Article type : Research Article

Risks of Feature Leakage and Sample Size Dependencies in Deep Feature Extraction for Breast Mass Classification

Ravi K. Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A. Helvie

Department of Radiology, University of Michigan, Ann Arbor

Short title: Feature leakage risks in deep learning

Correspondence:
Ravi K. Samala, Ph.D.
Department of Radiology
University of Michigan
1500 E. Medical Center Drive
C474 Med-Inn Bldg
Ann Arbor, MI 48109-5842
Telephone: (734) 647-8556
Fax: (734) 615-5513
E-mail: rsamala@umich.edu

ABSTRACT

Purpose: Transfer learning is commonly used in deep learning for medical imaging to alleviate the problem of limited available data. In this work we studied the risk of feature leakage and its dependence on sample size when using pre-trained deep convolutional neural network (DCNN) as feature extractor for classification breast masses in mammography.

Methods: Feature leakage occurs when the training set is used for feature selection and classifier modeling while the cost function is guided by the validation performance or informed by the test

performance. The high-dimensional feature space extracted from pre-trained DCNN suffers from the curse of dimensionality; feature subsets that can provide excessively optimistic performance can be found for the validation set or test set if the latter is allowed for unlimited reuse during algorithm development. We designed a simulation study to examine feature leakage when using DCNN as feature extractor for mass classification in mammography. 4,577 unique mass lesions were partitioned by patient into three sets: 3,222 for training, 508 for validation and 847 for independent testing. Three pre-trained DCNNs, AlexNet, GoogLeNet, and VGG16, were first compared using a training set in four-fold cross validation and one was selected as the feature extractor. To assess generalization errors, the independent test set was sequestered as truly unseen cases. A training set of a range of sizes from 10% to 75% was simulated by random drawing from the available training set in addition to 100% of the training set. Three commonly used feature classifiers, the linear discriminant, the support vector machine, and the random forest were evaluated. A sequential feature selection method was used to find feature subsets that could achieve high classification performance in terms of the area under the receiver operating characteristic curve (AUC) in the validation set. The extent of feature leakage and the impact of training set size were analyzed by comparison to the performance in the unseen test set.

Results: All three classifiers showed large generalization error between the validation set and the independent sequestered test set at all sample sizes. The generalization error decreased as the sample size increased. At 100% of the sample size, one classifier achieved an AUC as high as 0.91 on the validation set while the corresponding performance on the unseen test set only reached an AUC of 0.72.

Conclusions: Our results demonstrate that large generalization errors can occur in AI tools due to feature leakage. Without evaluation on unseen test cases, optimistically biased performance may be reported inadvertently, and can lead to unrealistic expectations and reduce confidence for clinical implementation.

Keywords: breast cancer classification, feature leakage, generalization error, pre-trained DCNN, sample size

I. INTRODUCTION

Machine learning using deep convolutional neural network (DCNN) requires training of a large number of parameters. Typically, millions of training samples are needed to train DCNNs for computer vision tasks. Through representation learning, a DCNN learns to extract features from the input image via the shallow to the deep convolutional layers. It has been shown that the DCNN extracted features are more generic, such as lines and edges, in the shallow layers, and become progressively more specific to the target task as the layers get deeper. The feature extraction capability is incorporated in the weights of the convolutional filters. Due to the limited availability of medical image data, transfer learning is often used to train DCNNs for medical imaging tasks. Transfer learning from *source* to *target* tasks in medical imaging has been implemented using different strategies but generally starting with the transfer of weights from the *source* task. During transfer learning, the dense layers at the DCNN output may be replaced or new layers added to be trained with the target domain images, while the convolutional filter weights may be frozen at different levels [1] and the remaining unfrozen layers are allowed to be fine-tuned. When the *target* domain data set is small, the pre-trained DCNN may be used directly as feature extractor and the extracted features are weighted and classified by an external classifier trained with the *target* domain data. Since features can be extracted from any layers of the DCNN, the dimensionality of the extracted feature space can be extremely high and a subset of the features are often selected before or during formulation of the classifier.[2-4] The high dimensionality of the extracted deep feature space coupled with the limited medical imaging data available creates the 'curse of dimensionality' problem.

Due to the large learning capacity, DCNNs can be over-trained to fit to the patterns or characteristics in the training set rather than learning generalizable features.[5,6] Small training sets increases the risk of over-training. Over-training results in drop in performance on unseen test cases. Regularization methods have been developed to mitigate the risk but cannot eliminate it. Identifying the balance between learning and over-fitting is not trivial. In machine learning, usage of validation set is recommended to guide training, but the overfitting may then be directed to the validation set. Over-training can also occur when a DCNN is used as feature extractor and

a subset of features is selected as input predictor variables for an external classifier. The feedback from the validation set in the form of 'data' and 'features' is used to optimize the machine learning methods on the training set. In this work, we studied the 'features' feedback as 'feature leakage'[7] between the training set and the validation set or test set that is allowed unlimited reuse.

In machine learning, 'data' or 'feature' leakage between the training and the validation/test partitions results in overly optimistic results. An example of data leakage is when data from the same patient spread across the training and test partitions. Data within each patient is highly correlated resulting in the leakage. An example of feature leakage is when feature selection on the training partition is influenced by the performance on the test set. Feature leakage can occur unintentionally even if the training and test sets are separated with independent samples. If the test set is not sequestered and the algorithm developer can reuse the same test set to evaluate the algorithm performance unlimited times during model development, the test set essentially becomes a part of the validation set that guides the feature selection and model formulation [8,9]. Feature leakage can be particularly serious in the context of DCNNs because of the high-dimensional nature of the problem. Similar problem can occur in radiomics where hundreds or thousands of texture features can be extracted and a small subset is selected to build predictive models. The risk of curse of dimensionality compounded with the issue of feature leakage in machine learning can potentially lead to overly optimistic reporting of the performance of clinical decision support tools.

The goal of this study is to demonstrate the hazards of feature leakage in the process of selecting deep features and classifier modeling, the generalization errors when the trained model is applied to truly unseen cases, and the effects of sample size on the problem.

Recent deep learning related work in medical imaging has renewed interest in developing machine learning, or artificial intelligence (AI), methods for various applications in healthcare. Transfer learning is an important technique of developing these tools, especially the use of a DCNN pre-trained with large *source* domain data as a feature extractor to alleviate the data shortage problem in medical domain. The extracted features of the *target* domain data are then used as input predictor variables to train an external classifier for the *target* task using the

available data set. The limited available data may be split into a training set and a validation set with or without another held-out test set. *K*-fold cross validation, leave-one-out, or a single split are basically the same data set partitioning in principle except for the differences in the number of partitions to split so that we will focus on the single split approach as example in the current study.[10,11] An important issue of this developmental process is to understand whether and how feature leakage occurs and the impact of the sample size on feature leakage and generalization error, i.e., bias on the predicted performance relative to true performance on unseen test data. To study this process, we use a relatively large labeled data set of malignant and benign breast masses from mammograms. A subset is sequestered as an unseen test set that is not used in any process during training. Another independent subset is drawn as a validation set. The remaining cases are used to randomly drawn training sets that simulate a range of sample sizes. Sensitivity analysis within the variability of the data is also studied by repeated experiments at each training set size. Although we use the task of classifying malignant and benign masses in mammography as example, it can be expected that the observed trends are applicable to other similar tasks.

As described above, DCNNs learn low-level functions that transform an input image to an output class through representation learning but at a large scale producing hundreds of thousands or millions of these functions [12]. Through back propagation, the layers closer to the input break down the image to build basic descriptors of the input domain, and the layers closer to the output amplify the attributes important for classification and suppress the rest. Thus the deep features in the DCNN are organized to transition from *generic* to *specific* to the *source* task. By using the DCNNs as feature extractors, these low-level functions, which can amount to thousands, can extract pertinent characteristics of the *source* domain images. The number of these deep features far exceeds that of the handcrafted features. Studies have shown that the knowledge of extracting representative features can be successfully transferred from the ImageNet 1000-class classification task to other domains.[13] The discriminability of these deep features compared to traditional radiomics features and DCNN trained on *target* task data could depend on the complexity of the task, the quality of the extracted features and the training/validation data sizes.

The effectiveness of the extracted features is also influenced by the DCNN architecture. For example, VGG16 with convolutional filters of smaller receptive field and deeper

convolutional layers compared to AlexNet was found to be superior in the ImageNet 1000-class classification task. GoogLeNet consisting of inception blocks with parallel convolutional layers and vastly smaller number of trainable weights, was found superior to both AlexNet and VGG16. For transfer learning from the *source* task to the *target* tasks in medical imaging, the quality of the deep features will further be influenced by whether and how fine-tuning with target domain data is performed and how large the training target data set is. In our previous study[1] using transfer learning with fine-tuning, where some deeper layer weights were fine-tuned in a two-stage multi-task transfer learning process, we showed that transferring 'knowledge' from ImageNet-trained weights to mammography and then to digital breast tomosynthesis (DBT) resulted in higher performance when compared to direct transfer learning from ImageNet to DBT. However, we also observed that at small training set sizes, both with mammography and DBT, the risk of overfitting increased substantially. Thus, for the current study, to focus on our goal of studying feature leakage, we used the pre-trained DCNN as feature extractor without fine-tuning. This would keep the extracted deep feature set constant and facilitate the study of the impact of feature selection without the additional variabilities due to the strategy and training set size used for fine tuning. The observed relative trends of feature leakage analyzed in this study, however, should not be dependent on the feature extractor used because we studied the feature leakage that occurred after the deep feature extraction process.

In the following sections, we will give a detailed description of the mammography data used in the study, compare three pre-trained DCNN structures as feature extractor for the mass classification task, describe the design of the simulation study and analyze feature leakage using deep features extracted from one of the pre-trained DCNNs. Preliminary results of the study were published in a conference proceedings [14].

II. METHODS AND MATERIALS

Pre-trained DCNN can be used to extract deep features in the order of thousands from each mass lesion in mammography images. In principle, the idea behind DCNNs is that the convolutional layers extract spatial features and the dense layer combines these spatial activation maps from the convolutional layer by assigning weights as determined by the loss function. In this study, we extracted deep features only at the first dense layer to keep the feature space

dimensionality relatively low, although it is still in the thousands. Since the curse of dimensionality and the feature leakage generally get worse as the dimensionality increases, the trends observed in this study will be conservative but still serve the purpose of demonstrating the problem. The following sections give a description of the data set and the partitions used in the study, and the different DCNNs used to study the behavior of deep features and training external classifiers.

II.A. Data Set

A total of 4,577 unique mass lesions from the mammograms of 1882 patient cases were used in the study as shown in Table 1. The lesions were split into 3,222 training, 508 validation and 847 independent test set by patient case. The mammography cases were collected from the University of Michigan Health System (UMHS) archives and the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [15,16]. The cases collected from the UMHS included digitized screen-film mammography (SFM) and digital mammography (DM). The cases from CBIS-DDSM included SFM cases. The mass lesions in the UMHS cases were identified by an MQSA approved breast radiologist with over 30 years of experience in breast imaging. All masses from the UMHS cases were biopsy-proven as malignant or benign. The malignancy or benignity of the masses from the DDSM database were described in the DDSM website[16]. A region-of-interest (ROI) of 256x256 pixels in size was extracted from images of 100μm x 100μm pixel size centered over the radiologist-provided bounding box. The labeled ROI were extracted in the same size and resolution from the CBIS-DDSM data set. All the ROIs were background corrected to reduce the intensity inhomogeneity due to x-ray exposure and reduced the dynamic range variation across the different image sources [17-19]. The improvement in the generalizability due to the reduction of the dynamic range variation and the advantages of multi-task learning from SFM and DM tasks were studied in our previous works [19,20]. The ROI of each mass lesion was duplicated in the RGB channel input of the pre-trained DCNN.

Table 1. Distribution of data in the training, validation, and independent sets. The partitioning is by patient so that the three sets contained independent cases.

	Training	Validation	Test

	M	B	M	B	M	B
Unique mass lesions	1,550	1,672	239	269	363	484
Total in each set	3,222		508		847	
Total	4,577					

M: Malignant, B: benign

II.B. Selection of DCNN for the simulation study

We first compared the deep features extracted by three DCNNs for breast mass classification before selecting one for this study: (a) AlexNet, (b) GoogLeNet and (c) VGG16 (Fig. 1), all pre-trained on 1.2 million ImageNet 1000-class object classification task [21-23]. AlexNet has five convolutional layers and three dense layers. VGG16 has thirteen convolutional layers and three dense layers. GoogLeNet has three convolutional layers, nine inception modules and a single dense layer. Each inception module has four convolutional layers arranged in parallel and two convolutional layers in series. GoogLeNet has the highest number of layers at 22 compared to AlexNet at 8 and VGG16 at 16 layers. However, GoogLeNet has the lowest number of trainable parameters among the three DCNNs. These three DCNNs achieved the lowest ImageNet classification error (top 5) between the years 2012 and 2014 challenges. They were commonly used in the literature[24] and represented the basis on which other more complex DCNNs were built.

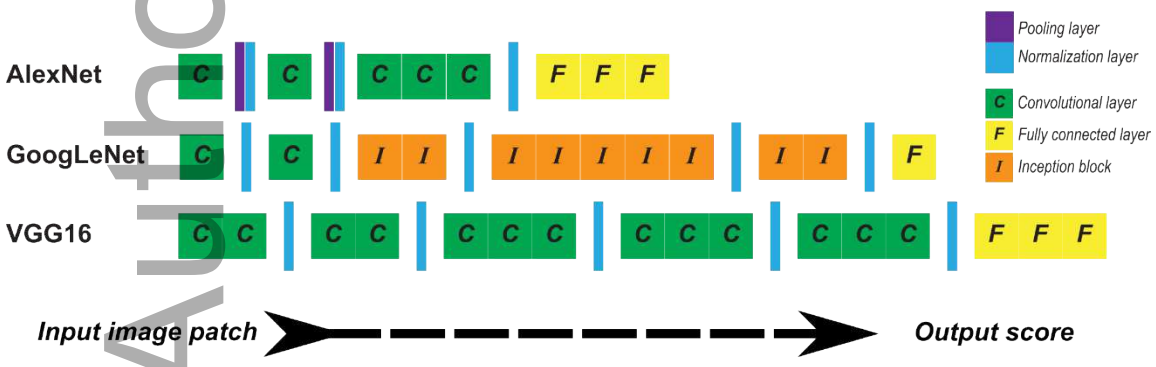


Fig. 1. Three DCNNs: AlexNet, GoogLeNet and VGG16 for comparison as feature extractor. The input image patch size is a 256x256 extracted from a mammography

image at 100μmx100μm resolution. Feature extraction is performed at the first dense layer of all three DCNNs.

For the comparison of the three DCNNs as feature extractors, the validation and the test sets described in Section II.A were not used. Only a subset of the training lesions consisting of 886 UMHS-SFM, 337 UMHS-DM and 1446 CBIS-DDSM totaling 2,669 unique lesions was used in a four-fold cross-validation for this part of the study to avoid feature leakage at this preparatory step. The features were extracted from the first dense layer of each DCNN and a random forest classifier was trained. The folds were split by patient case, and the validation fold was not involved in any step of the training process, again to ensure no feature leakage. To reduce the influence of experimental uncertainties, all experiments were repeated ten times with different random seeds to introduce randomness in the generation of the random forest trees.

The random forest classifier [25] is an ensemble algorithm that builds and aggregates the votes from multiple decision trees. The advantage of this approach is aggregation from many decision trees that reduces overfitting risk even when presented with a high dimensional feature space. Each decision tree is trained on a bootstrapped training data using randomly selected features, thus also avoiding the explicit need for feature selection. Because of the randomness in initializing multiple trees, the method scales well for large dimensional space. Due to these characteristics, both the feature selection and large dimensional feature space were internally handled thus avoiding potential bias by the developer. Note that this step of selecting the DCNN is only a precursor to the simulation study. The experimental setup for the selection of DCNN is deliberately focused on selecting the ImageNet pre-trained DCNN that could provide effective deep features for classification of masses on mammograms.

As a reference, we also trained a DCNN directly for the mass classification task with and without transfer learning from the ImageNet data, rather than as a feature extractor. The AlexNet was chosen for this comparison.

II.C. Simulation study – feature leakage

Fig. 2 shows the approach used to simulate feature leakage between training and

validation sets. The data set partitions have been described in Section II.A. The test set is kept independent of any training process so that the evaluation of the trained classifier on the test set can serve as a reference of its performance in truly unseen cases. The deep features extracted from the training set are used to build the classifier model including feature selection and weight training in a wrapper-mode. The cost function for selecting the classifier, and therefore the feature combination, at each iteration depends on the performance of the trained classifier applied to the validation set. The classification performance is assessed using the area under the receiver operating characteristic curve (AUC) estimated by the trapezoidal rule for fast calculation. This process is implemented in an automatic algorithm that searches through a large number of feature combinations to identify selected feature subsets that can provide high AUC values in the validation set. The classifier models that are found to have high AUC values will then be applied to the independent test set to assess the generalization error, estimated as the potential optimistic bias on the validation performance with reference to the true performance on unseen cases.

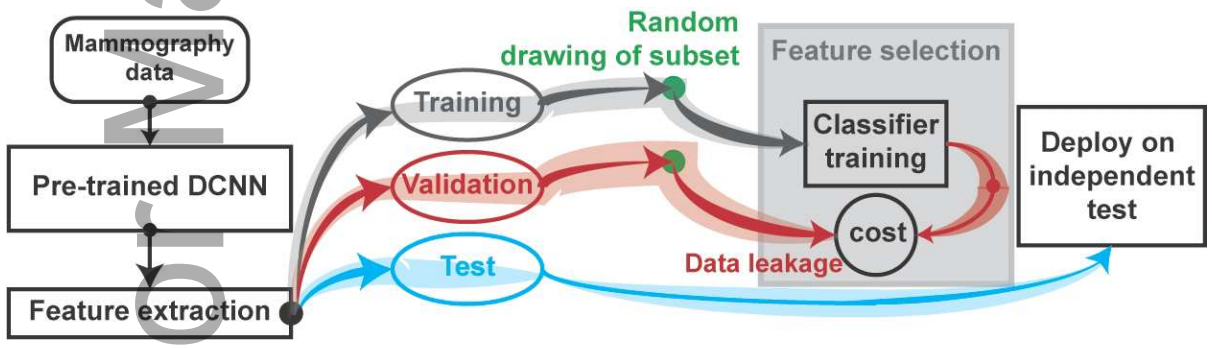


Fig. 2. Simulation study for data leakage using cross-validation and independent testing. An ImageNet-pre-trained DCNN was used to extract deep features for mass classification. The simulated sizes of training and validation data sets were varied over a range (10% to 100% of the available training and validation sets) by random drawing from the original 100% sets. The independent test set is fixed. Data leakage was introduced by directing the cost function of feature selection by the performance on the validation set. The optimistic bias, or generalization error, on

the validation set was assessed with reference to the classifier performance on the truly unseen independent test set.

To study the impact of training and validation set sizes on feature leakage, we simulated a range of available sample size by randomly drawing a subset from the original training set (3,222 at 100%) and validation set (508 at 100%), respectively. The simulated sample sizes at 75%, 50%, 25% and 10% were studied in addition to the 100%. To keep the ratio of malignant and benign masses about the same as the original set, the desired percentages were separately drawn from the two classes. The performance on the independent test set was assessed at the 100% size (847 masses) for all conditions such that the generalization errors can be compared on the same set of unseen cases as reference. The experiment at each sample size was repeated 10 times by randomly drawing from the original set to simulate sample variations.

Sequential feature selection (SFS) is a simple and fast feature selection algorithm used in wrapper-based mode. It belongs to the family of greedy search algorithms, the AUC of the classifier on the validation set iteratively guides the SFS to find the "best" feature combinations. In our implementation, it starts by finding a single best feature (n=1) from the available feature pool containing a total of M features, and adds a new feature sequentially at a time until it is terminated by a preset maximum number of features to be included. For a given subset of n features that has already been selected, all new combinations of (n+1) features obtained by combining the n features with one of the remaining (M-n) features in the feature pool are compared, resulting in (M-n) feature combinations as input to the classifier being evaluated for their AUC values. Each of this process is called an iteration in the following discussion. The combination yielding the best performance based on a cost function is then identified to be the subset of (n+1) selected features and continues onto the selection of the next feature. Thus, SFS is ideal for this study because we can implement the feature selection process with an automated algorithm to search through a large number of feature combinations efficiently and systematically. This can simulate an AI developer optimizing the deep feature selection and classifier modeling while checking the performance on the validation set, or retesting on the test set many times until a satisfactory performance is found. Further, the range of overfitting observed through this analysis provides the extent of the risks from the perspective of a developer with computer-assisted search.

Three classifiers were chosen to verify if there are any advantages of using a simple classifier like linear discriminant analysis (LDA) classifier or more complex non-linear classifiers like the random forest or support vector machine (SVM). An LDA classifier models the class conditional distribution of the training data to generate a discriminant function for classification. LDA is optimal for multivariate normally distributed feature spaces with equal covariance matrices and the coefficients of the transformation function provide simple interpretation. LDA is not effective for complex feature spaces but is least prone to overfitting compared to complex classifiers when the training set is small.[26] Support vector machine (SVM) classifier constructs hyperplanes in the multidimensional feature space that maximizes the separation of classes. We used the SVM with radial basis function (RBF) kernel to map the input space. In comparison to the LDA, SVM with RBF kernel is known to handle high dimensional feature space and control the effects of outliers.[27] SVM also has the advantage of interpreting the transformation function. However, SVM due to the algorithmic complexity do not scale well with large-scale tasks. In addition, as the dimensionality of the feature space increased, due to limited samples, the feature space becomes sparse, affecting the construction of the hyperplanes in SVM. We chose fixed gamma and *C* values at 0.1 and 1.0, respectively, for the RBF. The random forest classifier described in Section II.B was used as the third classifier. We had evaluated the random forest parameters for the mass classification task in a previous study [28] and chose the total number of trees and the tree depth at 100 and 10, respectively, which were therefore also chosen for the experiments in the current study.

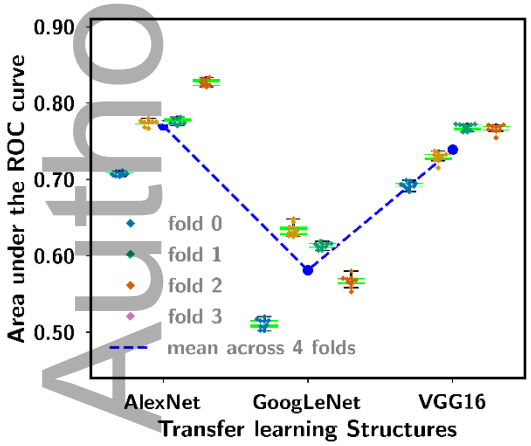
For reference, a commonly used SVM classifier with linear kernel was also evaluated with the same experimental settings except that no feature selection was performed but the regularization parameter (*C*) was chosen based on the validation performance. This type of classifier parameter optimization guided by validation performance is a common approach during classifier design. The performance curves would reveal the impact of sample size without explicit feature leakage on this classifier, and the difference between the performances on the validation set and the sequestered test set demonstrates the generalization error if the validation performance is reported without independent testing or if the "test" set is repeatedly used.

III. RESULTS

To limit the number of conditions in this study, we first compared the deep features extracted by three commonly used DCNNs for mass classification, and selected one as the feature extractor for this study. Using the deep features from the selected DCNN, the simulation study is performed with the SFS feature selection method and three classifiers while varying the sample sizes for the training and validation sets.

III.A. Selection of pre-trained DCNN for mass classification in mammography

A subset of the training partition containing 2,669 masses was used in four-fold cross-validation to evaluate the deep features from the pre-trained DCNN. Features from the first dense layer of all the DCNNs were extracted, resulting in 4096, 1024 and 4096 features from AlexNet, GoogLeNet and VGG16, respectively, for each mass. As discussed in Section II.B, we used the random forest classifier for this relative performance comparison. The four-fold cross-validation was repeated ten times with different seed initialization of the random forest classifier to estimate the variance. The results from the cross-validation are shown in Fig. 3 in a box plot. The average test performance across the four folds is indicated by a blue dot, which shows AlexNet performed the best while GoogLeNet the worst. The mean AUC for AlexNet, GoogLeNet and VGG16 were 0.77±0.04, 0.58±0.05 and 0.74±0.03, respectively. Thus, AlexNet was chosen for the simulation study.



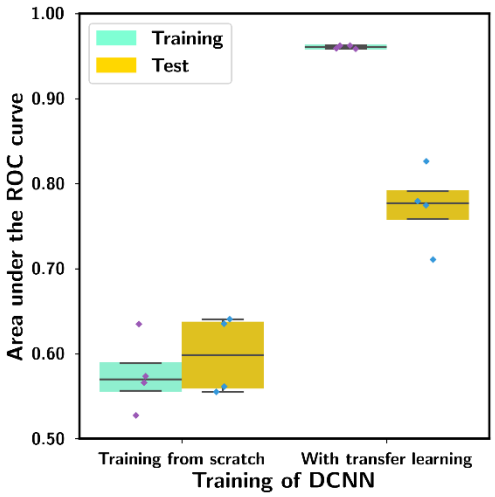


Fig. 3. Selection of DCNN among three common DCNNs for classification of masses in mammograms. Each pre-trained DCNN was used as feature extractor without fine-tuning by training data from the mass classification task and random forest was used as classifier. Four-fold cross-validation with the training set was used in the experiments. All experiments were repeated ten times using different stochastic initializations. The error bars indicate 95% confidence interval. The average performance of the test AUCs over the four folds is indicated by a blue dot, linked by a blue dotted line to facilitate visualization.

Fig. 4. Comparison of AlexNet training from scratch and with transfer learning from ImageNet data. The box plot shows the AUCs from the four-fold cross-validation.

The advantage of transfer learning from non-medical imaging domain with large data set to medical imaging can be seen in the comparison between training DCNN from scratch and with transfer learning for AlexNet shown in fig. 4. Training from scratch was unstable for all four folds and the test performance was much lower than that with transfer learning. It is interesting to note that, for the AlexNet with transfer learning, the average test AUCs were similar when it was used as a feature extractor with an external random forest or used directly for the mass classification task.

III.B. Simulation study: Feature leakage

To simulate feature leakage, the classifier weight training or formulation was performed on the training set, while the trained classifier was applied to the validation set for each feature

combination examined by SFS so that the feature search was informed by the classifier performance on the validation set. Fig. 5 illustrates the SFS feature search process to select up to a maximum of 75 features from the 4096 deep features extracted by the AlexNet, using 100% of the training and validation data and an LDA classifier. About 300K different combinations of features were assessed (i.e., about 300K iterations as plotted) before arriving at the best 75 features. We set a maximum of 75 selected features to limit the time required for the experiments and also we observed that this was large enough to reach a validation AUC greater than 0.8, which is sufficiently high to demonstrate feature leakage for the mass classification task in this study. Fig. 5(a) and 5(b) show the increasing number of features selected and the variation of the validation AUCs for all evaluated feature combinations as the SFS method sequentially selected the best combination of features. The histogram of the validation AUCs is shown in fig. 5(c) and the top 20K AUCs were highlighted. Fig. 5(d) shows the corresponding AUCs when deployed on the independent unseen test set, where the AUCs corresponding to those highlighted in fig. 5(c) were also highlighted. Despite the wide range of AUCs in the validation set, the AUCs on the unseen cases distributed in a much narrower range. The highlighted AUCs on the validation set had an average AUC of 0.91, whereas the corresponding AUCs for the independent test set reached an average AUC of only 0.72. The large difference between the average AUC values on the validation and on unseen cases is indicative of the optimistic bias on the classifier, and thus the large generalization error. In addition, for a given training set, although higher and higher validation AUCs may be found by further feature search, the AUCs on the unseen cases are relatively stable, indicating that searching for extremely high validation performance only increases generalization errors without benefiting performance in unseen cases. To observe trends over different experimental conditions, these average AUC values were tracked for each experiment, as described next.

III.C. Simulation study: Effects of training and validation sample size

Feature leakage results in large generalization error as shown in fig. 5. However, the extent of generalization error caused by feature leakage also depends on the sample size. To study the effects of the training and validation sample sizes on the feature leakage, LDA, SVM and random forest classifiers were trained while using SFS to select features. The training set and validation set sizes were varied together by the same percentage; the size of the test set was fixed

400 at 100% (847 masses) to avoid introducing more variables and provide a consistent reference for comparison.

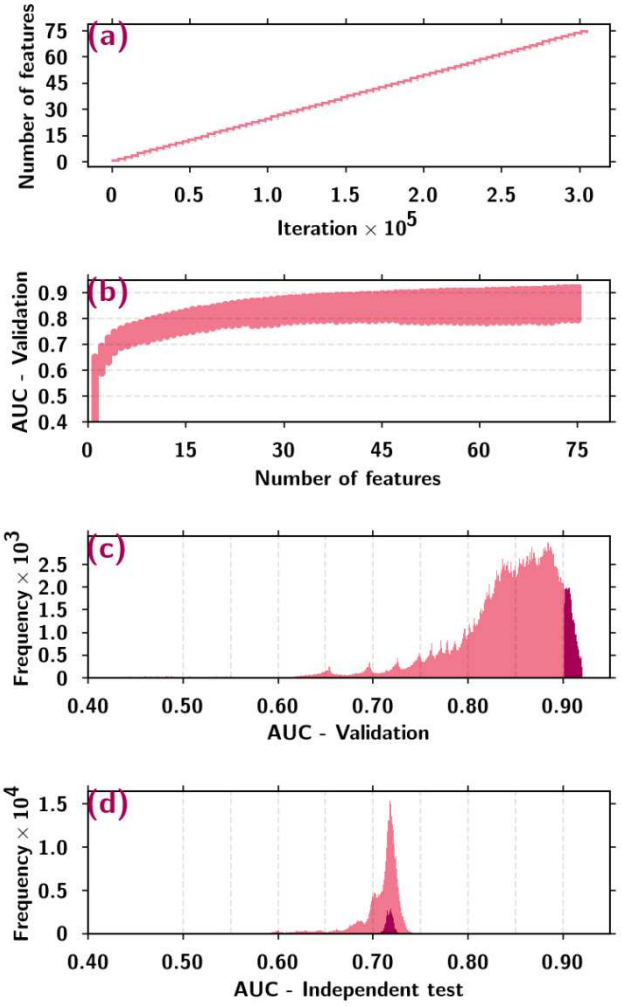


Fig. 5. Feature selection based on high AUCs from the validation set using SFS for feature selection and LDA for classification. The training and validation partitions were used at 100%. (a) Number of features selected as the number of iterations increased in the SFS method. (b) Performance on the validation set for the various feature combinations as the number of iterations increased in the SFS method. (c) Histogram of all the validation AUCs. The top 20K AUC values are highlighted. (d) Histogram of the corresponding test AUCs, in which the highlighted AUC values corresponded to those highlighted in (c). The average AUC values in the highlighted region for validation and test set are 0.91 and 0.72, respectively.

Fig. 6 shows the AUC obtained from the best selected feature set for the three classifiers when the feature set was increased by one single feature at a time using the SFS algorithm. For demonstration purpose, we tracked the validation and the test AUCs up to 150 selected features from the available 4096 features for two validation set sizes at 10% and 100% for this experiment. At the point of 75 features selected, the validation AUCs ranged from 0.78 to 0.99 whereas the test AUCs ranged from 0.55 to 0.72. The changes were relatively gradual after 75 features so that we chose 75 to be the maximum number of features for the rest of the experiments. Of the three classifiers, the random forest appeared to be least prone to overfitting with the smallest generalization errors, especially when the sample size was large (see random forest curves at 100%).

At each simulated sample size of 75%, 50%, 25% and 10%, the smaller training set and validation set were obtained by random drawing without replacement from the respective original 100% sets. Ten repeated experiments were performed for each condition. As the simulated sample size increased, the variation in the data set for the 10 repeated experiments decreased. One major reason is that a larger and larger subset was drawn from the original set and more cases would overlap among different drawings, which contributed to the smaller spread in the repeated experiments in the boxplot. At 100%, the entire set was used in a single experiment. Another reason is that the variance of the performance decreases with increasing training set size.[29] In each experiment, the maximum number of selected features were fixed at 75 and the top 20K validation AUCs in the feature selection process were averaged to obtain an average AUC, as described in Section III.B. The average AUC of the validation and the corresponding average AUC for the test set were tracked for each experiment and used to plot the boxplot in fig. 7. The dotted line shows the mean performance over the 10 experiments at each sample size and the difference between the two mean curves from validation and testing indicates the average generalization error. These results show that feature leakage resulted in generalization error over the range of sample sizes studied. The generalization error increased as the validation sample size decreased. At a validation set size of 10% (about 50 masses) the validation AUCs were 0.9 or higher while the AUC in unseen cases were about 0.5 to 0.65. Even with a large validation set (508 masses at 100%), the generalization errors in AUC were

still as large as 0.1 to 0.2. The scenario of reporting these validation results when independent unseen test set is not available can set unreasonably high expectations for the AI tool being reported.

435

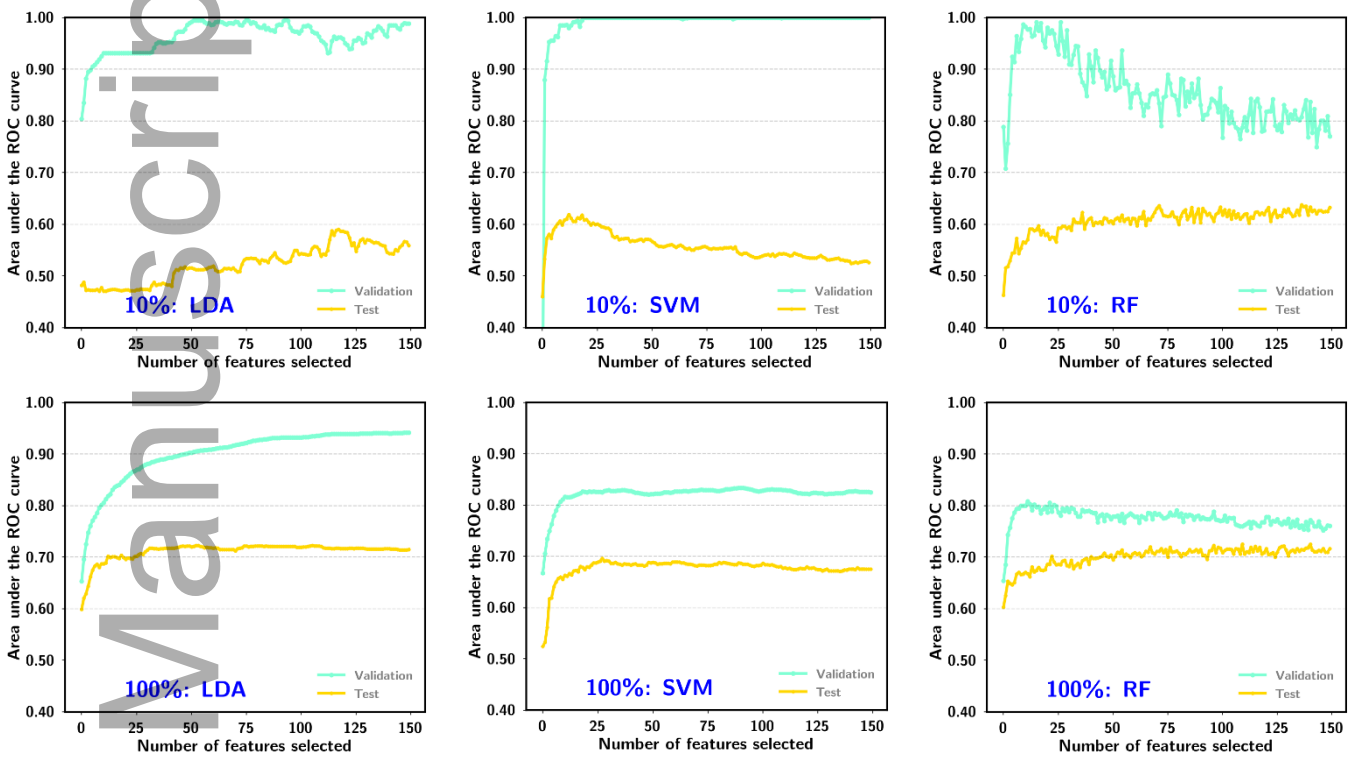


Fig. 6. The AUC performance by selecting the best set of N features that provided the highest AUC on the validation set. The number of selected features was tracked up to 150. The performances on the validation and independent test sets are shown for the simulated training and validation sample sizes of 10% and 100% of the respective available data sets in this study. The test set was fixed at 100%.

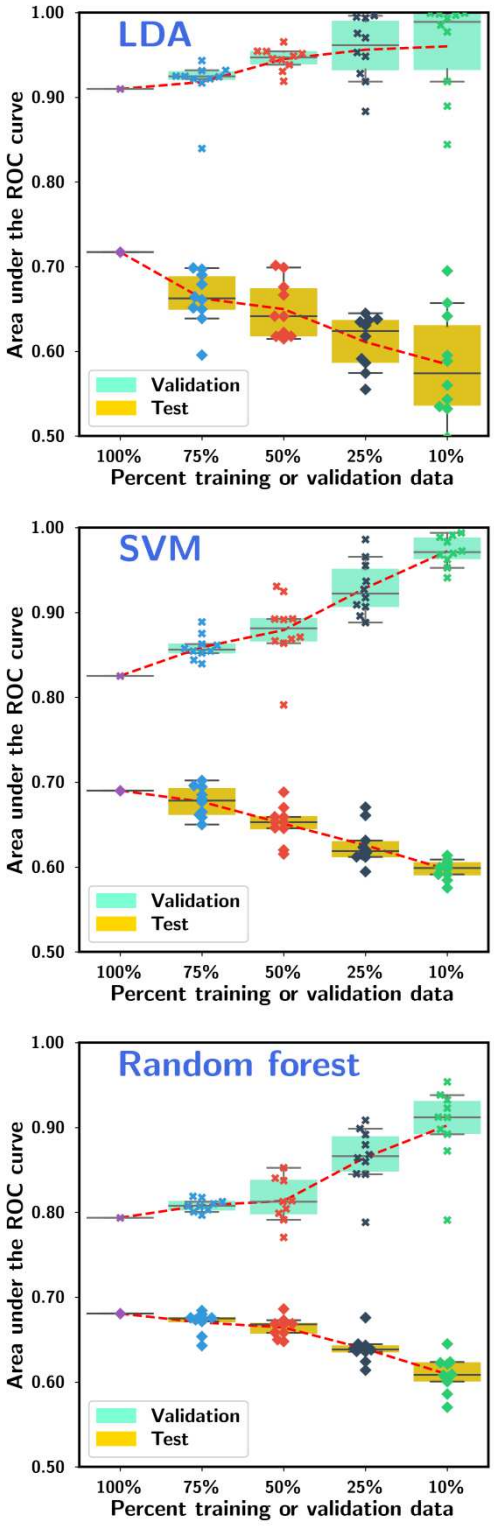


Fig. 7. Performance curves to study the feature leakage and finite sample size dependency for three classifiers. The box plot was obtained from ten repeated experiments for each condition. In each experiment the training and validation sets were randomly

drawn from the respective original 100% sets to analyze the sensitivity of the results to variations in the data set. The test set was fixed at 100%. The red dotted line indicates the mean performance.

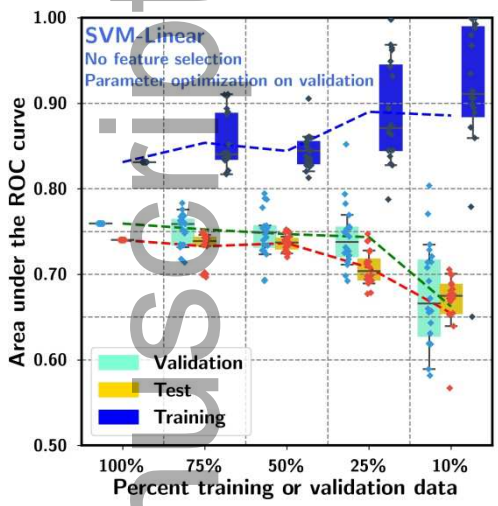


Fig. 8(a) Performance curves to study the finite sample size dependency for SVM-linear classifier without feature selection but with optimization of regularization parameter (*C*) based on the validation set performance. The box plot was obtained from 20 repeated experiments for each condition. In each experiment the training and validation sets were randomly drawn from the respective original 100% sets to analyze the sensitivity of the results to variations in the data set. The test set was fixed at 100%. The blue, green and red dotted lines indicate the mean performance on the training,

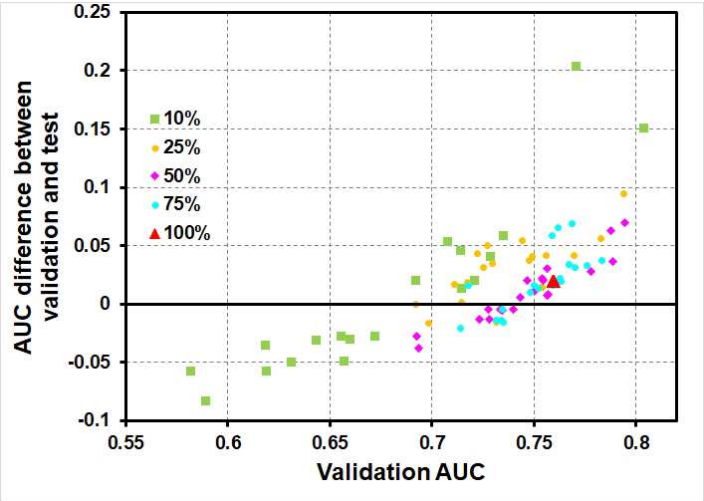


Fig. 8(b) Plot of the difference in the corresponding validation and test AUCs as a function of the validation AUC for each experiment in fig. 8(a). Five marker types with different colors were used to indicate the sample sizes at 100%, 75%, 50%, 25% and 10%.

validation and test set, respectively. The data points of the three curves were plotted offset to facilitate visualization.

The performance curves of the SVM with linear kernel trained under the same experimental conditions using all 4096 deep features as input without feature selection are shown in fig. 8(a). For each experiment, the selection of the regularization parameter (*C*) of the SVM classifier with linear kernel was guided by the validation AUC. The difference between the corresponding validation AUC and test AUC as a function of the validation AUC for each experiment is plotted in fig. 8(b). Although the mean generalization errors appeared to be only about 0.02 to 0.04 over the sample size range studied, fig. 8(b) revealed a strong trend of sample size dependence. The generalization error was between -0.1 and 0.2 at 10% sample size and about 0.02 at 100% sample size, with the variance decreasing with increasing sample size. Thus, even without feature leakage due to feature selection, classifier parameter selection guided by validation performance or repeated use of test set will still introduce generalization errors.

IV. DISCUSSION

The important issues of limited data availability and the repeated use of test data leading to bias in the predicted performance of machine learning algorithms have been recognized before the era of deep-learning[9,30] but have not been systematically studied. Due to the large number of parameters for training and the large number of features that can be extracted with deep-learning models, these issues have exacerbated and it is important to understand the risks and take necessary caution to avoid over-training. In this work, we studied the importance of data usage in the context of deep learning using classification of masses in mammography as an example.

Deep learning with transfer learning has the potential to develop robust computer-assisted tools in medical imaging. However, unlike the traditional feature engineering approaches where significant domain knowledge was needed to develop and achieve presentable results, deep

learning, due to representation learning and ease of using developer tools, can be used with minimal effort and thus prone to oversight of fundamental issues in machine learning field. In this work, we studied an important application with transfer learning, where DCNN is used to extract thousands of deep features, from which predictor variables are selected to build classifier model for a classification task. With a limited available data set for model development, it is often split into training and validation sets without an independent test set. The validation set is used both for model optimization and performance reporting. Even when a test set is reserved, if it is reused numerous times for testing when high performance cannot be achieved at a few trials, the test result essentially is used to guide model selection. The feature leakage between the training and validation/test set during the numerous trials lead to over-training of the classifier and overfitting to the validation/test set. We designed a simulation study to demonstrate the effects of feature leakage in the cross-validation scenario, where the developer optimizes the model while using the feedback from the validation/test set in the process. We showed that over-training can occur for all three classifiers studied, and the bias or generalization error of the predicted performance increases as the training set size decreases. Even in the case of classifier modeling without explicit feature selection, using the validation/test performance to guide parameter or model selection will still introduce bias. Without a sequestered representative test set for evaluation of the model generalizability, the reported validation/test results will be overly optimistic.

Three popular DCNNs in medical imaging, AlexNet, GoogLeNet and VGG16 were compared for the mass classification task and AlexNet was chosen as the deep feature extractor for the current study. Although we selected one feature extractor to simulate how DCNN may be used to extract deep features, it is expected that the feature leakage trends observed in this study should be applicable to similar problems when features are selected from a large feature space with thousands of features, regardless of the specific feature extractor used or whether the pre-trained DCNN is fine-tuned with target domain data before it is used to extract deep features. In fact, even manually extracted texture features in radiomics that can easily add up to hundreds or thousands will suffer the same high risk of feature leakage problem while selecting features and building predictive model using a limited data set.

We chose SFS as the feature selection method for this study due to its efficiency. However, the feature leakage problem exists regardless of the feature selection method used. According to the well-known "curse of dimensionality" problem in the machine learning field, there always exist feature subsets in a high-dimensional feature space that can provide excessively high classification performance for a small data set, e.g., the validation set. The role of the feature selection algorithm is only to search for these existing feature combinations, but not to create them. We tried other feature selection methods such as genetic algorithm and found that it is very slow and often trapped in local maxima, depending on the choice of parameters (number of chromosomes, crossover rate, mutation rate, etc) so that it is extremely time consuming. Exhaustive search is impractical because of the astronomical number of possible feature combinations needed to be checked when selecting up to 75 features from the 4096 deep features. On the other hand, the SFS method searches for the feature combinations systematically without depending on choices of parameters. It can be expected that the trends observed in this study do not depend on the feature selection method as long as some high-performance feature combinations can be found, and this simulation study is possible only if these feature combinations can be found within a reasonably achievable computation time due to the large number of experiments for the various conditions we examined. In reality, the SFS may not have found the highest performance feature combinations. However, as can be seen from our analysis, we only used the average AUCs from the top 20K feature combinations for comparison and demonstrated that feature leakage can occur even without reaching the extreme situations.

Fig. 7 shows that all three classifiers could achieve overly optimistic performance specific to the validation set, resulting in large generalization errors. The generalization errors of all classifiers follow a similar trend, increasing with decreasing sample size. At 10% of the training and validation set size, LDA had the largest generalization error while random forest classifier had the smallest. At 100% sample size, the LDA, SVM and random forest classifier had AUCs of 0.91, 0.83, 0.80 on the validation set and 0.72, 0.69, 0.68 on the unseen independent test set, respectively. Ideally, if the validation sample size can increase well beyond the sample size available for this study (i.e., >>100%) to sufficiently cover all characteristics of unseen cases in the population, optimizing the feature classifier to the validation set should be the same as optimizing for the population so that the bias should eventually reduce to negligibly small. However, many studies in medical imaging only have small data sets and may not reserve

independent unseen cases or cannot avoid reuse of the independent test set repeatedly during model development. Optimistic bias on the reported results due to feature leakage can be substantial.

There are limitations in the study. To reduce the complexity and the computational resource required for the experiments, we kept the number of variables and conditions manageably small. First, we compared three popular DCNNs in medical imaging as deep feature extractor and chose only one, AlexNet, for further analysis. Second, we chose the SFS method to search for the feature subsets that can provide high classification performance without exhaustively searching for those with the highest performances. Third, the breast mass classification task was used as example because of the availability of a large data set. However, as discussed above, the trends observed in this study should not be specific to the feature extractor, the feature selection method used, or the target classification task. If a different DCNN or target task is used for deep feature extraction, the deep features may have different distributions and different discriminative power, but the bias exists as long as a feature subset is selected from the high-dimensional feature space based on excessive search for high performance on a small data set. Fourth, we chose three commonly used feature classifiers (LDA, SVM, and random forest) as examples, and the user-selected parameters for the latter two were fixed. Since all three classifiers show similar trends, i.e., feature leakage leads to optimistic bias and the bias increases with decreasing sample size, it can be expected that the trends would not be very different for other classifiers or other parameters. Overall, although the amount of generalization error between the validation set and the unseen data may depend on the deep feature space or classifier used, the trends of optimistic bias and its dependence on the sample size used for guiding the model design should be general.

V. CONCLUSION

Development of AI tools in medical imaging is an important area of research. With the limited data set available in many clinical applications, use of pre-trained DCNN as feature extractor and selecting feature subsets from the large-dimensional feature space to form an external classifier is a common approach. In order to maximize the training data, AI developers may not reserve independent unseen cases or cannot avoid using a single independent test set

repeatedly during model development. Feature leakage is one of the potential sources leading to generalization errors due to over-training of the classification methods in order to achieve high performance on the validation set or the test set.

We studied feature leakage in a cross-validation approach over a range of sample sizes used by developers in medical imaging where data set sizes are small compared to other computer vision areas. When the validation set is used for guiding feature selection and classifier optimization, the validation performance is optimistically biased relative to the performance in independent unseen cases. It can be expected that guidance can also occur indirectly if an independent test set is not sequestered and is allowed to be reused unlimited times. The generalization error worsens as the data set size decreases. A similar problem occurs in radiomics where a small feature subset is often selected from a high dimensional feature space to build predictive model for clinical decision support. Optimistic biases on the reported performance may drive unrealistic expectations in the field and thus proper rules in machine learning should be followed when developing AI tools and reporting their performances. AI users should also be aware of the potential biases and evaluate the generalizability of a predictive model in the patient population of interest properly before implementation for clinical use.[31,32]

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA214981. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. RKS is also supported by the Basic Radiological Science New Investigator Award from the Department of Radiology at the University of Michigan.

REFERENCES

1. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Richter CD, Cha K. Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging.* 2019;38(3):686-696.
2. Lao J, Chen Y, Li Z-C, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports.* 2017;7(1):1-8.
3. Paul R, Hawkins SH, Balagurunathan Y, et al. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. *Tomography: a journal for imaging research.* 2016;2(4):388.
4. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics.* 2016;21(1):31-40.
5. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research.* 2014;15(1):1929-1958.
6. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Richter C. Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis. *Physics in Medicine & Biology.* 2020;65(10):105002.
7. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD).* 2012;6(4):1-21.
8. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research.* 2010;11(Jul):2079-2107.
9. Petrick N, Sahiner B, Armato III SG, et al. Evaluation of computer-aided detection and diagnosis systems. *Medical physics.* 2013;40(8):087001.
10. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media; 2009.
11. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics.* 2005;21(15):3301-3307.

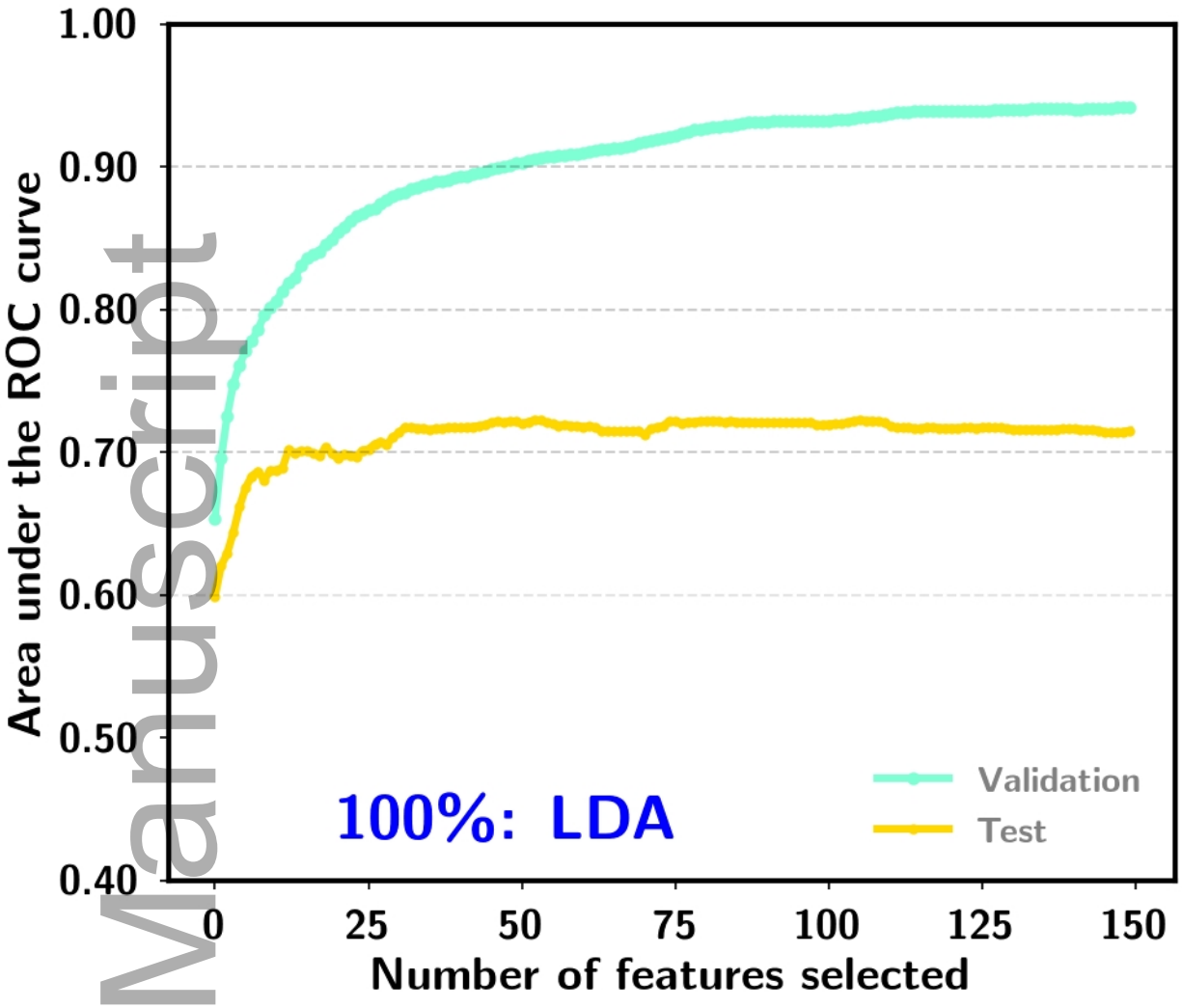
12. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
13. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: an astounding baseline for recognition. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition workshops.* 2014.806-813.
14. Samala RK, Chan H-P, Hadjiiski L, Koneru S. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. *Proc SPIE medical imaging.* 2020;11314:1131416.
15. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data.* 2017;4:170177.
16. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P. The digital database for screening mammography. In: Yaffe MJ, ed. *Digital Mammography; IWDM 2000.* Toronto, Canada: Medical Physics Publishing; 2001:457-460.
17. Chan H-P, Wei D, Helvie MA, et al. Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space. *Physics in Medicine and Biology.* 1995;40:857-876.
18. Sahiner B, Chan H-P, Petrick N, et al. Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images. *IEEE Transactions on Medical Imaging.* 1996;15:598-610.
19. Samala RK, Chan H-P, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics.* 2016;43(12):6654-6666.
20. Samala RK, Chan H-P, Hadjiiski LM, Helvie MA, Cha K, Richter C. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine and Biology.* 2017;62:8894-8908.
21. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems.* 2012.1097-1105.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv:14091556.* 2014.
23. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 2015.1-9.

24. Chan H-P, Samala RK, Hadjiiski LM. CAD and AI for breast cancer—recent development and challenges. *The British Journal of Radiology.* 2019;92:20190580.
25. Breiman L. Random forests. *Machine learning.* 2001;45(1):5-32.
26. Chan H-P, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. *Medical Physics.* 1999;26:2654-2668.
27. Lausser L, Kestler HA. Robustness analysis of eleven linear classifiers in extremely high–dimensional feature spaces. Paper presented at: IAPR Workshop on Artificial Neural Networks in Pattern Recognition2010.
28. Samala R-K, Chan H-P, Hadjiiski L, Szerman N. Comparison of transfer learning and deep feature extraction strategies for breast cancer classification in mammography using deep neural networks. *RSNA Program Book.* 2018.SSG13.
29. Fukunaga K. *Introduction to Statistical Pattern Recognition.* 2nd ed. New York: Academic Press; 1990.
30. Gur D, Wagner RF, Chan HP. On the repeated use of databases for testing incremental improvement of computer-aided detection schemes. *Academic Radiology.* 2004;11(1):103-105.
31. Chan HP, Hadjiiski LM, Samala RK. Computer-aided diagnosis in the era of deep learning. *Medical Physics.* 2020;47(5):e218-e227.
32. Huo Z, Summers RM, Paquerault S, et al. Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use. *Medical physics.* 2013;40(7):077001.

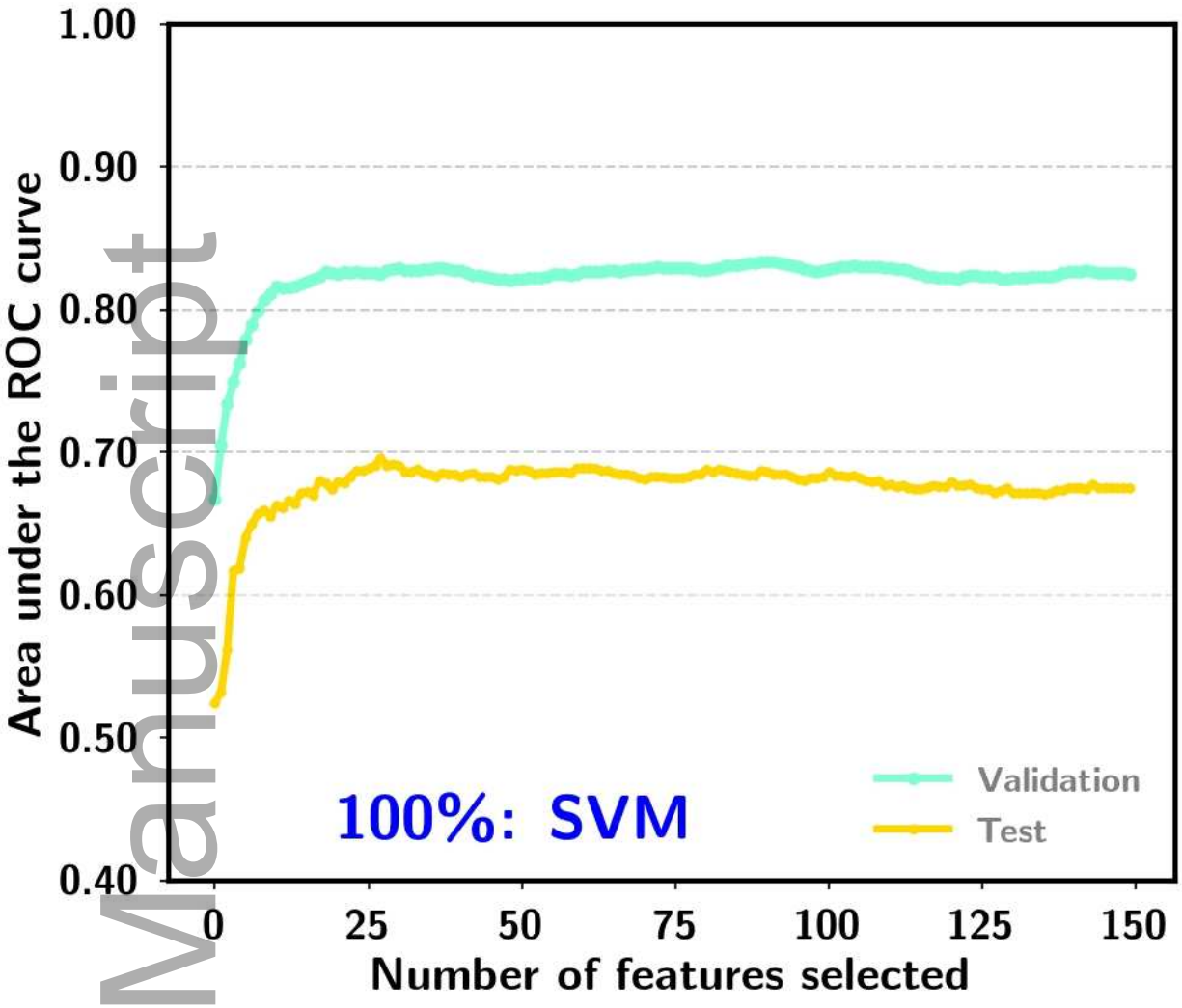
Table 1. Distribution of data in the training, validation, and independent sets. The partitioning is by patient so that the three sets contained independent cases.

	Training		Validation		Test	
	M	B	M	B	M	B
Unique mass lesions	1,550	1,672	239	269	363	484
Total in each set	3,222		508		847	
Total	4,577					

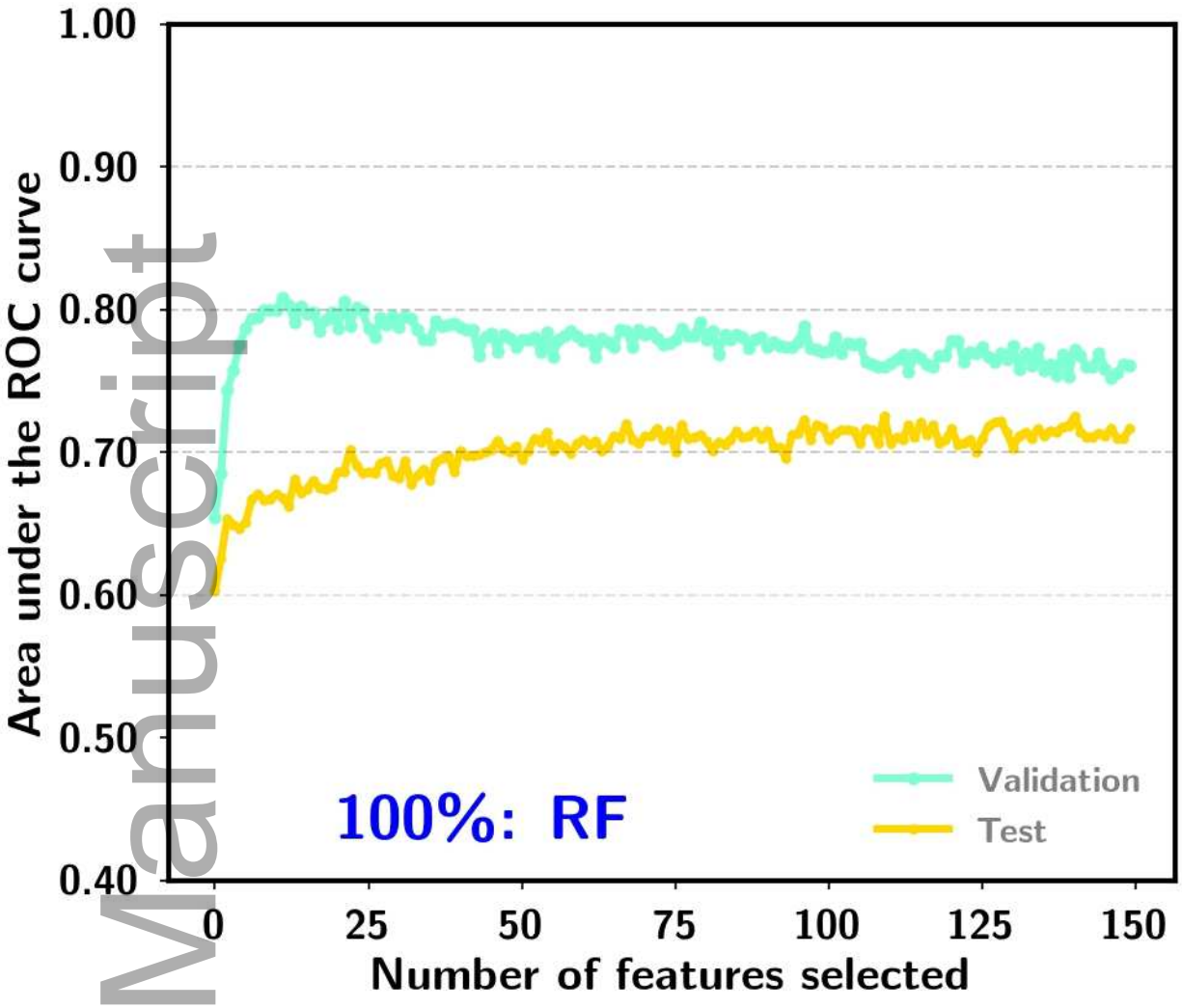
M: Malignant, B: benign



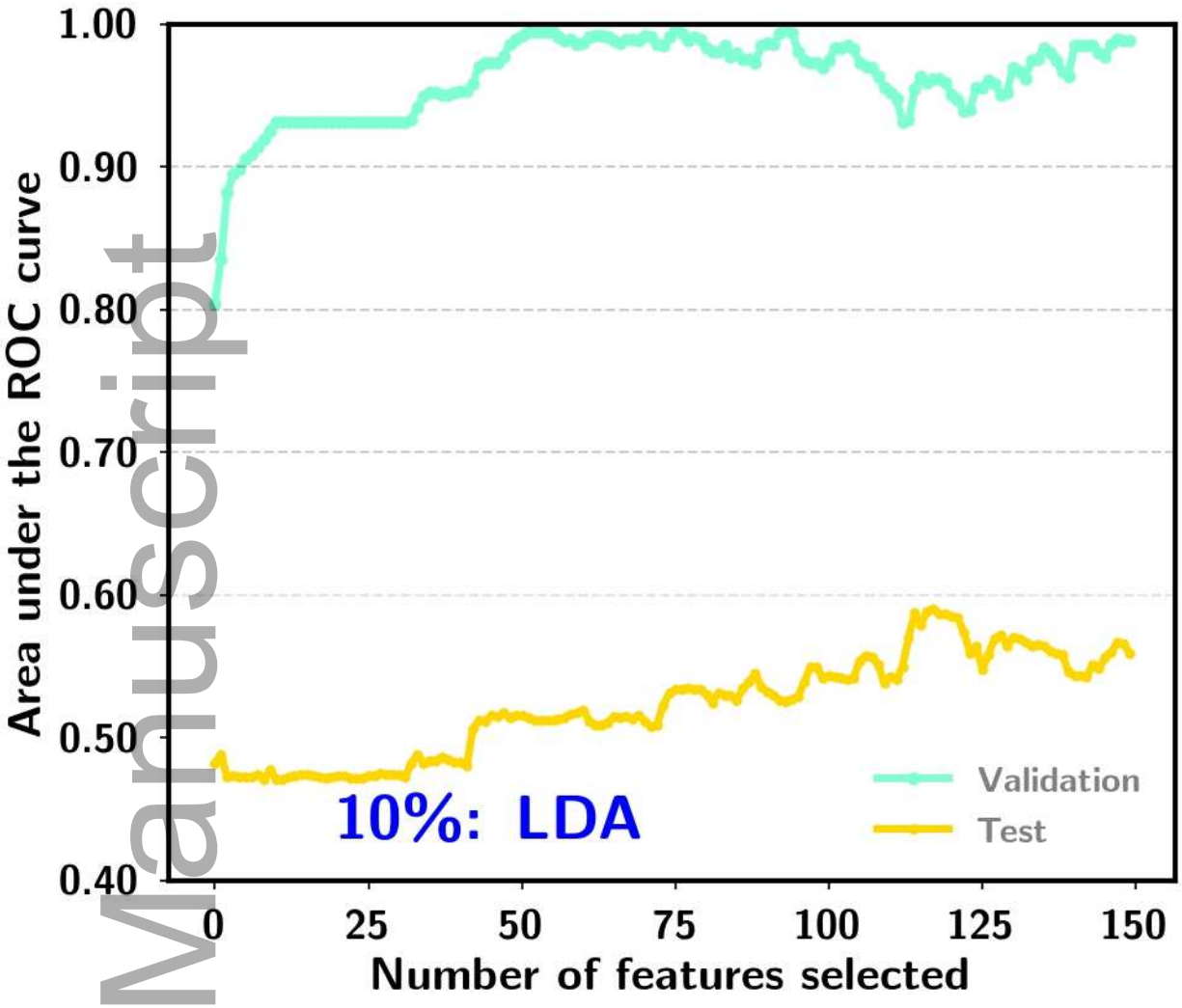
mp_14678_f6__bottom_left.jpg



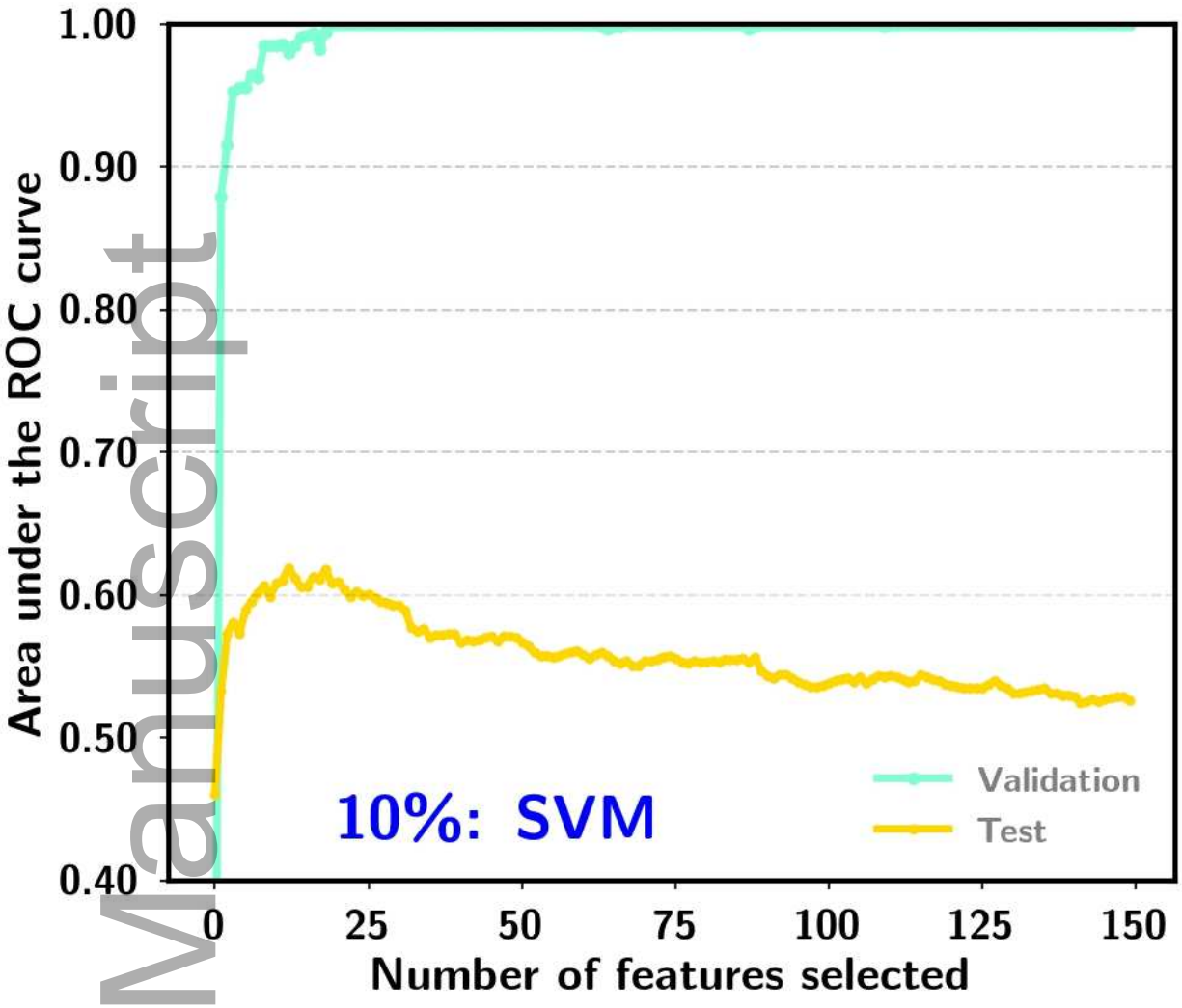
mp_14678_f6__bottom_middle.jpg



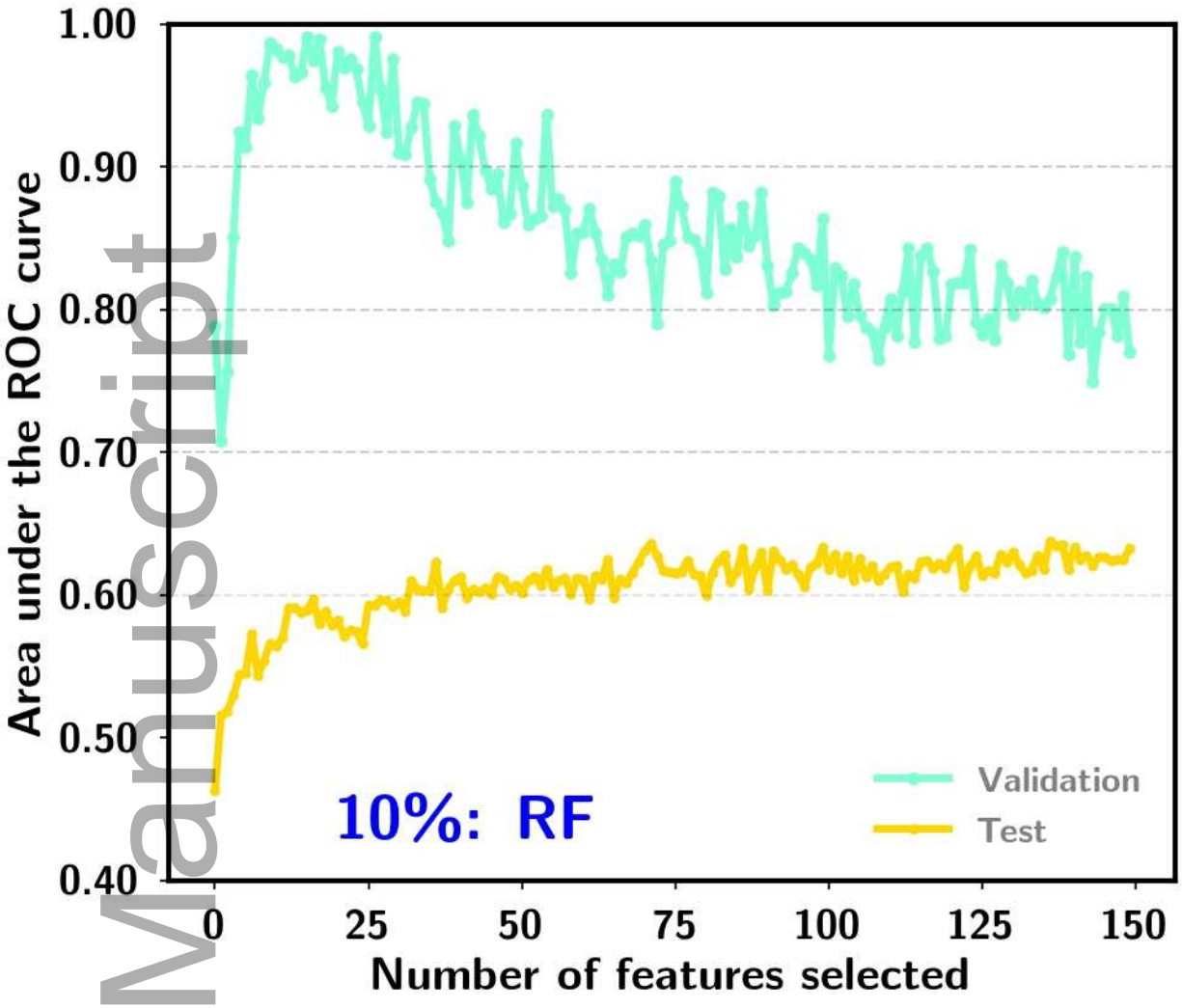
mp_14678_f6__bottom_right.jpg



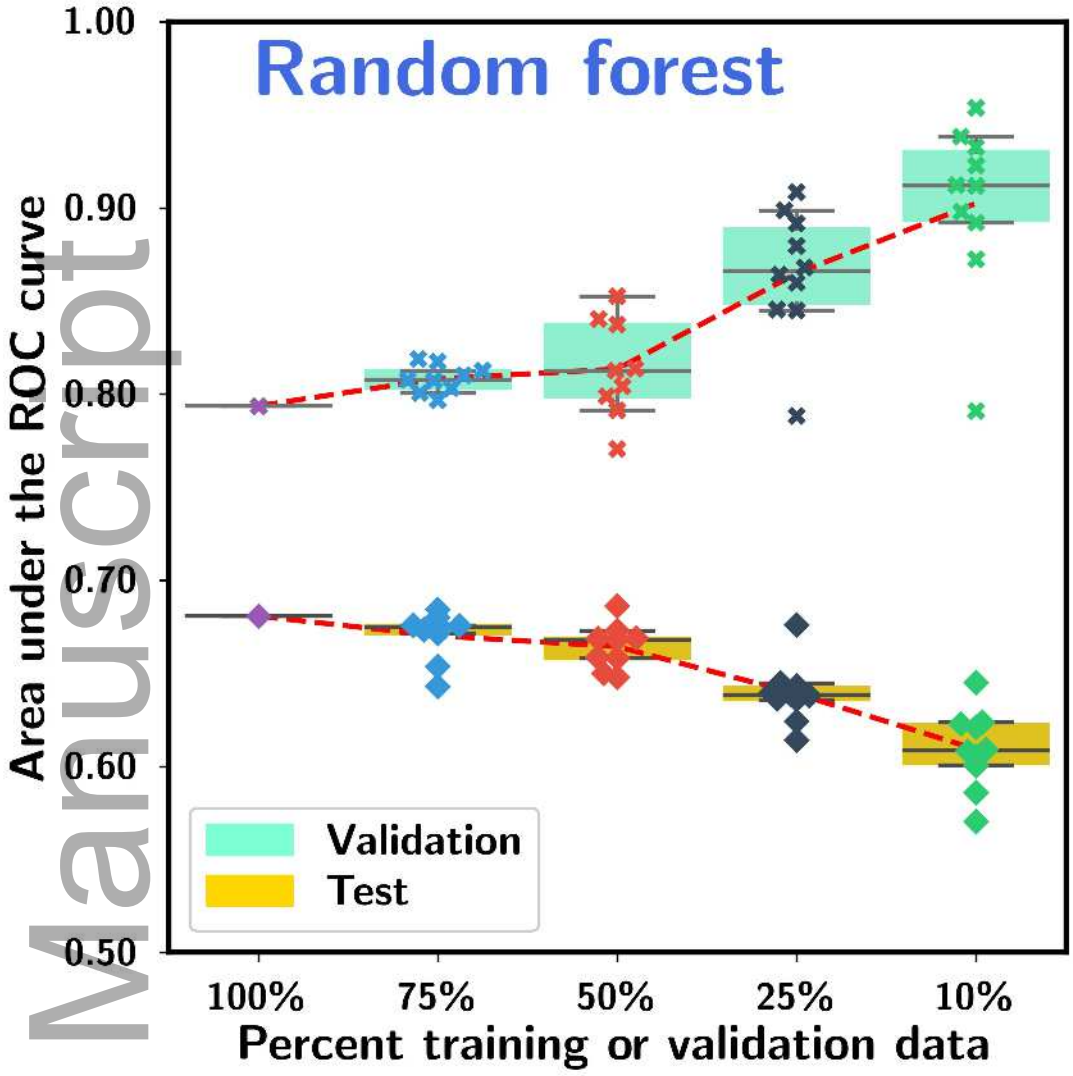
mp_14678_f6__top_left.jpg



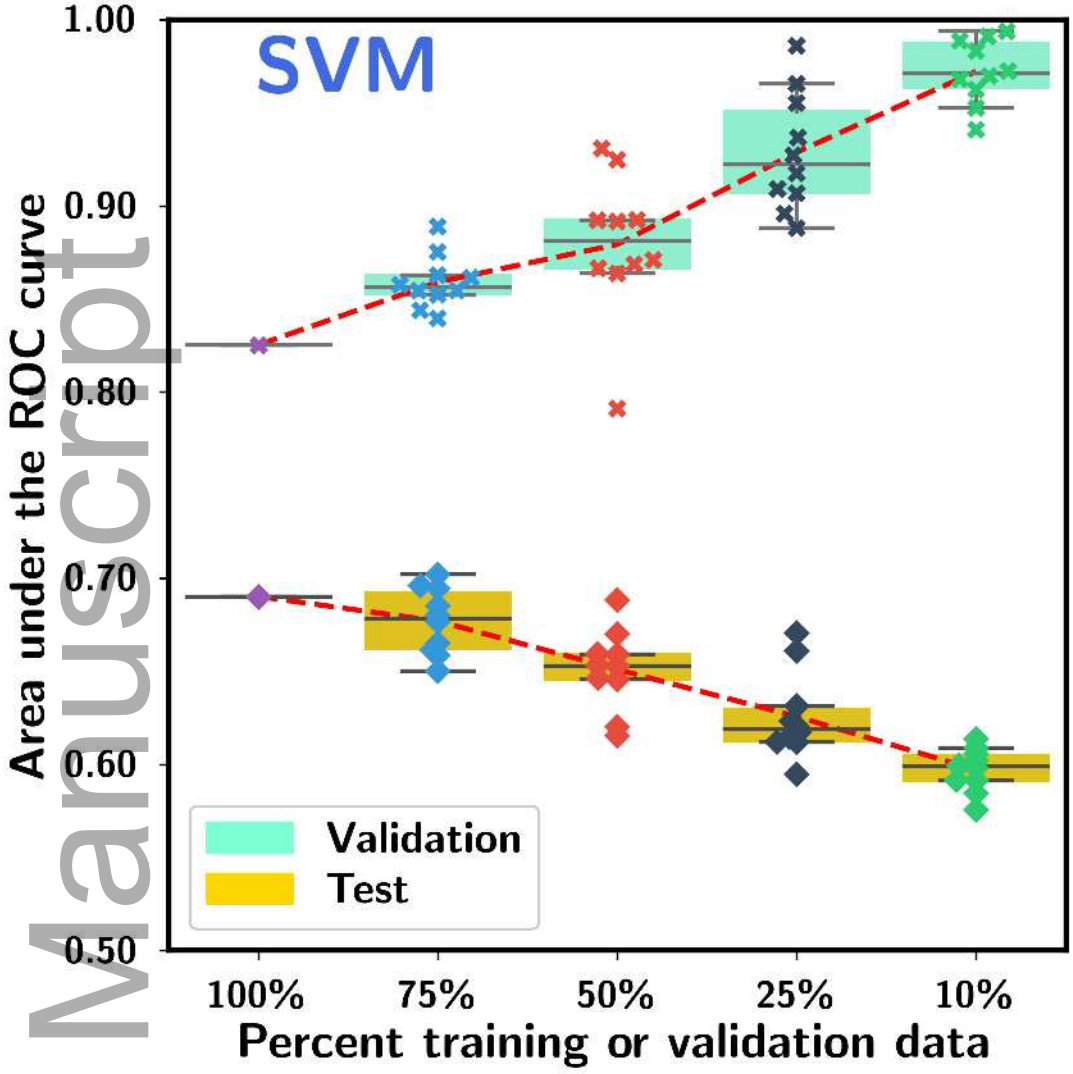
mp_14678_f6__top_middle.jpg



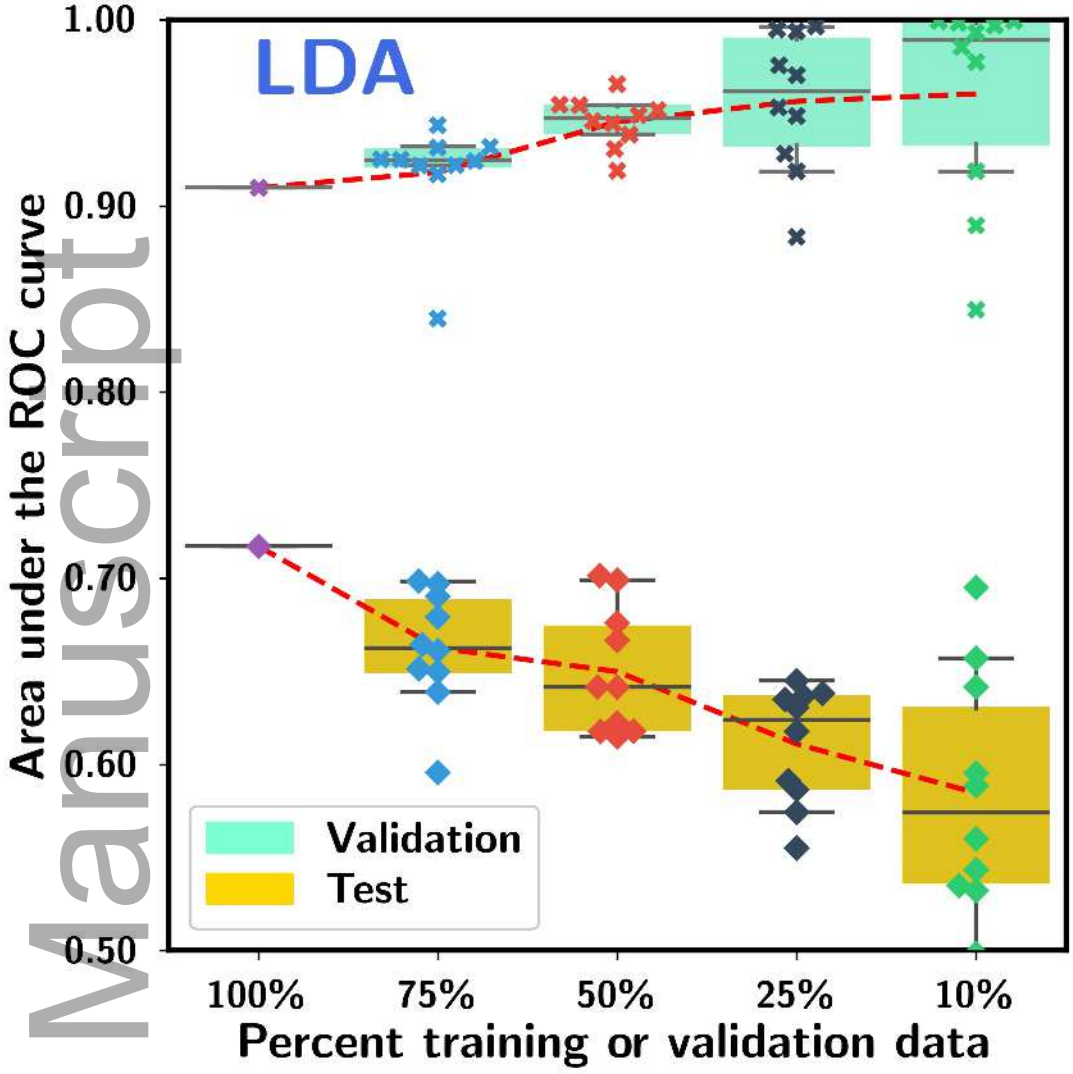
mp_14678_f6__top_right.jpg



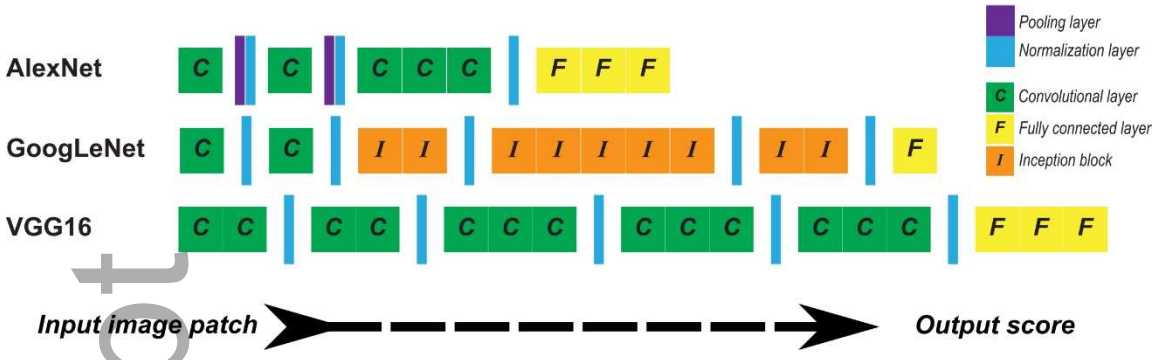
mp_14678_f7__bottom.jpg



mp_14678_f7__middle.jpg

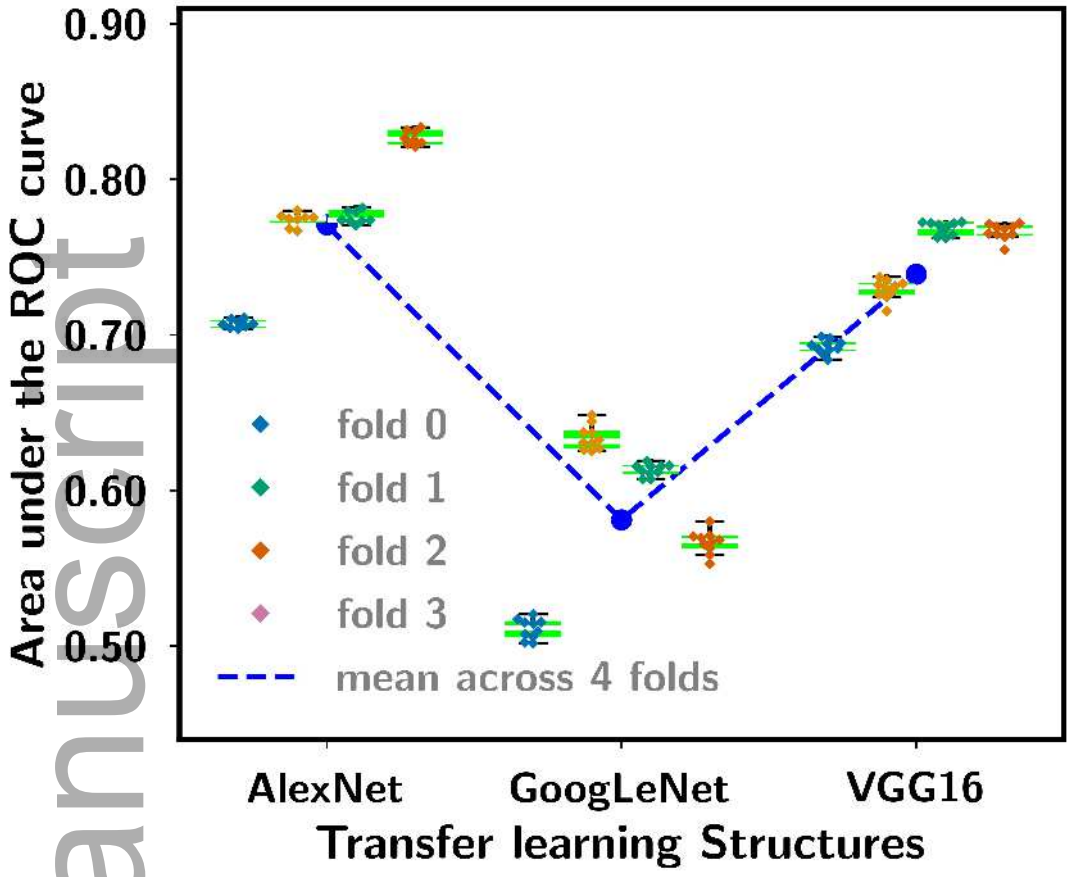


mp_14678_f7__top.jpg

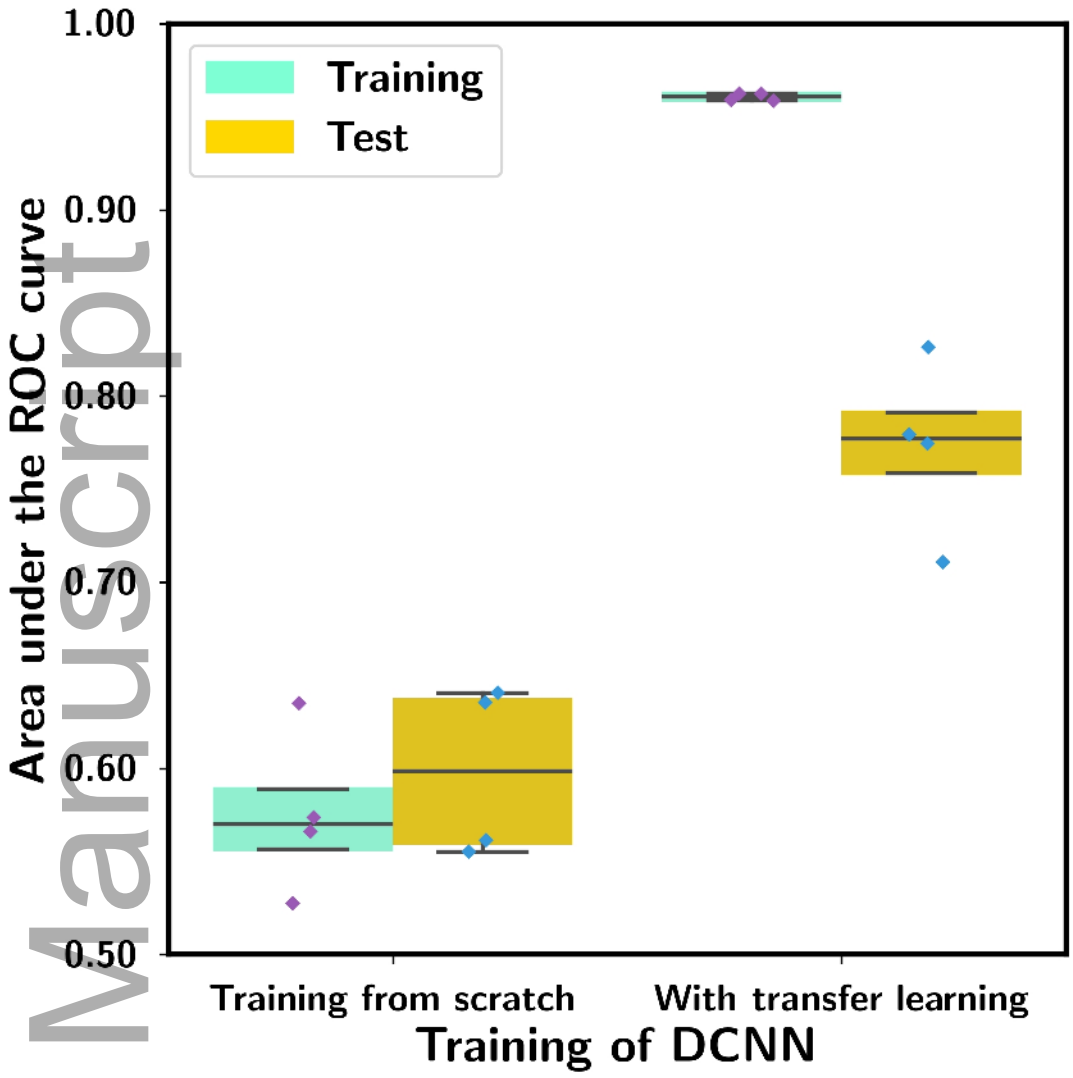


mp_14678_f1.jpg

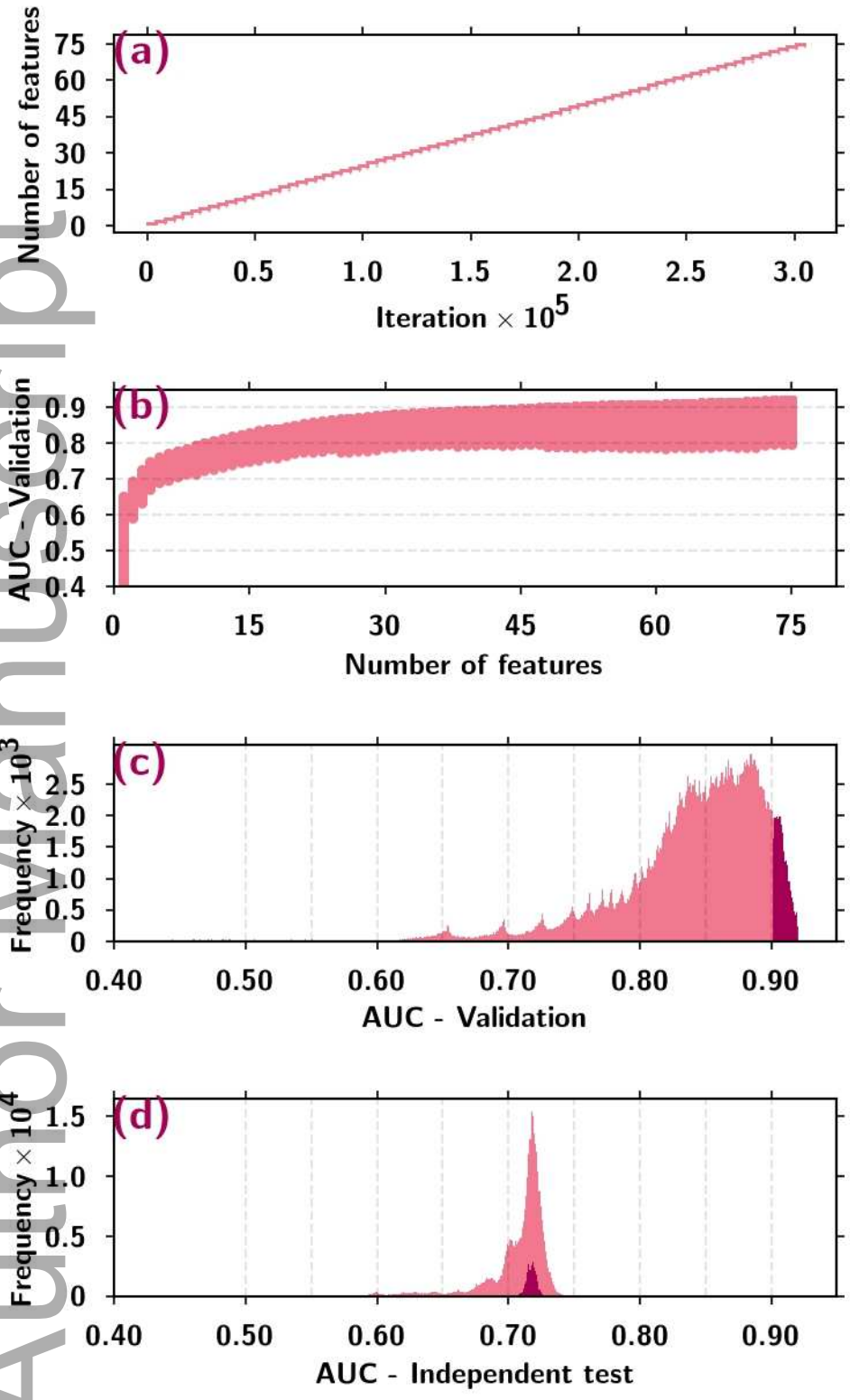
mp_14678_f2.jpg



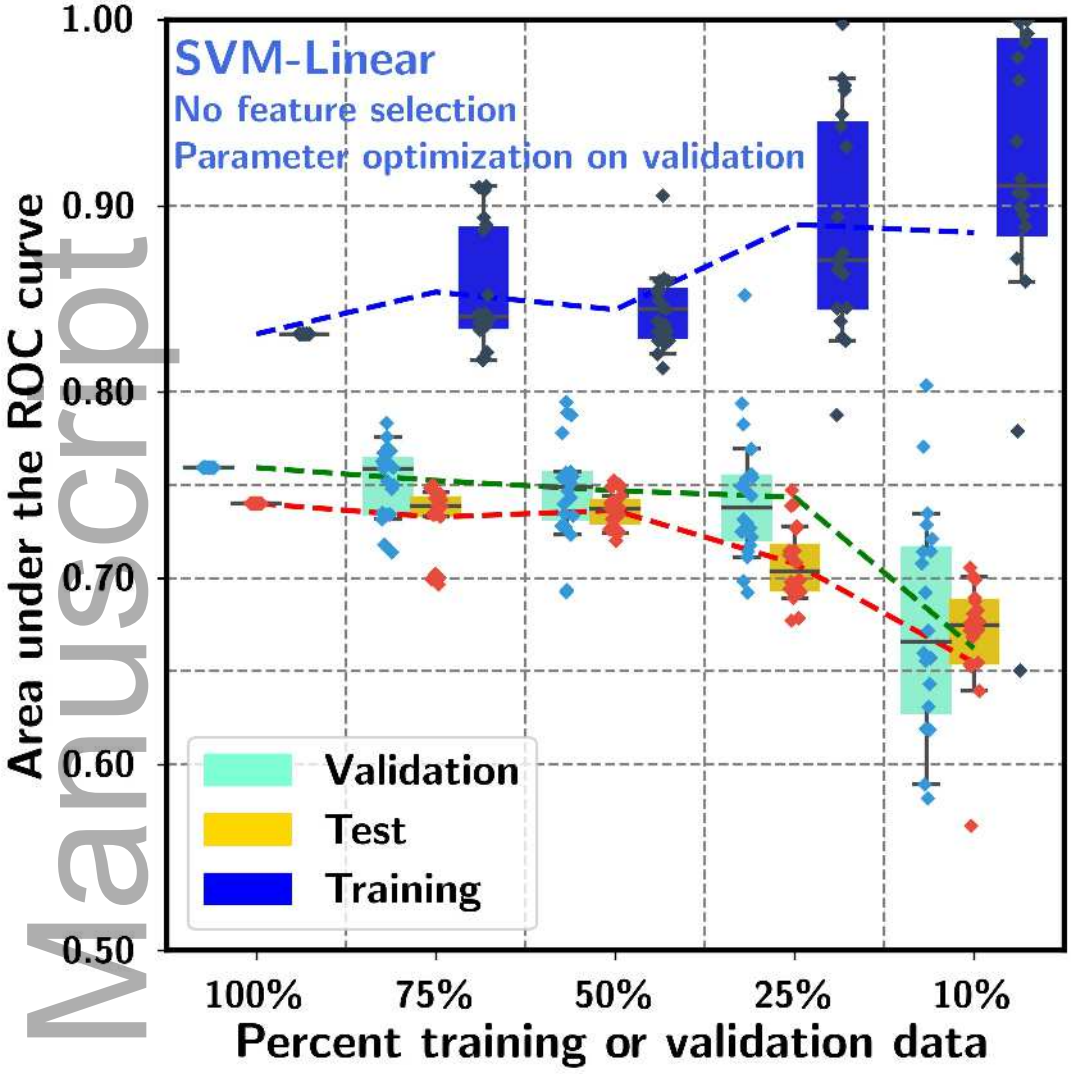
mp_14678_f3.jpg



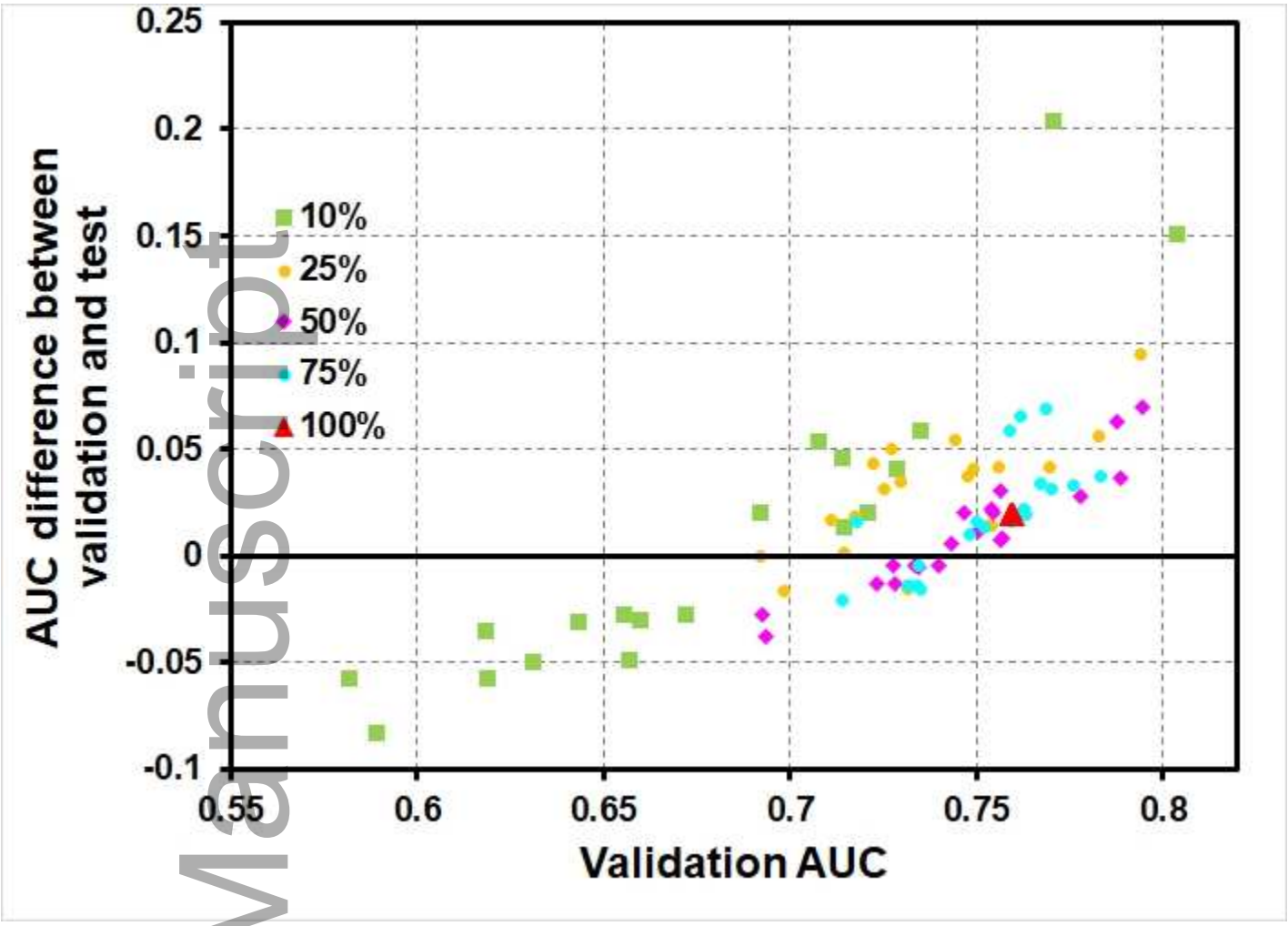
mp_14678_f4.jpg



mp_14678_f5.jpg



mp_14678_f8a.jpg



mp_14678_f8b.jpg