these additional labels to the performance of AI models. We comprehensively evaluated 11 different SSL techniques to label a public reference dataset of 2801 samples. To have comparable scores among classes and better estimates of

the confidence of each prediction, we implemented an L2-penalized, multinomial logistic regression model to calibrate the raw SSL labels' scores. This approach was also adapted to calibrate the RF classification scores [8]. After calibration, the labels obtained from the best SSL model (self-train with editing method using support vector machine as the base learner – SETRED-SVM) were used to train the random forest (RF) and artificial neural net (NN) baseline models. The other six RF and NN models were trained with additional SSL labels from the independent GSE109379 data set. This study showed that CNS tumor classification accuracy of supervised models was statistically increased when training data were augmented with additional SSL labeled samples. While this study indicates that SSL is highly effective for labeling CNS tumors using DNA-methylation profiles, SSL is likely to be more broadly applicable in other types of cancer and biomedical data.

Results

Cohort characteristics

As an initial evaluation of SSL on methylation data, we used a cohort of previously published training data from a comprehensive brain tumor classification model (GSE90496 [8]). We applied five different training functions with four different base learners (Table S1). This cohort has several characteristics that make it ideal for this task. The cohort includes 2801 labeled tumors comprising 75 methylation families (MCF) from tumor and standard brain samples, representing 82 distinct CNS tumor methylation subclasses (MC) and nine control groups. The subclasses represent closely related molecular groups within individual tumor types, and a comparison of the family and subclass labeling performance is a good surrogate for tumor relabeling. An additional 1104 unlabeled samples (GSE109379 [8]) were also available.

Semi-supervised models for evaluation

We evaluated 11 semi-supervised learning models (Table S1) using five different training techniques, i.e., self-training (SELFT [18]), self-training with editing (SETRED [19]), self-training nearest-neighbor rule using cut edges (SNNRCE [20]), tri-training (TRITRAIN [11]), and democratic co-learning (DEMO [21]). Each training function used either of these supervised classifiers as the base learner: one nearest neighbor (oneNN), decision tree C5.0, and support vector machine (SVM). For SNNRCE, oneNN was the fixed built-in learner. For DEMO, all three base learners were used. The learner parameters for SVM were C-classification and radial kernel. Euclidian distances were used for SSL with oneNN base learner.

SSL models performance on tumor family classification

The train and test data for SSL models are depicted in Figure 1. The performance of each model was summarized and evaluated based on average accuracy, specificity, precision, and recall after predicting 75 MCFs for 842 (30% of 2801 samples) inductive and 986 transductive samples.

All models performed well on transductive and inductive data sets at the family classification level with an accuracy above 0.84 (Figure 2A). Training functions with the C5.0 decision tree as the base learner

had the lowest average accuracy (< 0.87) (Figure 2A) as well as precision (< 0.77) and recall (< 0.73) (Figure 2B). Those with 1NN learners had consistent accuracies above 0.93 (Figure 2A), while their precision and recall were around 0.92 and 0.87, respectively (Figure 2B). Among the 11 models, SETRED with SVM base learner performed the best with mean accuracy above 0.95 (Figure 2A), precision greater than 0.96, and recall above 0.91 (Figure 2B). In addition, all models had very high specificity (above 0.99), suggesting that SSL could identify true negative at a high rate (Figure 2C).

SSL models performance on tumor subclass classification

The performance of the 11 SSL models was also assessed based on their prediction of 91 methylation subclasses (MC). Figure 3 shows the prediction of the inductive and transductive data sets for each training function combined with different supervised-based learners. Similar to the methylation family classification, training function with 1NN and SVM (accuracy $> 91\%$, precision $> 91\%$, and recall $> 83\%$) performed better than those with C5.0 decision tree (accuracy ranges from 83-86%, precision ranges from 70-76%, and recall ranges from 66-84%) (Figure 3A and 3B). Among the 11 SSL models, SETRED with SVM base learner performed the best with accuracy $\sim 97\%$, precision $> 96\%$, and recall $> 91\%$ for inductive and transductive testing (Figure 3A, B). All models had specificity at around 99% (Figure 3C).

Obtaining high confidence scores from the best SSL model

After evaluating 11 SSL models, SETRED-SVM yielded the highest accuracy, precision, and recall and was chosen as an application model for further analysis. This model assigned a predicted score to each of the 91 classes for each sample, resulting in an aggregated raw score (Figure 4A). To obtain class probability estimates that can measure the confidence in the class assignment, we fitted a multinomial logistic regression calibration model to the raw scores. This calibration process produced calibrated scores that allowed us to compare between classes despite differences in raw score distributions (Figure 4B).

Receiver operating characteristic (ROC) curve analysis of the maximum raw and calibrated scores demonstrated that calibration improved the area under the curve (AUC) from 0.73 to 0.855, respectively (Figure 5A). To obtain a cutoff score with high confidence for predicting a matching class, we performed threshold analysis that utilized maximization of the Youden Index (balanced between sensitivity and specificity) [22]. The analysis suggested a threshold of 0.8 for MC calibrated scores with the specificity of 78.3%, sensitivity of 79.5%, and precision of 98.8% (Figure 5B). Single family specificity and sensitivity for threshold 0.8 were 95.9% and 79.8%, respectively (Figure S1).

Effect of supervised classification model performance after augmenting training data with semi-supervised labeled data

To demonstrate the utility of our labeling method, we tested whether adding samples labeled using SSL would improve the performance of supervised random forest (RF) and neural net (NN) classification models. First, we built baseline models with 35% data from GSE90496. Then, we trained the baseline models with additional SSL pseudo-labels from the remaining 35% GSE90496 and the 1104 GSE109379

samples (Figure S2). Compared to the baseline model, all RF models showed a statistically significant increase in balanced accuracy (up to 7%), while the NN models showed a significant increase in balanced accuracy (0.7%) when only high confidence MCF labels were included (Figure 6A). Overall, the NN model yielded a much higher balanced accuracy (92.9% and 97.5%) compared to the RF classifier (70.9% and 72.3%) at predicting methylation subclass and family, respectively (Figure 6 and 7). Unlike the RF classifier, the performance of the NN on the hold-out test set appeared to be very robust (Figure 6A and 7A). The RF prediction accuracy on the hold-out test set closely resembled its cross-validation performance. In contrast, the accuracy of the NN appeared to be unchanged regardless of its cross-validation performance for both subclass and family levels (Figure 6A and 7A).

To further understand how the quality of semi-supervised labels affect the supervised classifiers, we have defined labels predicted by SETRED-SVM with calibrated scores above the thresholds as high confident (HC) while the remaining would be considered as low confident (LC) SS labels. The numbers and their corresponding proportions to the referent labeled samples were categorized into low frequency (< 10 samples per group, LF-green bars), high frequency (≥ 10 samples per group, HF-red bars), as shown in Figure 6B and 7B. We observed that adding more data to the original training data set showed a statistically significant improvement in the performance of the RF classifier (Figure 6A and 7A). NN performance showed statistically significant improvement in the presence of high confident SS labels only, while its accuracy significantly decreased in the addition of only low confident data. This suggests that NN was more sensitive to the data quality than the RF classifier. In conclusion, semi-supervised labels improved the prediction of both RF and NN classifiers. The NN classifier had a higher accuracy prediction; however, its performance was more sensitive to low confident labels than the RF classifier.

Discussion And Conclusions

This study demonstrates that semi-supervised (SS) learning is a valuable and efficient approach to label or relabel large cohorts of unclassified samples using their DNA-methylation profiles. Further, we show that the pseudo-labels from an SSL model can effectively be utilized during training to improve the performance of supervised classification models. This approach substantially reduces the subjectivity and effort from domain experts associated with class label generation.

Among the 11 SSL models (Table S1) evaluated, SETRED-SVM, a data editing method using support vector machine as the base learner, outperformed other models for labeling methylation family or subclass. SETRED utilizes an active-learning-like technique to identify and remove mislabeled examples from the self-labeled data and has been demonstrated to be robust to noise in self-labeled data [19]. While this is the first instance of SSL being applied to methylation array data, other investigators have also found SETRED to perform well in other biomedical SSL tasks such as biomedical image classification [23].

One concern when using SSL methods to label data is the potential to spuriously label out-of-distribution samples with one of the provided seed labels [24]. Implementation of SSL for clinical classification tasks likely will require concomitant sample distribution analysis. In our instance, we calibrated the SETRED-

SVM predictive scores using a multinomial logistic calibration model. The logistic model produced probability estimates that could be used to assess the confidence in class assignment. This approach was previously adapted to calibrate the RF classification scores [8]. We showed that this step is critical in creating a high-quality data set for subsequent supervised classifier training. Receiver operating characteristic (ROC) curve analysis of SETRED-SVM maximum raw and calibrated scores demonstrated that calibration improved the area under the curve (AUC) from 0.73 to 0.855, respectively (Figure 5A). Both the RF and NN classifiers showed significant degradation when low-confidence samples were utilized for training, even in the presence of a substantial increase in overall samples (Figure 6 and 7).

One problem with the chosen method of distribution analysis was it yielded relatively low scores for tumors with very low representation (< 10 samples) among the seed labels. For instance, tumor classes such as pituitary adenoma, prolactin (PITAD, PRL) and low-grade glioma, desmoplastic infantile astrocytoma/ganglioglioma (LGG, DIG/DIA), which had representation from only eight tumors per methylation class, each failed to produce high-confidence pseudo-labeled tumors (Figure S3 and S4). We suspect that polynomial regression models may show better performance in determining the class confidence from under-represented tumor classes because they are more suitable for high-dimensional data [25] and provide plausible results for small sample size data (< 8) [26].

In our supervised analysis, the RF and NN classifiers demonstrated variability in performance when using pseudo-label samples. While both models benefited from the additional SS labeled data, the NN had at least 20% higher accuracy in all models compared to the respective RF models. As our models utilized a significantly reduced set of samples compared to the entire training set described by Capper et al. [8], this likely reflects the ability of the NN classifier to learn with fewer total examples per class even at baseline. Notably, the RF models appeared less sensitive to noisy data in the training set, whereas the NN models showed degradation from baseline with the addition of only low-confidence pseudo labels (Figure 6 and 7).

Overall, our analysis shows that SSL can be used to label or relabel methylation array samples for clinical diagnostics. In the clinical environment, this approach may be helpful to (1) improve the performance of supervised classifiers by utilizing public methylation data, (2) provide ground truth labels for validation purposes when other traditional approaches appear to be more costly, or (3) modify labels after more granular classification schemes are introduced.

Methods

Processing methylation data

All semi-supervised models were trained on genome-wide DNA methylation data from the CNS tumor reference cohort (GSE90496) consisting of 2,801 samples from 75 methylation class families (MCF), which include 91 methylation subclasses (MC) [8]. All methylation data, including those from GSE90496 and GSE109379, were processed in R (<http://www.r-project.org>, version 4.0.2), using several packages

from Bioconductor and other repositories. Specifically, array data were preprocessed using the *minfi* package (v.1.36.0) [27]. Background correction with dye-bias normalization was performed for all samples using noob (normal-exponential out-of-band) with “single” dye method [28] with preprocessFunNorm. Probe filtering was performed after normalization. Specifically, probes located on sex chromosomes containing a nucleotide polymorphism (dbSNP132 Common) within five base pairs of and including the targeted CpG-site, or mapping to multiple sites on hg19 (allowing for one mismatch), and cross-reactive probes were removed from the analysis. After the filtering process, 438370 probes remained.

Training and validation of semi-supervised learning (SSL) models

The *ssc* R package (v2.1-0) was used to build and train SSL models. First, the standard deviation for each probe across all 2,801 samples from GSE90496 was calculated. Input features for SSL models were the 5072 probes with a standard deviation greater than 0.3. The train and test data for SSL models are depicted in Figure 1. Briefly, 30% of DNA methylation data from GSE90496 were kept aside as an inductive testing data set to independently evaluate the performance of each SSL model. The remaining 70% of the data was used as a training set. The training data were proportionally partitioned into a labeled and an unlabeled set. Specifically, 50% (of the 70%) training data were used as labeled examples, while the remaining 50% (of the 70%) were used as unlabeled examples or as a transductive testing set. The partitioning process was bootstrapped seven times. Each MC or MCF was proportionally sampled for every bootstrap to ensure that the class distributions were similar to the original class distributions.

Performance metrics

After predicting MC and MCF for both transductive and inductive testing sets, the one-vs-all multiclass performance for each model was evaluated based on the average accuracy, specificity, precision, and recall of all splits. A model with the highest balanced accuracy, specificity, precision, and recall in both testing data sets was selected as the final model for performing threshold analysis and prediction for GSE109379 samples. Balanced accuracy was used to compare performance across models as it is a better judge in the imbalanced class setting, in which some tumor classes are a lot rarer than others and have much smaller sample sizes (< 10 samples in a group). Balanced accuracy is insensitive to imbalanced class distribution, and it gives more weight to the instances coming from minority classes. On the other hand, accuracy treats all instances alike and usually favors the majority class [29]. First, recall for the algorithm on each class was computed, then the arithmetic mean of these values was calculated to find the final balanced accuracy score. It is “balanced” because each class is represented by its recall; thus, it has the same weight and the same importance, regardless of its size [30].

Semi-supervised classifier score calibration

The distribution of the scores generated from the best semi-supervised model (SETRED-SVM) varied between classes, making inter-class comparisons difficult. These scores appeared to spread widely from 0 to 1 when they were used to assign the correct class labels (Figures 4A), suggesting that they do not

reflect well-calibrated class probabilities nor certainties of predictions. To have comparable scores among classes and better estimates of the confidence of individual predictions, we performed a calibration process using an L2-penalized, multinomial, logistic regression model described in Capper et al. [8]. The model was fitted with the methylation class as the response variable and the transductive prediction scores as explanatory variables using *glmnet* (v.4.1-2) R package [31]. The penalization parameter was determined by running 10-fold cross-validation, and the λ that gave minimum mean cross-validated error was chosen.

Threshold analysis

Finding an optimal cutoff for diagnostic tests is usually necessary to maintain confidence in clinical settings. For example, in brain tumor classification, the cost of false negatives (i.e., a tumor cannot be classified) is usually more tolerable than the cost of false positives (i.e., a tumor is falsely predicted to a methylation class). As such, high specificity (high true negative rate) is preferable. If there are no preferences regarding specificity and sensitivity, the optimal cutoff can be chosen by maximizing the Youden Index (specificity + sensitivity – 1) [22].

In this study, the goal was to demonstrate the utility of SSL in improving a supervised classifier performance; hence, no specific metric was preferred. To find a common cutoff for all MC/MCF, ROC analysis [32] was performed, and the optimal cutoff was chosen by maximizing the Youden Index (balance between specificity and sensitivity). Since ROC analysis is usually designed for binary classification problems, we converted the multiclassification problem here into a binary problem so that a common threshold could be defined for all classes instead of having 91 individual thresholds for 91 methylation classes. Based on the maximal calibrated MC/MCF scores, a binary classification problem was defined as follows: samples that were correctly classified at these scores were considered "classifiable," while samples that were incorrectly classified were considered "non-classifiable". Under this definition, an optimal threshold (≥ 0.8) was selected, and the area under the curve (AUC) [33], specificity, sensitivity, and precision were evaluated using the *pROC* (v. 1.17.0.1) [34] and *ROCR* (v.1.0-11) [35] R packages.

Training and testing supervised and deep learning classifiers

A random forest (RF) and a neural net (NN) classifier were built using the python *scikit-learn* [36] and *keras* [37] libraries. The NN was constructed as a sequential model with three dense layers. The first and second layers contained 1000 and 500 units, respectively, with a *he_uniform* kernel initializer and *relu* activation function. The last layer consisted of a softmax layer mapping to either 75 methylation families or 91 subclasses. NN models were fit with a batch size of 6 and 20 epochs. Figure S2 showed the four combinations of training and test sets that were used to build eight supervised models and provided a schematic view of the training and testing process for the supervised models. Briefly, 30% of GSE90496 samples (841 samples), representing the inductive test set for the SSL models, were held out as an independent test set for each RF model. The first RF model was trained with the same 35% GSE90496

data (labeled samples used during the training of the best SETRED-SVM classifier). This model is referred to as the baseline model for evaluating SSL pseudo labels. All other RF models were trained using additional SSL labels from the 35% remaining GSE90496 samples plus SSL labels from the GSE109379 samples with or without threshold constraints. Balanced accuracy and weighted recall were computed for each RF after five repeated stratified 3-fold cross-validation and after predicting the labels of the inductive testing set at *random_state* = 123456. The 70-30 split was resampled seven times to create seven independent hold-out test sets to better estimate the accuracy and errors of the supervised models.

Abbreviations

AI	artificial intelligence
AUC	area under the curve
MC	methylation subclass
MCF	methylation class family
NN	neural net
RF	random forest
ROC	receiver and operating characteristic
SETRED-SVM	self-train with editing method using support vector machine as the base learner
SSL	semi-supervised learning

Declarations

Ethics approval and consent to participate: not applicable

Consent for publication: not applicable

Availability of data and materials: data were downloaded from the NCBI database with GSE90496 and GSE109379. R and Python codes are uploaded to https://github.com/stjude/Semisupervised_Learning.git and available upon request.

Competing interests: The authors declare no competing interest.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author's contributions: QTT and BAO designed the study. QTT implemented the approach in R and Python and maintained the code in GitHub. QTT and BAO drafted the manuscript. MZA involved in the discussion and helped refine the study design. QTT, BAO, and MZA interpreted the results, reviewed and edited the manuscript.

Acknowledgements: not applicable

References

1. Balyen L, Peto T: **Promising Artificial Intelligence-Machine Learning-Deep Learning Algorithms in Ophthalmology.** *Asia Pac J Ophthalmol (Phila)* 2019, **8**(3):264-272.
2. Currie G, Hawk KE, Rohren E, Vial A, Klein R: **Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging.** *J Med Imaging Radiat Sci* 2019, **50**(4):477-487.
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts H: **Artificial intelligence in radiology.** *Nat Rev Cancer* 2018, **18**(8):500-510.
4. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, Birgand G, Holmes AH: **Machine learning for clinical decision support in infectious diseases: a narrative review of current applications.** *Clin Microbiol Infect* 2020, **26**(5):584-595.
5. Shimizu H, Nakayama KI: **Artificial intelligence in oncology.** *Cancer Sci* 2020, **111**(5):1452-1460.
6. Orozco JIJ, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jalas JR, Wilmott JS, Hothi P, Wang X, Takasumi Y *et al*: **Epigenetic profiling for the molecular classification of metastatic brain tumors.** *Nat Commun* 2018, **9**(1):4627.
7. Rauschert S, Raubenheimer K, Melton PE, Huang RC: **Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification.** *Clin Epigenetics* 2020, **12**(1):51.
8. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE *et al*: **DNA methylation-based classification of central nervous system tumours.** *Nature* 2018, **555**(7697):469-474.
9. Koelsche C, Schrimpf D, Stichel D, Sill M, Sahm F, Reuss DE, Blattner M, Worst B, Heilig CE, Beck K *et al*: **Sarcoma classification by DNA methylation profiling.** *Nat Commun* 2021, **12**(1):498.
10. Camargo G, Bugatti PH, Saito PTM: **Active semi-supervised learning for biological data classification.** *PLoS One* 2020, **15**(8):e0237428.
11. Zhou Z, Li M: **Tri-training: exploiting unlabeled data using three classifiers.** *IEE* 2005, **17**(Transactions on Knowledge and Data Engineering):1529-1541.
12. Blum A, Mitchell T: **Combining labeled and unlabeled data with co-training.** In: *Proceedings of the eleventh annual conference on Computational learning theory.* COLT' 98 edn. New York, NY, USA: Association for Computing Machinery; 1998: 92–100.

13. Chapelle O, Scholkopf B, Zien AE: **Semi-Supervise Learning** (Chapelle, O. et al., Eds.; 2006) [Book review]. 2009, **20**(3):542.
14. Kim J, Shin H: **Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data.** *J Am Med Inform Assoc* 2013, **20**(4):613-618.
15. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics.** *Nat Rev Genet* 2015, **16**(6):321-332.
16. Sheikh Hassani M, Green JR: **A semi-supervised machine learning framework for microRNA classification.** *Hum Genomics* 2019, **13**(Suppl 1):43.
17. Ge C, Gu IY, Jakola AS, Yang J: **Deep semi-supervised learning for brain tumor classification.** *BMC Med Imaging* 2020, **20**(1):87.
18. Yarowsky D: **Unsupervised word sense disambiguation rivaling supervised methods.** In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics; Cambridge, Massachusetts.* Association for Computational Linguistics 1995: 189–196.
19. Li M, Zhou Z: **Setred: Self-training with editing.** In: *Advances in Knowledge Discovery and Data Mining.* vol. 3518: Springer Berlin Heidelberg; 2005: 611-621.
20. Wang Y, Xu X, Zhao H, Hua Z: **Semi-supervised learning based on nearest neighbor rule and cut edges.** *Knowledge-Based Systems* 2010, **23**(6):547-554.
21. Zhou Y, Goldman S: **Democratic co-learning.** *IEEE 2004*(IEEE 16th International Conference on Tools with Artificial Intelligence (ICTAI)):8.
22. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF: **Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection.** *Biom J* 2008, **50**(3):419-430.
23. Wei Z, Wang H, Zhao R: **Semi-supervised multi-label image classification based on nearest neighbor editing.** *Neurocomput* 2013, **119**:462–468.
24. Cohen N, Abutbul R, Hoshen Y: **Out-of-Distribution Detection without Class Labels.** In. CoRR: arXiv; 2021.
25. Wan X: **The Influence of Polynomial Order in Logistic Regression on Decision Boundary.** In: *IOP Conference Series: Earth and Environmental Science.* 2019.
26. Jenkins DG, Quintana-Ascencio PF: **A solution to minimum sample size for regressions.** *PLoS One* 2020, **15**(2):e0229345.
27. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: **Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.** *Bioinformatics* 2014, **30**(10):1363-1369.
28. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD: **Low-level processing of Illumina Infinium DNA Methylation BeadArrays.** *Nucleic Acids Res* 2013, **41**(7):e90.
29. Brodersen KH, Ong CS, Stephan KE, Buhmann JM: **The Balanced Accuracy and Its Posterior Distribution.** In: *20th International Conference on Pattern Recognition: 23-26 Aug 2019.* 2010: 3121-3124.

30. Grandini M, Bagli E, Visani G: **Metrics for multi-class classification: an overview.** *arXiv preprint arXiv:200805756* 2020.
31. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33**(1):1-22.
32. Hanley JA: **Receiver operating characteristic (ROC) methodology: the state of the art.** *Crit Rev Diagn Imaging* 1989, **29**(3):307-335.
33. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology* 1982, **143**(1):29-36.
34. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: **pROC: an open-source package for R and S+ to analyze and compare ROC curves.** *BMC Bioinformatics* 2011, **12**:77.
35. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21**(20):3940-3941.
36. Pedregosa F, Varoquaux Ge, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Duborg V *et al*: **Scikit-learn: Machine learning in Python.** *Journal of Machine Learning Research* 2011, **12**:2825-2830.
37. Gulli A, Pal S: **Deep learning with Keras:** Packt Publishing Ltd; 2017.

Figures

Figure 1

Training and testing scheme for the 11 evaluated SSL models. After preprocessing the GSE90496 methylation data, probes with a standard deviation greater than 0.3 across all 2,801 samples were selected as features used in the 11 SSL models. Thirty percent of the samples were kept aside as an inductive testing data set to independently evaluate the performance of each SSL model. The remaining 70% of the data was used as training sets. The training data were proportionally partitioned into a labeled and an unlabeled set. Specifically, 50% (of the 70%) training data were used as labeled examples, while the remaining 50% (of the 70%) data were used as unlabeled examples or as a transductive test set. The partitioning process was bootstrapped seven times. Each molecular methylation group was proportionally selected for every bootstrap to ensure that the class distributions were similar to the original class distributions.

Figure 2

Box plots summarizing SSL performance of 11 SSL models for predicting 75 methylation class families (MCF) of inductive and transductive testing data. (A) Accuracy; (B) Precision and recall; (C) Specificity

across all models. Box plots showing the results of all seven bootstraps. Asterisks represent the outliers. 1NN – pink; combine 1NN/SVM/C5.0 – green; C5.0 – cyan; and SVM – purple.

Figure 3

Box plots summarizing SSL performance of 11 SSL models for predicting 91 methylation subclasses (MC) of inductive and transductive testing data. (A) Accuracy; (B) Precision and recall; (C) Specificity across all models. Box plots showing the results of all seven bootstraps. Asterisks represent the outliers. 1NN – pink; combine 1NN/SVM/C5.0 – green; C5.0 – cyan; and SVM – purple.

Figure 4

Distributions of predicted scores for 91 methylation subclasses. (A) Before calibration (B) After calibration for classifiable cases.

Figure 5

ROC and threshold analysis for classifiable and non-classifiable methylation class (MC) cases. (A) ROC analysis of maximal raw (red) and maximal calibrated (blue) predicted scores with the area under the curve (AUC) and the corresponding specificity and sensitivity that maximized the Youden Index. (B) Specificity, sensitivity (recall), and precision at different thresholds using calibrated scores. The black vertical line represents the suggested threshold (≥ 0.8) at which we had the best balance between sensitivity and specificity using calibrated scores.

Figure 6

Prediction performance of random forest (RF) and neural net (NN) classifiers at the family level when trained with different combinations of reference samples and semi-supervised (SS) predicted labeled samples. (A) Balanced accuracy of RF and NN. (B) Proportion (left panel) and count (right panel) of high (≥ 10 samples, red) and low (< 10 samples, green) frequency referent labels, high confident (HC) SS

labels with high frequency (blue) and low frequency (orange) families, and low confident (LC) SS labels in high frequency (yellow) and low frequency families (purple). Asterisks indicate statistically significant difference performed by the Tukey Honest Significant Difference test at 0.05 alpha level.

Figure 7

Prediction performance of random forest (RF) and neural net (NN) classifiers at sub-class level when trained with different combinations of reference samples and semi-supervised (SS) predicted labeled samples. (A) Balanced accuracy of RF and NN. (B) Proportion (left panel) and count (right panel) of high (≥ 10 samples, red) and low (< 10 samples; green) frequency referent labels, high confident (calibrated SSL scores ≥ 0.8 , HC) SS labels with high frequency (blue) and low frequency (orange) subclasses, and low confident (calibrated SSL scores < 0.8 , LC) SS labels in high frequency (yellow) and low frequency (purple) subclasses. Asterisks indicate statistically significant difference performed by the Tukey Honest Significant Difference test at 0.05 alpha level.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TranSupplementaryFiguresandTables.docx](#)