Published in final edited form as:

Inf Process Med Imaging. 2011 ; 22: 85–96.

Characterizing Spatially Varying Performance to Improve Multi-Atlas Multi-Label Segmentation

Andrew J. Asman¹ and Bennett A. Landman^{1,2,3}

Andrew J. Asman: andrew.j.asman@vanderbilt.edu; Bennett A. Landman: bennett.landman@vanderbilt.edu

¹ Electrical Engineering, Vanderbilt University, Nashville, TN, USA 37235

² Biomedical Engineering, Vanderbilt University, Nashville, TN, USA 37235

³ Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA 21218

Abstract

Segmentation of medical images has become critical to building understanding of biological structure-functional relationships. Atlas registration and label transfer provide a fully-automated approach for deriving segmentations given atlas training data. When multiple atlases are used, statistical label fusion techniques have been shown to dramatically improve segmentation accuracy. However, these techniques have had limited success with complex structures and atlases with varying similarity to the target data. Previous approaches have parameterized raters by a single confusion matrix, so that spatially varying performance for a single rater is neglected. Herein, we reformulate the statistical fusion model to describe raters by regional confusion matrices so that co-registered atlas labels can be fused in an optimal, spatially varying manner, which leads to an improved label fusion estimation with heterogeneous atlases. The advantages of this approach are characterized in a simulation and an empirical whole-brain labeling task.

Keywords

Simultaneous truth and performance level estimation (STAPLE); Statistical fusion; Classifier fusion; Rater performance; Automated segmentation

1 Introduction

Knowledge of the connections and relationships between biological structures and function is essential to scientific and clinical interpretation of medical images. Segmentation plays a pivotal role in building understanding of these relationships as it enables association of quantitative and functional observations with structural counterparts. Fully automated segmentation methods are important tools for achieving robust, high-throughput analysis, yet imaging and anatomical variability render these challenging objectives. The gold standard approach is to have a human expert manually label each image; yet manual segmentation (labeling) suffers from intra- and inter-rater reliability concerns. It is natural to have multiple experts (i.e., raters) perform the segmentation task so that the output of each individual can be combined (“fused”) to form an estimate of the “ground truth.” Manual labeling can be extraordinarily time consuming and cost prohibitive, so alternative approaches have been developed to efficiently derive multiple possible segmentations. In both manual and automated cases, a central challenge becomes finding the optimal method to fuse the possible segmentations so that a consistent, reliable and accurate estimate of the true segmentation can be obtained.

A practical and simple approach to fusion is using majority vote, where the “ground truth” estimate is obtained by declaring the label for each voxel that was reported most frequently [1, 2]. This approach, however, 1) does not guarantee a unique majority label, 2) does not provide information about the likelihood of the estimated segmentation and 3) does not provide any information about rater performance. The Simultaneous Truth and Performance Level Estimation approaches (aka, STAPLE) were recently presented to optimally estimate observations based on rater reliability [3, 4]. When raters are collectively unbiased and independent, this algorithm increases the accuracy of a single labeling by probabilistically fusing multiple less accurate delineations, e.g., following the general theory of statistical boosting.

Multi-atlas segmentation through non-rigid registration and label transfer represents a fully automated method for performing segmentation through label fusion based on existing label sets (i.e., the atlases). Using a database of manually segmented atlases, this segmentation technique has been shown to be reasonably robust and accurate. As with the manual labeling method, the problem becomes determining the optimal method to combine the observed segmentations. Statistical fusion techniques have been widely used in multi-atlas labeling [5-8]. However, when applied to multiple, intricate labels, voting techniques have been shown to dramatically outperform statistical fusion techniques [5, 9, 10]. Additionally, applications have largely focused on specific anatomical features with relatively few labels [7, 8].

Optimality of the statistical fusion approaches hinges on the validity of the labeling process model (e.g., the stochastic model of how a rater makes a mistake). Existing STAPLE approaches have used spatially invariant models — the probability distribution of error does not change voxel-by-voxel. Intuitively (and empirically – see Figure 1), there should exist certain regions where the registration (and label transfer) is more accurate than in other regions. Furthermore, the pattern of spatial agreement will vary on an atlas-by-atlas basis. Traditionally, single measures of rater performance that describe the quality of the registered atlases are an amalgamation of many sub-regions of varying accuracy. In the context of voting, regional and intensity based fusion techniques have emerged [7, 9]. However, other than simply ignoring consensus voxels [4], spatially varying characteristic have not been derived for the statistical fusion approach. Herein, we generalize the statistical fusion algorithm (STAPLE) to seamlessly account for spatially varying performance (Spatial STAPLE) by inclusion of regional performance level estimations (confusion matrices).

This manuscript is organized as follows. Section II details the extension of traditional STAPLE theory to include spatially varying performance levels. Section III illustrates the advantages of Spatial STAPLE on a simulation example and in an empirical task of whole-brain labeling. Finally, Section IV provides brief concluding remarks.

2 Theory

The following presentation of theory for Spatial STAPLE closely follows the approach of Warfield, et al [3].

2.1 Problem Definition

Consider an image of N voxels with the task of determining the correct label for each voxel in that image. Consider a collection of R raters (registered atlases) that provide an observed delineation of all N voxels exactly once. The set of labels, $\mathbf{L}$, represents the set of possible values that a rater can assign to all N voxels. Let $\mathbf{D}$ be an $N \times R$ matrix describing the labeling decisions of all R raters at all N voxels where $D_{ij} \in \{0, 1, \dots, L - 1\}$. Let $\mathbf{T}$ be a

vector of N elements that represents the hidden true segmentation for all voxels, where $T_i \in \{0, 1, \dots, L-1\}$.

A characterization of the R raters performance at each voxel is characterized by θ . Each element of θ , θ_{jm} , describes an $L \times L$ confusion matrix and is defined for rater j and region B_m , where B_m is a vector that indicates the voxels over which θ_{jm} is defined. Let M be the total number of regions, and let the union of all regions equal the full volume, (i.e. $\cup_{m=1}^M B_m = \{1, \dots, N\}$) and all regions are mutually disjoint (i.e. $B_i \cap B_j = \emptyset, \forall i \neq j$). Let the complete data be $(\mathbf{D}, \mathbf{T})$ and let the probability mass function of the complete data be $f(\mathbf{D}, \mathbf{T}|\theta)$.

Figure 2 illustrates the regional confusion matrix approach where an observation is divided into quadrants. Each quadrant is described by a different confusion matrix. One of the observed quadrants is of significantly higher quality than the other quadrants. Traditional STAPLE would fail to recognize this phenomenon, while Spatial STAPLE is capable of detecting this regional quality variation.

2.2 Spatial STAPLE Algorithm

The goal of the Spatial STAPLE algorithm is to accurately estimate the true segmentation using the R raters segmentation decisions and the current estimate of the rater performance level parameters. The estimated performance level parameters will be selected such that they maximize the complete data log likelihood function

$$\hat{\theta} = \arg \max_{\theta} \ln f(\mathbf{D}, \mathbf{T}|\theta). \quad (1)$$

It is assumed that the segmentation decisions are all conditionally independent given the true segmentation and the performance level parameters, that is $(D_{ij}|T_i\theta_{jm}) \perp (D_{ij'}|T_i\theta_{j'm}) \forall j \neq j'$. This model expresses the assumption that the raters derive their segmentations of the same image independently from one another and that the quality of the segmentation by each rater is captured by the estimation of the performance level parameters.

Our solution to the expectation maximization (E-M) algorithm used to solve (1) is now presented. The complete data used to solve this E-M algorithm is the observed data, $\mathbf{D}$, and the true segmentation, $\mathbf{T}$. The true segmentation is regarded as the missing or hidden data, and is unobservable. Let θ_{jm} be the covariance, or confusion matrix associated with rater j and region B_m and let

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1R} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2R} \\ \dots & \dots & \dots & \dots \\ \theta_{M1} & \theta_{M2} & \dots & \theta_{MR} \end{bmatrix} \quad (2)$$

be the complete set of unknown parameters for the R segmentations and the M disjoint subsets of the full set of voxels.

Following the notation proposed by Warfield, et al [3], the estimation of the true segmentation (E-step) is represented using common E-M notation. Let $\mathbf{W}$ be an $L \times N$ matrix, where each element W_{si} represents the probability that the true label associated with voxel i is label s . The solution for $\mathbf{W}$ on iteration k is

$$\begin{aligned}
W_{si}^{(k)} &\equiv f(T_i=s|D_i, \theta_{jm}^{(k-1)}) \\
&= \frac{f(T_i=s) \prod_j f(D_{ij}=n|T_i=s, \theta_{jm}^{(k-1)})}{\sum_{s'} f(T_i=s') \prod_j f(D_{ij}=n|T_i=s', \theta_{jm}^{(k-1)})} \\
&= \frac{f(T_i=s) \prod_j \theta_{jms}^{(k-1)}}{\sum_{s'} f(T_i=s') \prod_j \theta_{jms'}^{(k-1)}}
\end{aligned} \tag{3}$$

where the value of m is selected such that $i \in B_m$ and $\theta_{jms}^{(k-1)}$ represents the probability that rater j observed label n given that the true label is s in region B_m . The prior probability, $f(T_i=s)$, can be either a global prior or a voxelwise prior. In the case of this paper $f(T_i=s)$ is a global prior that represents an *a priori* estimate of the fraction of voxels in the true segmentation that have label value s .

The estimate of the performance level parameters (M-step) is presented in (4). The main difference between the representation seen in (4) and the traditional STAPLE representation [3] is the fact that θ is an $R \times M$ set of confusion matrices, where each rater has M confusion matrices associated with it, each defined over a region space that is a subset of the full voxel set.

$$\begin{aligned}
\theta_{jm}^{(k)} &= \arg \max_{\theta_{jm}} \sum_{i \in B_m} E[\ln f(D_{ij}|T_i, \theta_{jm}) | D, \theta_{jm}^{(k-1)}] \\
&= \arg \max_{\theta_{jm}} \sum_{i \in B_m} \sum_s W_{si}^{(k)} \ln f(D_{ij}|T_i=s, \theta_{jm}).
\end{aligned} \tag{4}$$

Using the estimate of $\theta_{jm}^{(k-1)}$ for $f(D_{ij}|T_i=s, \theta_{jm})$ and the constraint that each column of the confusion matrix must be normalized, we obtain the final result

$$\theta_{jms's}^{(k)} = \frac{\sum_{i \in B_m} I(D_{ij}=s') W_{si}^{(k-1)}}{\sum_{i \in B_m} W_{si}^{(k-1)}} \tag{5}$$

where $\theta_{jms's}^{(k)}$ is the probability that rater j reports label s' when the true label is s at voxel i , and I is the indicator function. This quantity is determined over the voxel set defined by B_m .

2.3 Sliding Windows and Biasing Priors

As evidenced in Figure 1, the spatial quality of a given rater or registered atlas can vary dramatically in a relatively small region. Thus, the number of voxels contained within set B_m should be relatively small. However, as B_m is diminished to account for this phenomenon, the ability to accurately characterize rater performance is dramatically hampered due to the limited number of degrees of freedom when estimating $\theta_{jm}^{(k)}$. To compensate for this problem, we introduce the idea of using a given region, B_m , as a sliding window with significant amount of overlap with other regions. Due to the overlap of the sliding windows, a given voxel i may be an element in multiple region sets. As a result, we use nearest neighbor interpolation to determine the appropriate $\theta_{jm}^{(k)}$ for a given voxel i .

With the overlap of the sliding windows, the estimations at a given $\theta_{jm}^{(k)}$ should be relatively smooth and stable. For example, if a given label is only observed in region B_m a handful of

times, the estimation of a rater's quality at observing this label will be limited. As a result, we introduce the idea of using a whole-image estimate of performance level parameters from STAPLE for regularization. Within a maximum *a posteriori* approach, we would use an informed parametric prior based upon this estimate. Nevertheless, parametric characterization of the manifold of confusion matrices is involved and not strictly necessary. Rather, for computational and stability concerns, we introduce an implicit prior in the

following form. Let $\theta_j^{(0)}$ be the confusion matrix associated with rater j estimated from the STAPLE algorithm. The performance level parameters can then be calculated as

$$\theta_{jms's}^{(k)} = \frac{\sigma \theta_{js's}^{(0)} + \sum_{i \in B_m} I(D_{ij}=s') W_{si}^{(k-1)}}{\sigma \sum_{s'} \theta_{js's}^{(0)} + \sum_{i \in B_m} W_{si}^{(k-1)}} \quad (6)$$

where σ is a scale factor that is dependent upon the size of the region (sliding window) B_m , the number of voxels, and the number of raters. Our empirically derived expression for this scale factor is

$$\sigma = \kappa \frac{N}{N_w \ln R} \quad (7)$$

where κ is an arbitrary constant, N_w is the number of voxels per window B_m , R is the total number of raters and N is the total number of voxels. Note that σ is designed such that the value of κ should be close to unity for a given experiment.

2.4 Initialization and Convergence

The initialization strategies, convergence detection strategies and data-adaptive prior initialization strategies are essentially equivalent to the traditional STAPLE approach [3]. As with the STAPLE approach, each element of θ_{jm} , is initialized to approximately 0.99 along the diagonal of the confusion matrix. The off-diagonal elements are filled with normally distributed values around 0.01. The standard normalization criterion for each column of θ_{jm} is maintained. Convergence is detected when the average change in the on-diagonal values in θ is less than a constant (herein, 0.001). The data-adaptive prior, $f(T_i = s)$, present in the true segmentation estimation was initialized as a global prior such that each element represents the fraction of the total observed voxels ($N \times R$) that were equal to label s . Lastly, from Eq. 7, a value of $\kappa = 1$ was used for all experiments presented.

3 Methods and Results

All simulations and experiments were implemented in MATLAB (Mathworks, Natick, MA). All studies were run on a 64-bit quad-core 3.07 GHz desktop computer with 13 GB of RAM, running Ubuntu 10.04.

3.1 Spatially Varying Rater Quality Simulation

First, we consider a simulation in which raters exhibited spatially varying performance levels when segmenting a 3D volume ($80 \times 60 \times 60$) consisting of 15 embedded cubes (Figure 3). The volume was divided into 16 different equally sized regions. A total of 16 raters observed the volume where each rater was “perfect” in one of the 16 regions while exhibiting boundary-shift error behaviors in the remaining 15 regions. In the boundary-shift error regions, the level of boundary shift for each boundary point was chosen from a

Gaussian distribution with zero mean and a variance specific to each rater. Note that when all 16 raters are fused, there exists a single rater that is perfect at all voxels in the volume.

Spatial STAPLE used a sliding window of $10 \times 10 \times 10$ voxels for a total of 1600 performance level estimations evenly distributed throughout the volume for each rater. Ten Monte Carlo iterations were used to assess the accuracy of STAPLE, Spatial STAPLE and majority vote for fusing the observations between 3 and 16 raters. The sensitivity of the biasing prior was assessed for 8 raters.

Both STAPLE and Spatial STAPLE were superior to majority vote for all numbers of raters fused (Figure 4A). For low numbers of raters the STAPLE and Spatial STAPLE estimates were of approximately equivalent quality. However, as the number of volumes increases, the Spatial STAPLE estimate becomes far superior to the STAPLE estimate.

For high values of κ (Eq. 7) the accuracy of Spatial STAPLE converges to the STAPLE result (e.g., increased impact of prior), while very low values of κ resulted in poor segmentations due to limited sample size in small sliding windows (Figure 4B). Nevertheless, a wide range of κ , the biasing prior resulted in estimates that were significantly superior to the STAPLE estimates.

3.2 Multi-Atlas, Multi-Label Empirical Experiment

Second, we examine an empirical application of label fusion to multi-atlas labeling for extraction of the cortical gray matter using atlases with six tissue labels. Fifteen atlases were registered to a collection of 24 target atlases using the Vectorized Adaptive Bases Registration Algorithm (VABRA) [11] and corresponding whole-brain segmentations were correspondingly deformed to match each target. All segmentations were obtained from the Open Access Series of Imaging Studies (OASIS) [12]. The registered labels were then used as observations for fusion with STAPLE, Spatial STAPLE and majority vote. The Dice Similarity Coefficient (DSC) [13] was used to compare label volumes. A window size of $10 \times 10 \times 10$ voxels was used with 8000 performance level estimations for each rater rectilinearly distributed throughout the volume. Spatial STAPLE provides improvement over STAPLE in certain regions of the brain, particularly the insula (Figure 5). As the number of volumes fused increases, a consistent improvement in label agreement is evident.

Third, the performance of all three methods was characterized for an intricate whole-brain labeling with 41 labels otherwise using the same procedure as above. Spatial STAPLE exhibits significant improvements over STAPLE estimate (Figure 6). For several of the labels, particularly the gray matter and some smaller labels in the mid-brain, Spatial STAPLE shows visible improvement over the majority vote estimate (Figure 6). For more than 5 registered atlases, both Spatial STAPLE and majority vote are superior to STAPLE. The median performance of majority vote and Spatial STAPLE are very similar as can be seen in Figure 6F; however, Spatial STAPLE results in a non-trivial improvement over majority vote for a large number of outlier cases.

4 Discussion and Conclusion

Spatial STAPLE accounts for spatial quality variation through seamless integration of multiple confusion matrices with the established STAPLE theory. This algorithm presents a dramatic change in how we can view the implementation of statistical fusion algorithms. With only a single confusion matrix, one is limited to cases where each rater is consistent throughout an entire volume. Our approach provides a conceptually simple and computationally tractable method to account for spatially varying performance characteristics. As presented, highly irregular performance level variations (as observed in

Figure 1) can be accounted for and utilized to perform local weighting of atlas labels, and, in turn, improve estimations of the true segmentation and understanding of rater performance. The introduction of a data informed prior regularizes the estimation process and enables localized performance estimates without substantially decreasing precision. Nevertheless, important topics of future research persist, such as a more thorough formulation of the biasing prior and a characterization of the sensitivity of the algorithm to the initialization of the performance level parameters. Additionally, this algorithm will have to be compared to more sophisticated voting based fusion algorithms [5, 9] to more completely assess performance.

A subtle, but important, result of this work is that by accounting for spatially varying rater performance in these demonstrations, the often-observed reliability gap between statistical fusion and majority vote has been closed or overcome. Both the use of local performance characterization and the introduction of data-informed priors could be used with other recent innovations in statistical fusion (such as new data types or rater models). The degree of correlation and impact of data informed priors can be readily controlled through hierarchical or regional derivations. As presented, the framework estimates and exploits spatially varying structure, but the performance gains from this approach suggest that this proposed framework could also be used to estimate the structure of potential correlations in rater performance and, perhaps, exploit these characteristics in the design of statistical fusion paradigm.

Acknowledgments

This work was supported in part by NIH/NINDS 1R01NS056307 and NIH/NINDS 1R21NS064534.

References

1. Warfield S, Dengler J, Zaers J, Guttman CR, Wells WM 3rd, Ettinger GJ, Hiller J, Kikinis R. Automatic identification of gray matter structures from MRI to improve the segmentation of white matter lesions. *J Image Guid Surg.* 1995; 1:326–338. [PubMed: 9080353]
2. Kikinis R, Shenton ME, Gerig G, Martin J, Anderson M, Metcalf D, Guttman CR, McCarley RW, Lorensen W, Cline H, et al. Routine quantitative analysis of brain and cerebrospinal fluid spaces with MR imaging. *J Magn Reson Imaging.* 1992; 2:619–629. [PubMed: 1446105]
3. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging.* 2004; 23:903–921. [PubMed: 15250643]
4. Rohlfing T, Russakoff DB, Maurer CR. Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation. *IEEE Transactions on Medical Imaging.* 2004; 23:983–994. [PubMed: 15338732]
5. Langerak T, van der Heide U, Kotte A, Viergever M, van Vulpen M, Pluim J. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). *IEEE Transactions on Medical Imaging.* 2010
6. Lotjonen JM, Wolz R, Koikkalainen JR, Thurfjell L, Waldemar G, Soininen H, Rueckert D. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage.* 2010; 49:2352–2365. [PubMed: 19857578]
7. Isgum I, Staring M, Rutten A, Prokop M, Viergever M, van Ginneken B. Multi-atlas-based segmentation with local decision fusion—Application to cardiac and aortic segmentation in CT scans. *Medical Imaging, IEEE Transactions.* 2009; 28:1000–1010. on.
8. Depa M, Sabuncu M, Holmvang G, Nezafat R, Schmidt E, Golland P. Robust Atlas-Based Segmentation of Highly Variable Anatomy: Left Atrium Segmentation. *Statistical Atlases and Computational Models of the Heart.* 2010:85–94.
9. Sabuncu M, Yeo B, Van Leemput K, Fischl B, Golland P. A Generative Model for Image Segmentation Based on Label Fusion. *IEEE Transactions on Medical Imaging.* 2010

10. Kittler J, Alkoot F. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis And Machine Intelligence*. 2003;110–115.
11. Rohde GK, Aldroubi A, Dawant BM. The adaptive bases algorithm for intensity-based nonrigid image registration. *IEEE Trans Med Imaging*. 2003; 22:1470–1479. [PubMed: 14606680]
12. Marcus D, Wang T, Parker J, Csernansky J, Morris J, Buckner R. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*. 2007; 19:1498–1507. [PubMed: 17714011]
13. Dice L. Measures of the amount of ecologic association between species. *Ecology*. 1945; 26:297–302.

Fig. 1.

The spatial quality variation exhibited by registered atlases. A representative slice from the true labels (manually drawn) of a target brain is presented in (A). Four example observations of this slice can be seen in (B). The quality of the observations seen in (B) is compared to the true labels seen in (A) to construct the spatial quality variation heat maps presented in (C). Note that the spatial variation is independent of the actual labels of the brain.

Fig. 2.

A visual representation of the Spatial STAPLE algorithm. The images in (A – D) represent varying confusion matrices for the various regions of the presented observation. The confusion matrix presented in (D) is of significantly higher quality than the other confusion matrices as represented by the fact that it is nearly a diagonal matrix. The observation can be seen in (E) where the regions corresponding to each confusion matrix is specified.

Fig. 3. Spatially varying rater quality simulation. The 3D truth model used in the simulation can be seen in both (A) and (B). (A) shows each of the individual slices of the model and (B) shows the three main cross-sections. (C) presents an example observation of the truth model. Representative estimates from both Spatial STAPLE (D) and STAPLE (E) are shown using 10 raters.

Fig. 4. Simulation results for spatial varying performance. The results from the 10 Monte Carlo iteration simulation can be seen in (A). It is evident that both STAPLE and Spatial STAPLE outperform majority vote for all numbers of raters. For increasing numbers of raters Spatial STAPLE dramatically outperforms STAPLE. The sensitivity of the implicit prior can be seen in (B). Note that for high values, the Spatial STAPLE estimate converges to the STAPLE estimate. For low values, the Spatial STAPLE estimate is unstable and results in estimations that are dramatically worse than STAPLE.

Fig. 5.

Results from an empirical experiment using 6 labels. A representative truth slice from the 6-label model can be seen in (A). A cropped and rotated region (A) is presented in (B). The estimates seen in (C) and (D) represent the output from Spatial STAPLE (C) and STAPLE (D) using 8 volumes. The accuracy of the gray matter estimation (in terms of the difference in DSC values) for varying numbers of volumes can be seen in (E). For reference, the average STAPLE DSC was approximately 0.8.

Fig. 6.

Results from the empirical experiment using 41 labels. A representative slice from the truth model can be seen in (A). The majority vote, STAPLE and Spatial STAPLE estimates for this slice using 8 volumes can be seen in (B), (C) and (D), respectively. A comparison of the fraction of voxels correct when fusing 5 to 15 volumes to all 24 target atlases can be seen in (E). A per label comparison between Spatial STAPLE and majority vote can be seen in (F). Only the 36 labels that were consistent between the 24 target atlases are shown. Note that both Spatial STAPLE and majority vote outperform STAPLE for increasing numbers of volumes in (E). Lastly, note that there are a large number of outliers for which Spatial STAPLE outperforms majority vote in (F).