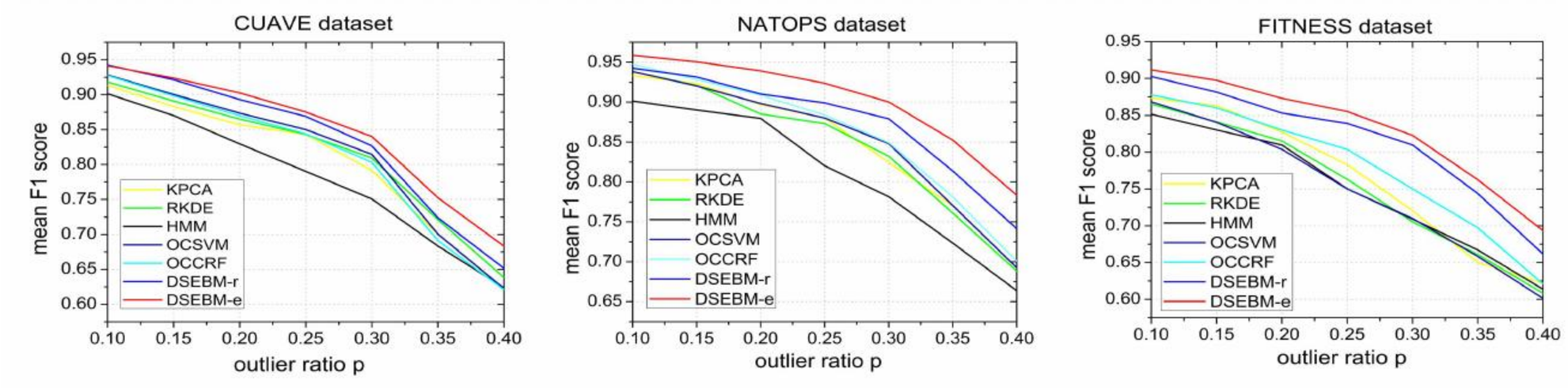

Figure 2. The means F_1 scores on the three sequential datasets with outlier ratio from 0.1 to 0.4.

Method	CUAVE ($\rho = 0.3$)			NATOPS ($\rho = 0.3$)			FITNESS ($\rho = 0.3$)		
	mPrec	**mRec**	**mF_1**	**mPrec**	**mRec**	**mF_1**	**mPrec**	**mRec**	**mF_1**
Kernel PCA	0.8531	0.7429	0.7916	0.8532	0.7842	0.8244	0.7518	0.6733	0.7206
RKDE	0.8479	0.7713	0.8094	0.8445	0.7139	0.7822	0.7417	0.6632	0.7054
HMM	**0.8793**	0.6202	0.7507	0.8693	0.5724	0.7215	0.7883	0.6321	0.7092
OC-SVM	0.8512	0.7726	0.8147	0.8741	0.8236	0.8481	0.7472	0.6774	0.7099
OCCRF	0.8529	0.7598	0.8032	0.8795	0.8582	0.8493	0.8088	0.7173	0.7501
DSEBM-r	0.8692	0.7381	0.8268	0.9035	0.8655	0.8792	0.8425	0.7710	0.8097
DSEBM-e	0.8754	**0.8033**	**0.8402**	**0.9178**	**0.8856**	**0.9022**	**0.8533**	**0.7873**	**0.8228**

Table 3. **CUAVE, NATOPS, FITNESS**: mean precision (mPrec), mean recall (mRec) and mean F_1 (mF_1) over over the sequential data sets of seven methods. For each column, the best result is shown in boldface.

Method	Caltech-101 ($\rho = 0.3$)			MNIST ($\rho = 0.3$)			CIFAR-10 ($\rho = 0.3$)		
	mPrec	**mRec**	**mF_1**	**mPrec**	**mRec**	**m F_1**	**mPrec**	**mRec**	**mF_1**
HR-PCA	0.8735	0.8025	0.8534	0.9278	0.9493	0.9352	0.8459	0.8217	0.8342
Kernel PCA	0.9005	0.9091	0.9022	0.9427	0.9576	0.9511	0.8552	0.8473	0.8506
RKDE	0.8904	0.8995	0.8952	0.9377	0.9218	0.9306	0.8319	0.8202	0.8261
OC-SVM	0.8598	0.8772	0.8674	0.9245	0.9432	0.9356	0.8332	0.8206	0.8259
UOCL	0.9203	0.9076	0.9135	0.9342	0.9198	0.9256	0.8613	0.8442	0.8520
DSEBM-r	**0.9184**	0.9037	0.9077	0.9597	0.9503	0.9536	0.8742	0.8603	0.8681
DSEBM-e	0.9175	**0.9042**	**0.9159**	**0.9788**	**0.9616**	**0.9689**	**0.8873**	**0.8647**	**0.8784**

Table 4. **Caltech-101, MNIST, CIFAR-10** datasets: mean precision (mPrec), mean recall (mRec) and mean F_1 (mF_1) over over the image data sets of eight methods. For each column, the best result is shown in boldface.

category as inliers, and sample images from the other categories with a proportion $0.1 \leq \rho \leq 0.4$. Each dataset is split into a training and testing set with a ratio of 2:1. We compare DSEBMs (with one convolutional layer + one pooling layer + one fully connected layer) with several baseline methods including: High-dimensional Robust PCA (HR-PCA), Kernel PCA (KPCA), Robust Kernel Density Estimator (RKDE), One-Class SVM (OC-SVM) and Unsupervised One-Class Learning (UOCL) (Liu et al., 2014). All the results are shown in Table 4 with $\rho = 0.3$. We see that DSEBM-e is the best performing method overall in terms of mean recall and mean F_1, with particularly large margins on large datasets (MNIST and CIFAR-10). Measured by F_1, DSEBM-e improves 3.5% and 2.3% over the best-performing baselines. Figure 3 with ρ varying from 0.1 to 0.4 also demonstrates consistent results.

5.4. Energy VS. Reconstruction Error

In terms of the two decision criteria of DSEBM, we observe that DSEBM-e consistently outperforms DSEBM-r on all the benchmarks except for the Thyroid dataset. This verifies our conjecture that the energy score is a more accurate decision criterion than reconstruction error. In addition, to gain further insight on the behavior of the two criteria, we demonstrate seven outliers selected from the Caltech-101 benchmark in Figure 4. For each image, the energy scores are displayed at the second row in red, followed by the reconstruction error displayed in green and the correct inlier class. Interestingly, all the seven outliers are visually similar to the inlier class and have small reconstruction errors (compared with the threshold). However, we are able to successfully identify all of them with energy (which are higher than the energy threshold).