# ACTIVE LEARNING TO EXTEND TRAINING DATA FOR LARGE AREA AIRBORNE LIDAR CLASSIFICATION

Nan Li [1,2,] *, Norbert Pfeifer [1]

[1] Department of Geodesy and Geoinformation, Technische Universität Wien, 1040 Vienna, Austria - (nan.li, norbert.pfeifer)@geo.tuwien.ac.at
[2] College of Survey and Geoinformation, Tongji University, 200092 Shanghai, China - linan123@tongji.edu.cn

**KEY WORDS:** active learning, semi-supervised classification, training data selection

**ABSTRACT:**

Training dataset generation is a difficult and expensive task for LiDAR point classification, especially in the case of large area classification. We present a method to automatically extent a small set of training data by label propagation processing. The class labels could be correctly extended to their optimal neighbourhood, and the most informative points are selected and added into the training set. With the final extended training dataset, the overall (OA) classification could be increased by about 2%. We also show that this approach is stable regardless of the number of initial training points, and achieve better improvements especially stating with an extremely small initial training set.

## 1. INTRODUCTION

LiDAR (Light Detection And Ranging) automatic classification has been an important study topic over years. Supervised statistical approaches, such as Support Vector Machines (SVM) (Secord and Zakhor, 2007) or Random Forest (Guo et al., 2011) have been widely applied and achieved good performance. Additionally, to incorporate the spatial contextual information, Markov Random Field (MRF) and Conditional Random Field (CRF) are successfully used for contextual classification and achieve smoother results than the classifications based on individual independent features (Niemeyer et al., 2014; Shapovalov et al., 2010). This research mostly focuses on site-specific classification for 3D points at a small scale. Only few papers were published on large area LiDAR classification.

Extensive 3D point clouds over large area would result in handcrafted features inhomogeneity, making automated points cloud classification difficult. This would bring further challenges for class separability when only small training data is available. Especially, supervised classifiers rely on the quality of the labeled training data. The training samples should be fully representatives of the class-type statistics to allow the classifier to find the correct solution. In the case of large area classification, this constraint makes the generation of an appropriate training set a difficult and expensive task that requires extensive manual interaction. This is a common problem for classification of large amounts of data, and only a small amount of reference points can be manual labelled due to the limited economical and temporal resources. Therefore, the classification model constructed on the collected small training data could show poor generalization capabilities when applied to the rest of large amount of data. Additionally, manual training set definition is usually done by visual inspection of the scene and the successive labeling of each sample. This phase is highly redundant as well as time-consuming.

A solution to the problem of training data extraction is represented by semi-automatic active learning methods. Its key idea is to select the samples whose inclusion in the training set would be beneficial to the classification performance. And the semi-automatic active learning already has shown to be effective for hyperspectral image classification. For instance, a combination of the SVM classifier is commonly used (Mitra et al., 2004; Tan et al., 2014), samples that are close to the hype-plane are selected into the training dataset. In order to be adaptive with any generative classifier, the maximum information gain (Rajan et al., 2008) and breaking tie (Luo et al., 2005) can also be used to select uncertain samples. A co-training approach proposed by (Romaszewski et al., 2016) scored samples by combining spatial and spectral features, an optimal training set would be learned by iteratively adding new samples with high scores.

In this paper, we aim to extend a small set of initial labelled samples during a process of label propagation. By adapting an optimal neighborhood selection, the knowledge about class labels from the training set can be correctly extended to their neighborhood. And one most informative point is selected by BT (breaking tie) and added into the training set. In this way, we extent the training dataset, and automatically label the newly added samples. Compared with original small training set, the new extended training set could be more representative for features and capable to improve the classification results.

The rest of the paper is organized as follows: Section 2 explains our method. Section 3 presents the experiment on real data and its results, while Section 4 describes the performance along iteration and the impact of the number of initial training points. Summaries are provided in Section 5.

## 2. METHODOLOGY

Normally, the active learning approach consists of two components. The first is the selection of the most useful unlabelled samples to the classifier, and the second is how to determine the class labels of these new selected samples. In this paper, we start with a small set of suboptimal training points. The

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-1033-2019 | © Authors 2019. CC BY 4.0 License.

breaking ties (BT) method (Luo et al., 2005) is applied to sample the informative unlabelled points. And the label of selected unlabelled point is determined by the spatial similarity. Since the class label is highly correlated with spatial similarity of points, we could assume that points located in the same neighbourhood are likely to have the same label with the center point. After adding those informative samples to the training dataset, the classification model is forced to focus on conflicting areas and to improve its generalization capabilities. The processing sequence is as follows:

Step 1: Based on the initial small training dataset, an initial classifier is built;
Step 2: For each training point, finding its' optimal neighbouring points;
Step 3: The classifier is applied to those neighbouring points, and one most informative point is selected by the minimal BT value and labelled by the current training point.
Step 4: Extending training dataset by adding new samples, and updating the classifier;
Step 5: Repeating step 2,3,4, until a maximum iteration number is met. Then, the final training set is used to refine the classifier;
Step 6: Finally, the classifier is used to predict labels for all unlabelled points.

The following section 2.1 describe the estimation of the optimal neighborhood, and section 2.2 induces the breaking ties

### 2.1 Label propagation by the optimal neighbourhood

By taking the advantage of spatial correlation of point cloud, the knowledge about class labels of training points can be extended to their neighborhood. To guarantee the accuracy of label propagation, an optimal neighborhood estimation method is applied (Li et al., 2019). The neighboring points are adaptively selected by weighted geometric similarity, so that all neighboring points that potentially belong to the same object with the concerned points could be included.

Here, the geometric similarity is measured by the angle between the normal vectors and point-to-plane orthogonal distances, while the weights are determined by the local surface variations (σ). To avoid lacking enough neighboring points for non-planar points, like vegetation, we assign larger weights to those non-planar points to increase the geometric similarity with neighbors. The weight function is defined in Eq. (1), and neighboring points that satisfy Eq. (2) are collected as the optimal neighbors of the concerned point:

$$Weight(p_0,p_i) = \begin{cases} 1 & if\ \sigma(p_0) \leq THR_\sigma \\ e^{\sigma(p_0)} \cdot e^{\sigma(p_i)} & else\ \sigma(p_0) > THR_\sigma \end{cases} \quad (1)$$

$$\begin{gathered} Weight(p_0,p_i) \cdot nv_{p_0} \cdot nv_{p_i} \geq \cos(THR_\alpha) \\ and\ Weight(p_0,p_i) \cdot \left|(p_0 - p_i) \cdot nv_{p_0}\right| \leq THR_d \end{gathered} \quad (2)$$

Where $\sigma(p_0)$ is the local surface variation in the point $p_0$. $THR_\sigma$ is a threshold to determine whether $p_0$ may belong to a planar object. $nv_p$ denotes the normal vector of point $p$ and $p = [x_p, y_p, z_p]$ denotes the 3D coordinates of point $p$. $THR_\alpha$ is the threshold of the normal vector-angle change and $THR_d$ is the threshold of the local point-to-plane orthogonal distance.

### 2.2 Sample selection by BT

The BT technique is focused on the diversity of the unlabeled samples, which is obtained by the minimum difference between the two highest posterior class probabilities. The more a point shows a similar posterior probability between the two most probable classes, the more it is uncertain and thus capable of providing useful information if added to the training dataset (Tuia et al., 2011). Thus the BT value of point $p_i$ is formed by Eq.(3):

$$\mathrm{BT}(p_i) = \max_{c \in C}\big(P(l_i = c|p_i)\big) - \max_{c \in C \backslash c^+}\big(P(l_i = c|p_i)\big) \quad (3)$$

Where $P(l_i = c|p_i)$ probability for class prediction $l_i$ of a point $p_i$, $c \in C$ corresponds to one class $c$ among the $C$ possible classes, and $c^+ = \max_{c \in C}\left(P\left(l_i = c\middle|p_i\right)\right)$ is the most probable class for point $p_i$.

After finding all optimal neighboring points for one training point, the point minimizing Eq.(3) is then taken and labeled by the current, certain training point. The procedure is implemented for all training points and repeated for several times, the final selected labeled training points are used to refine the classifier.

## 3. RESULTS

### 3.1 Datasets

The point cloud we used was a fully labelled airborne LiDAR dataset of Vienna, Austria. The selected area is 1270×200 m$^2$, and the average density is about 50 points/m$^2$. This area represents a complex urban scene, including a mixture of high and low vegetation, high-rise and small detached houses, and flat and sloped ground. Five domain classes were categorized for the Vienna dataset: *ground, vegetation, roofs, façades* and *others* that include fences, cars, street lights, power lines and so on.
To get an impression of the dataset, the percentage distribution of each class in the dataset are shown in Table. 1.

| Class | Percentage |
|---|---|
| Ground | 53.70% |
| Vegetation | 26.72% |
| Roofs | 14.04% |
| Facades | 1.54% |
| Others | 4.00% |

Table 1. Percentage distribution of each class in the dataset

### 3.2 Experiment setup and results

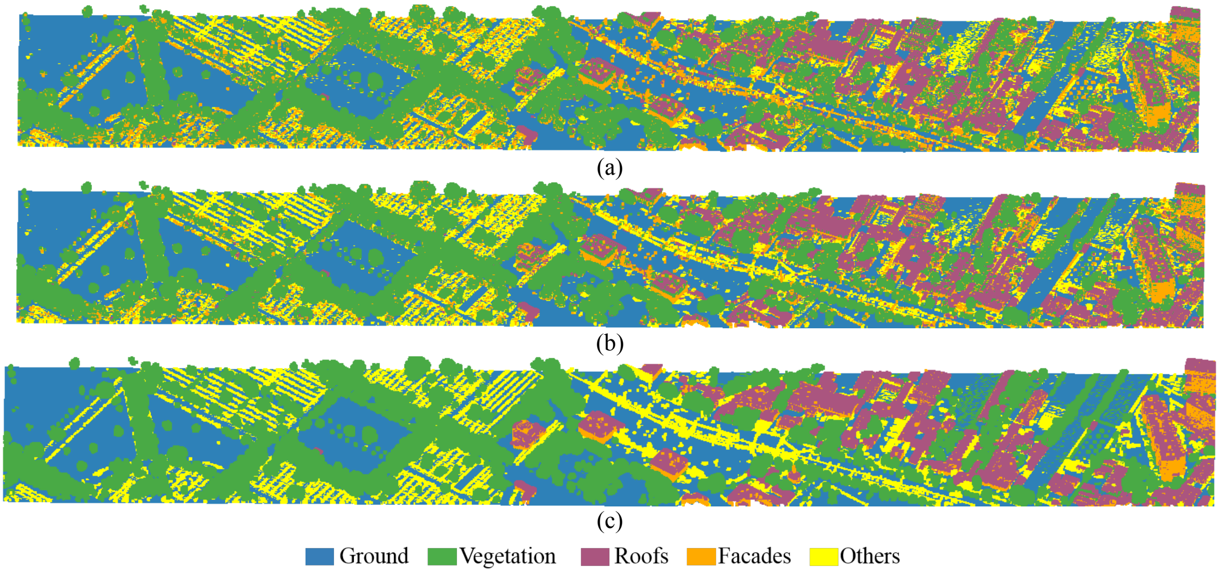
We used the random forest (RF) as the probabilistic classifier. For the optimal neighbourhood estimation, the spherical neighbourhood with radius of 2m was used for initial neighbouring points searching. Then the optimal neighbouring samples are selected by weighted geometric similarity and labelled by its neighbouring labelled training point. The initial number of training points is 100 per class, and the iteration was empirically set as 3 to extend the training data.

Figure. 1 shows the classification results using initial training dataset, extended training dataset and the reference dataset. From visual inspection, a more smooth classification result is achieved after training dataset extension. For instance, as shown in the marked area A in Figure. 2, more points are correctly classified as ground after the active learning, whereas those points are wrongly labelled as others in the initial training dataset. Another notable change appears in the marked area B in Figure. 2. There is a large amount of points misclassified as vegetation by the initial classification. After the active learning, most of those points' labels are changed into façades, this situation could be explained by a small error in the reference data (seen in the Figure. 2(c)).

From the Table. 2, the OA (overall) accuracy was increased by 1.7% after the active learning of training points. A relatively significant improvement was achieved for the class of vegetation and others, 4.00% and 3.37%, respectively. However, the accuracy of façade has dropped to 74.14% using the extended training dataset. 13.36% façade points are misclassified as vegetation which is 6.31% higher than the initial classification, and a few of façade points (2.41%) are misclassified as roofs since they are easily mixed up over the conjunctions of roofs and façades.

| Class | Initial accuracy | Accuracy after active learning |
|---|---|---|
| OA | 84.24% | 85.98% |
| Ground | 88.60% | 89.35% |
| Vegetation | 79.74% | 83.73% |
| Roofs | 81.82% | 84.85% |
| Façades | 84.59% | 74.14% |
| Others | 64.06% | 67.37% |

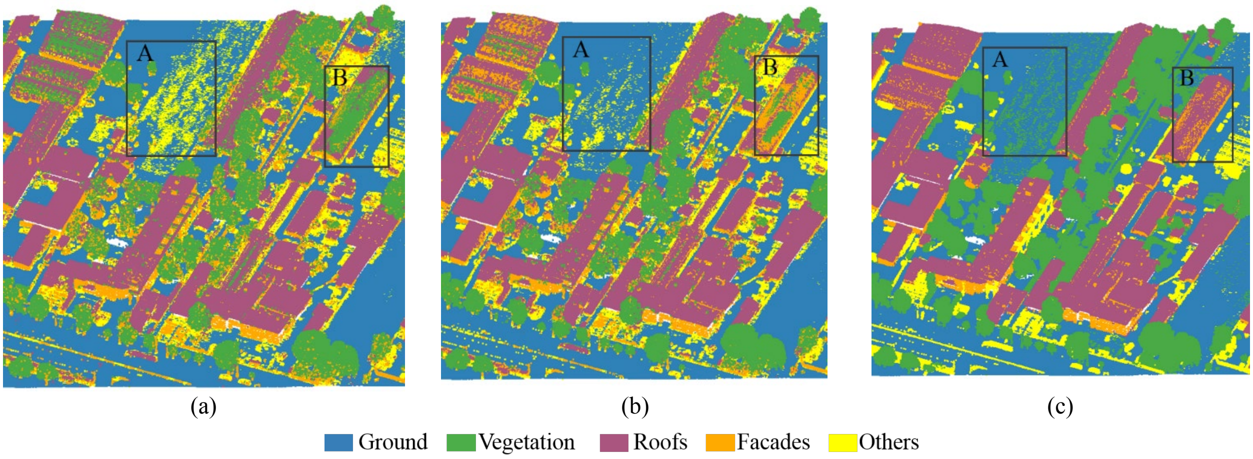
Table 2. Accuracy comparison of classification using initial training set and extended training set


(a)

(b)

(c)

Figure 1. The classification results. (a) using the initial training dataset; (b) using the extended training dataset; (c) the reference labelled data.


(a)

(b)

(c)

Figure 2. Detailed comparison of classification results. (a) using the initial training dataset; (b) using the extended training dataset; (c) the reference labelled data.

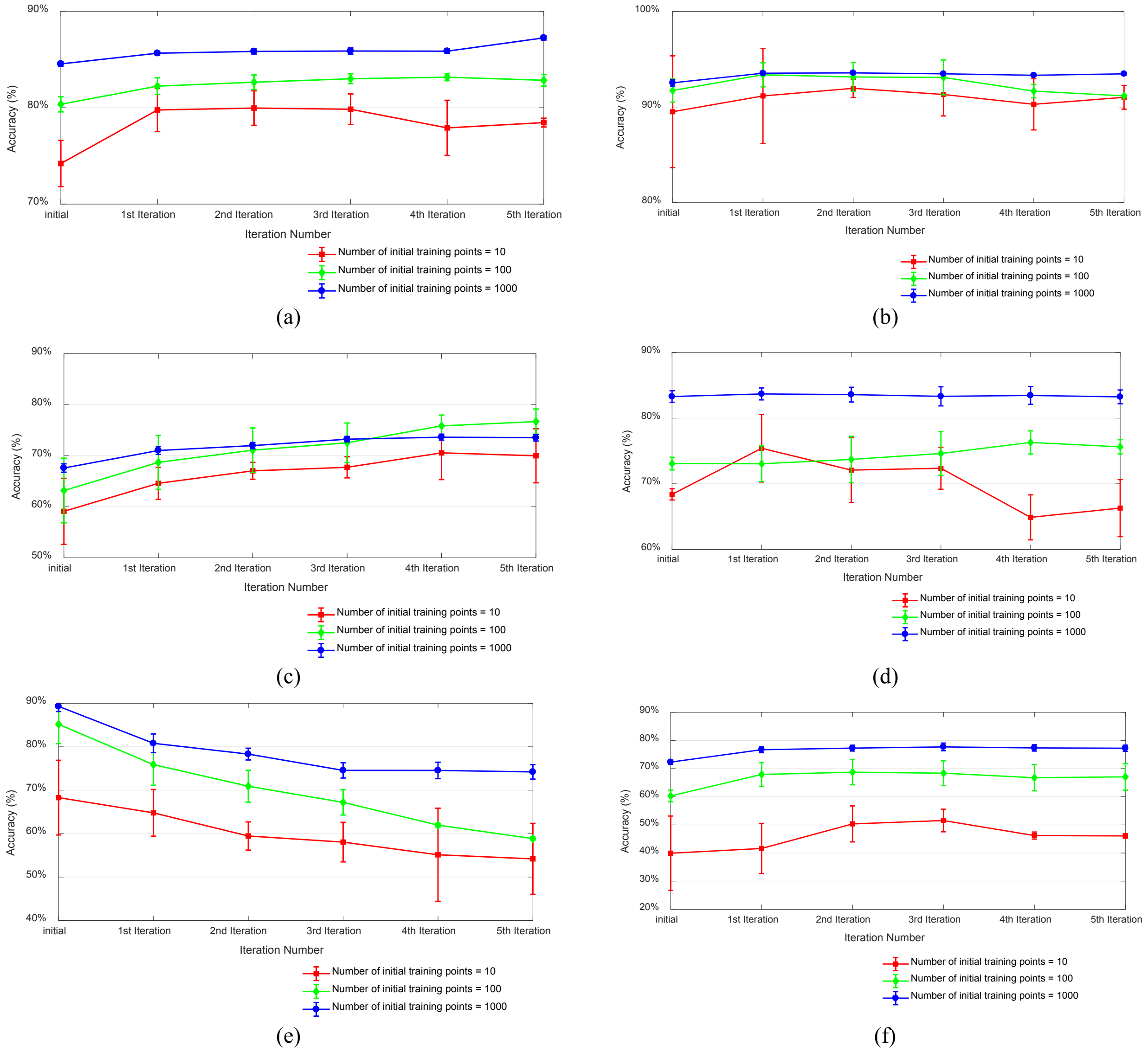
## 4. DISCUSSIONS

To access the stability of this active training data learning method, we started with different amounts of training dataset, which includes 10,100 and 1000 initial training points per class, respectively. Each experiment was repeated 3 times. The accuracy changes along the iterations are shown in Figure. 3, which are the average accuracy and its standard deviation over 3 experiments.

Compared to the initial classification results, the OA accuracies were all increased after the active learning (seen in Figure. 3(a)). Notably, the significant overall accuracy improvement was achieved by the smallest set of initial training data of 10 samples per class. It gained 5% higher OA accuracy than initial classification, while 2.4% and 1.5% OA increase for initial training points of 100 and 1000 per class, respectively. The representativeness of the extremely small training set is usually lacking strongly, thus the effect of adding new informative samples would be notable when it was started with a poor initial classification result. While the models trained by 100 and 1000 samples per class are already decent, the improvement would become moderate when the amount of the initial training set is raised. Also due to the incompleteness of small initial training set of 10, the variation is relatively larger than the other two initial training sets.

We also observed that the accuracy would be immediately improved by extending the training data in the 1st iteration, and the accuracy only has slight changes over iterations besides the accuracy of façade. It means that the samples that are selected during the first extension are the most informative and could be effective to increase classification ability, whereas other samples from the rest of iterations may have very similar feature vectors with samples that already exist in the training set. Therefore, they could not provide more useful information to achieve better accuracy. This is caused by the local neighbourhood we used for label propagation. However, the trend of accuracy change of façade is different from the others. Façade points tend to be misclassified into vegetation during this active learning procedure. Since generally the optimal neighbourhood favours points that are located in the same plane, vegetation points that lie in the same vertical plane would have similar feature vectors with the vertical façade points. Iteratively including those vegetation points into training dataset would lead to the confusion with façade.

Another interesting finding is that the performance of active learning would be fundamentally impacted by the initial amount of training samples. The accuracy with 100 initial training points per class reached 84.5% after 6 iterations, meanwhile the total number of training points per class is 1600. This result is still not as good as the initial classification accuracy using random 1000 training samples at first. But it is comparable to the classification by initial training samples of 300 per class (84.04%).



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3. The trend of accuracy changes over iterations. (a) overall accuracy; (b) ground accuracy; (c) vegetation accuracy; (d) roofs accuracy; (e) façades accuracy; (f) others accuracy.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-W13-1033-2019 | © Authors 2019. CC BY 4.0 License.

## 5. SUMMARY

We proposed an effective active learning method to automatically extend training points. Classification accuracy was increased by using the extended training dataset, which was significant especially starting with an extremely small set of 10 labelled points per class. An optimal training dataset would be achieved by a few of iterations. The reasonable amount of training samples also keep the classifier learning efficient. Due to the limitation of initial training sample, an exploration for initial samples selection will be considered in the further research.

## ACKNOWLEDGEMENTS

The Vienna airborne LiDAR dataset was provided by the ‘Stadtvermessung Wien’, the Magistrate of the City of Vienna.

## REFERENCES

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. *ISPRS J Photogramm Remote Sens* 66, 56-66.

Li, N., Liu, C., Pfeifer, N., 2019. Improving LiDAR classification accuracy by contextual label smoothing in post-processing. *ISPRS J Photogramm Remote Sens* 148, 13-31.

Luo, T., Kramer, K., Goldgof, D.B., Hall, L.O., Samson, S., Remsen, A., Hopkins, T., 2005. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research* 6, 589-613.

Mitra, P., Shankar, B.U., Pal, S.K., 2004. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern recognition letters* 25, 1067-1074.

Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. *ISPRS J Photogramm Remote Sens* 87, 152-165.

Rajan, S., Ghosh, J., Crawford, M.M., 2008. An active learning approach to hyperspectral data classification. *IEEE Trans Geosci Remote Sens* 46, 1231-1242.

Romaszewski, M., Głomb, P., Cholewa, M., 2016. Semi-supervised hyperspectral classification from a small number of training samples using a co-training approach. *ISPRS J Photogramm Remote Sens* 121, 60-76.

Secord, J., Zakhor, A., 2007. Tree detection in urban regions using aerial lidar and image data. *IEEE Geosci. Remote Sens*. Lett 4, 196-200.

Shapovalov, R., Velizhev, E., Barinova, O., 2010. Nonassociative markov networks for 3d point cloud classification. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XXXVIII, Part 3A. Citeseer.

Tan, K., Li, E., Du, Q., Du, P., 2014. An efficient semi-supervised classification approach for hyperspectral imagery. *ISPRS J Photogramm Remote Sens* 97, 36-45.

Tuia, D., Pasolli, E., Emery, W.J., 2011. Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment* 115, 2232-2242.