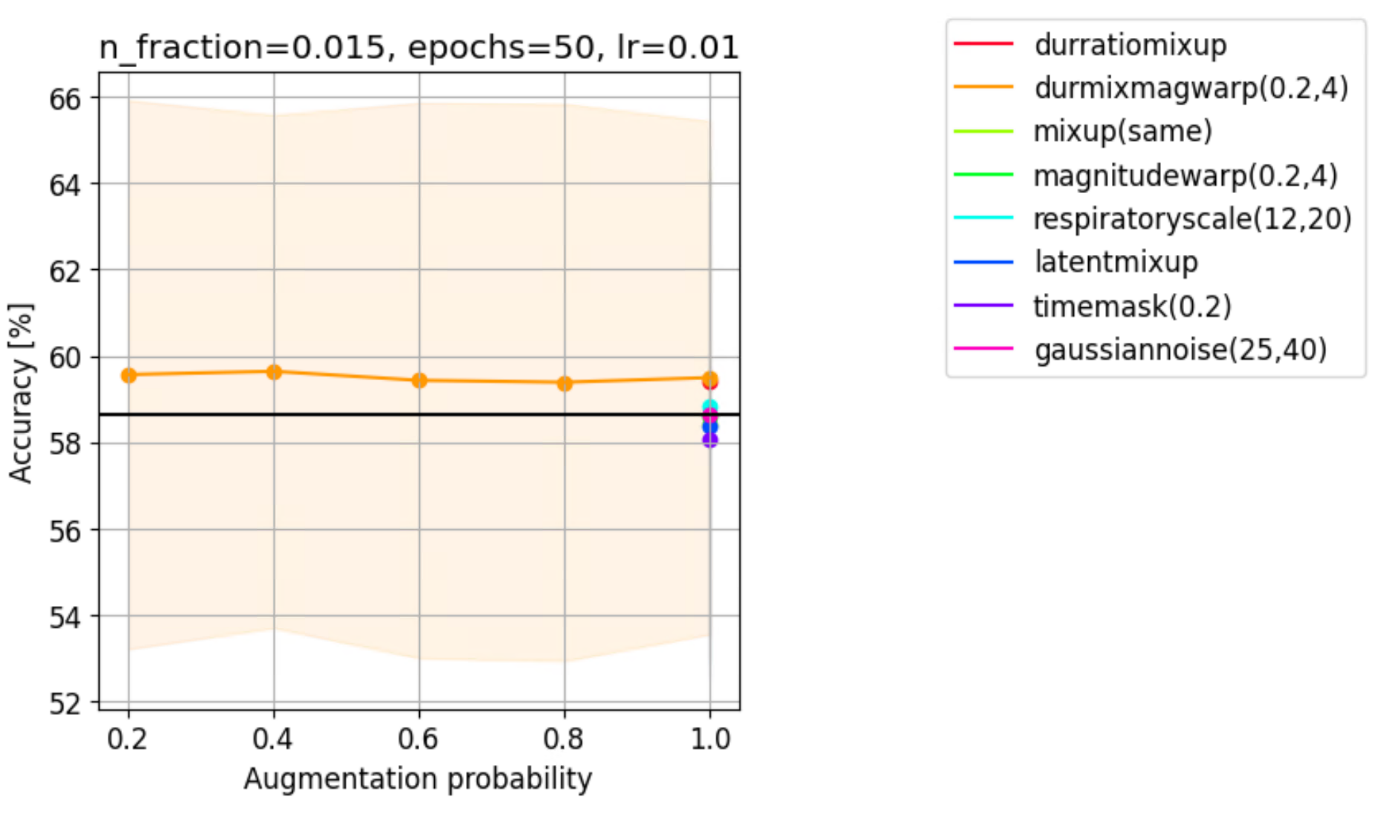
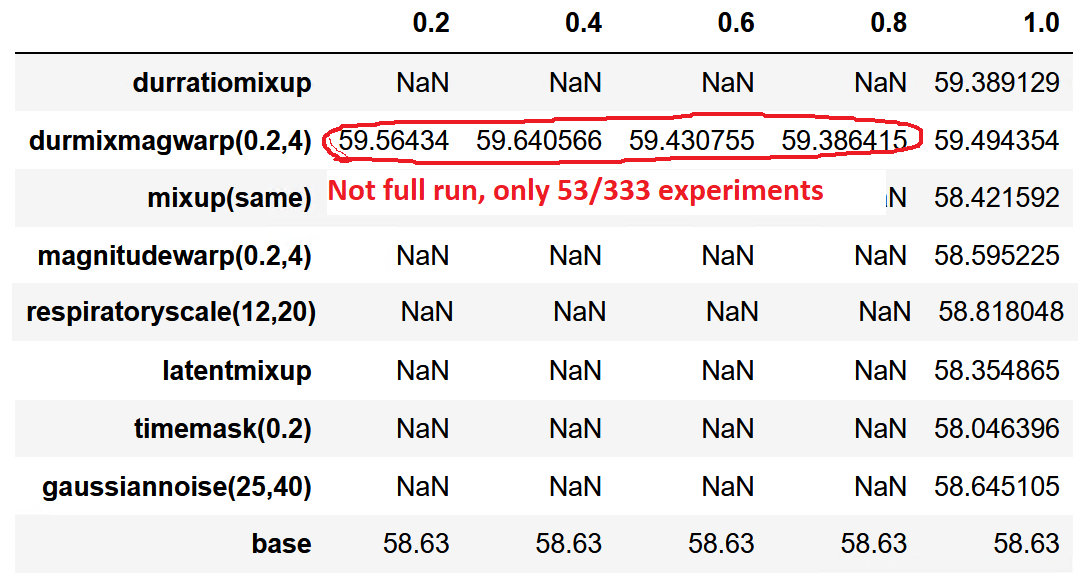
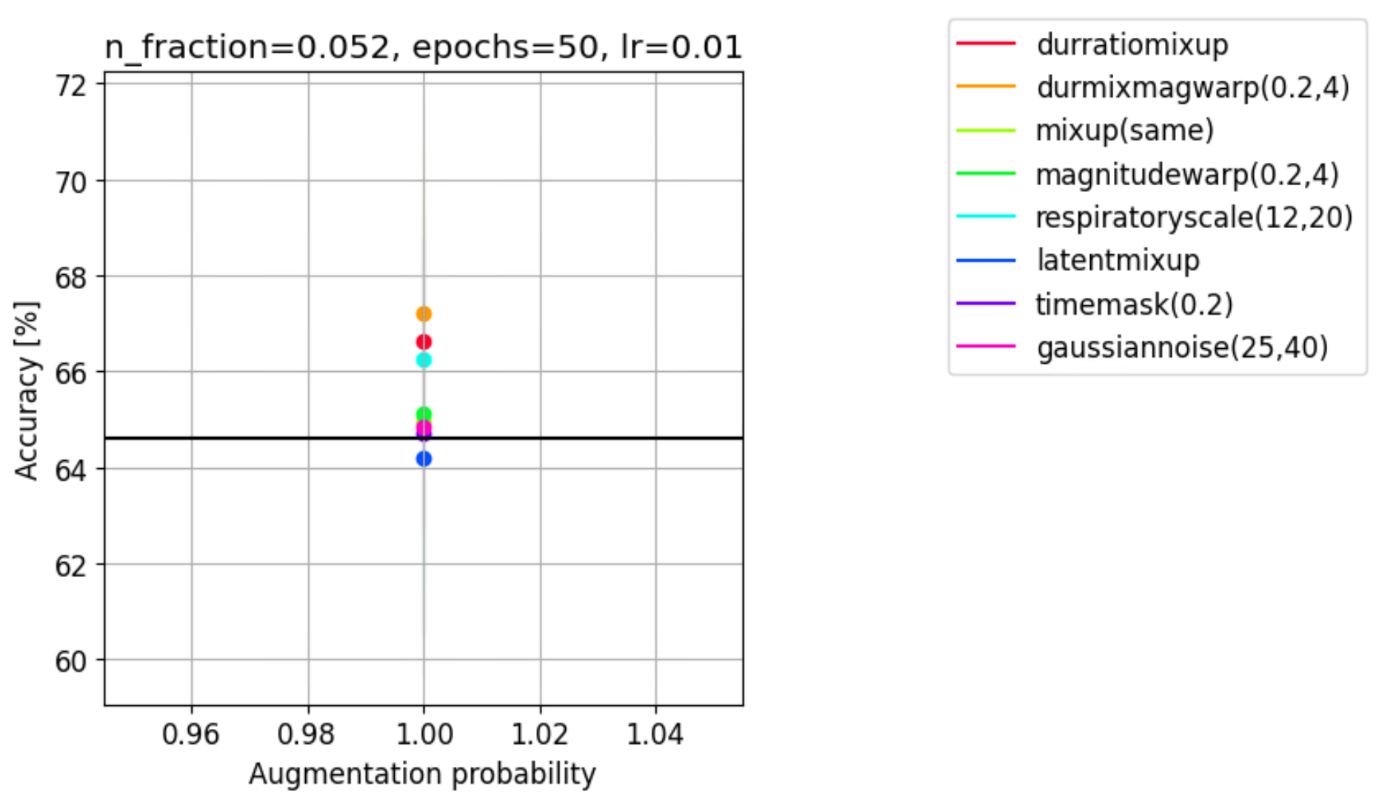
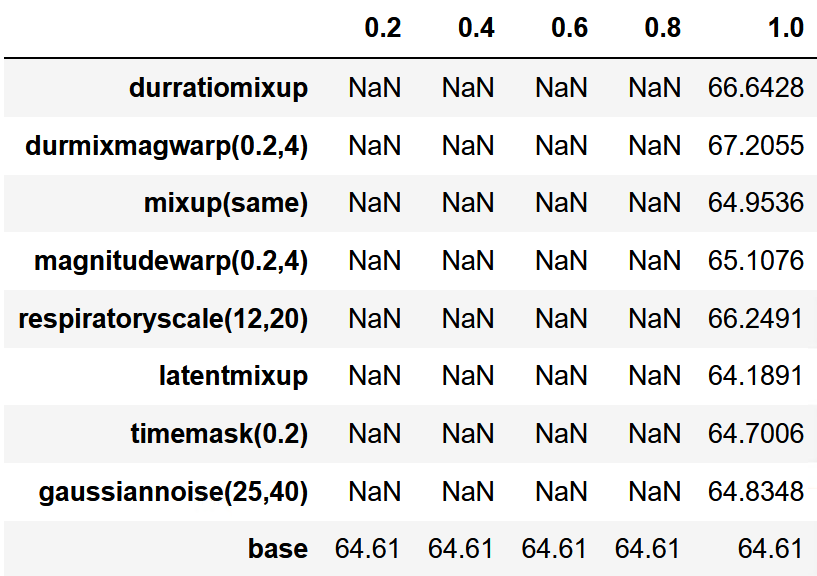
Final aug\_probas: nfrac=0.015->1.0, nfrac=0.052->1.0, nfrac=0.1->1.0, nfrac=0.2->0.8, nfrac=0.3->0.6

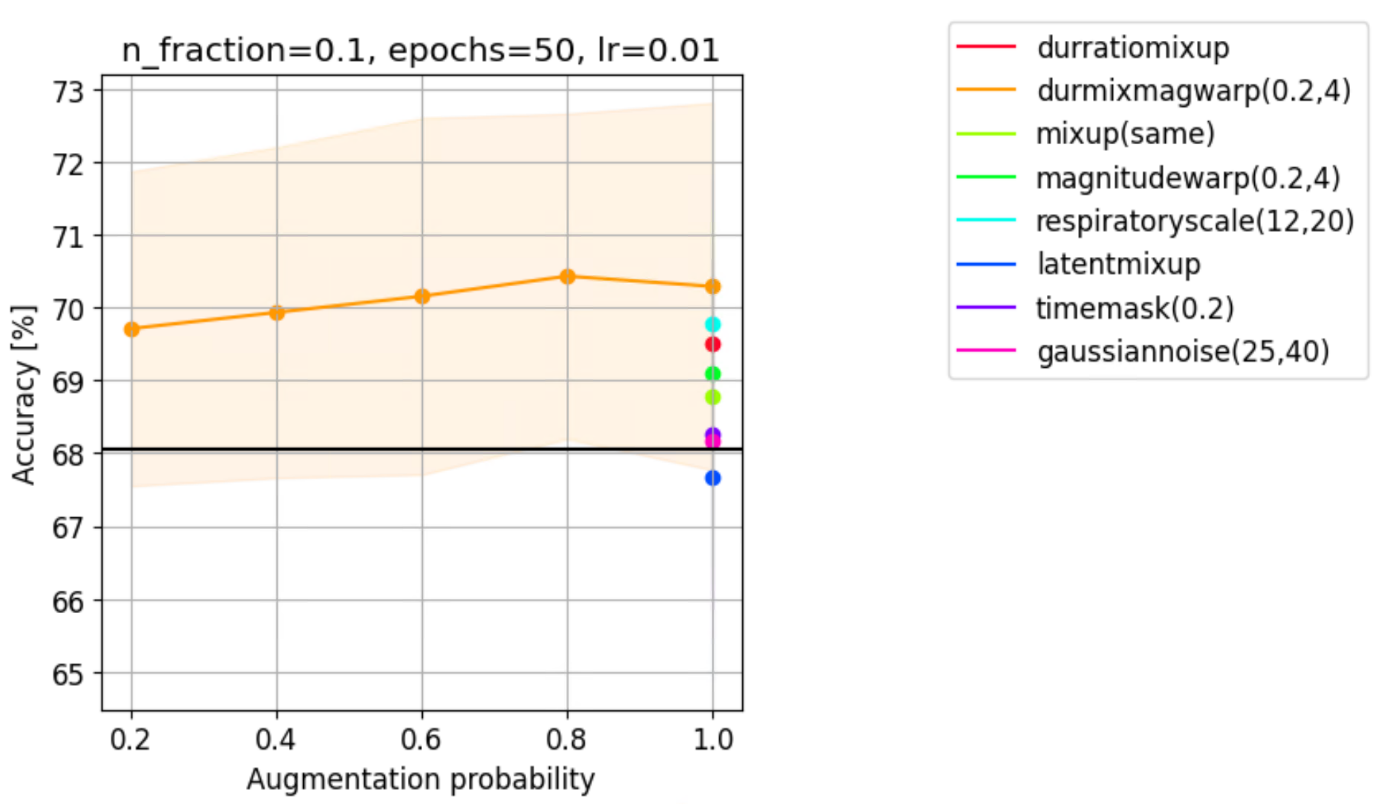
Potes

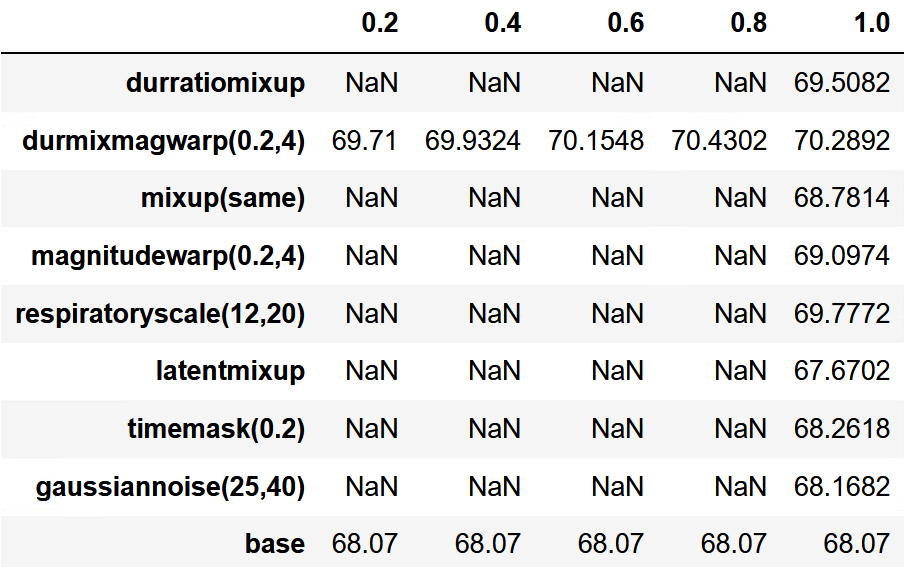


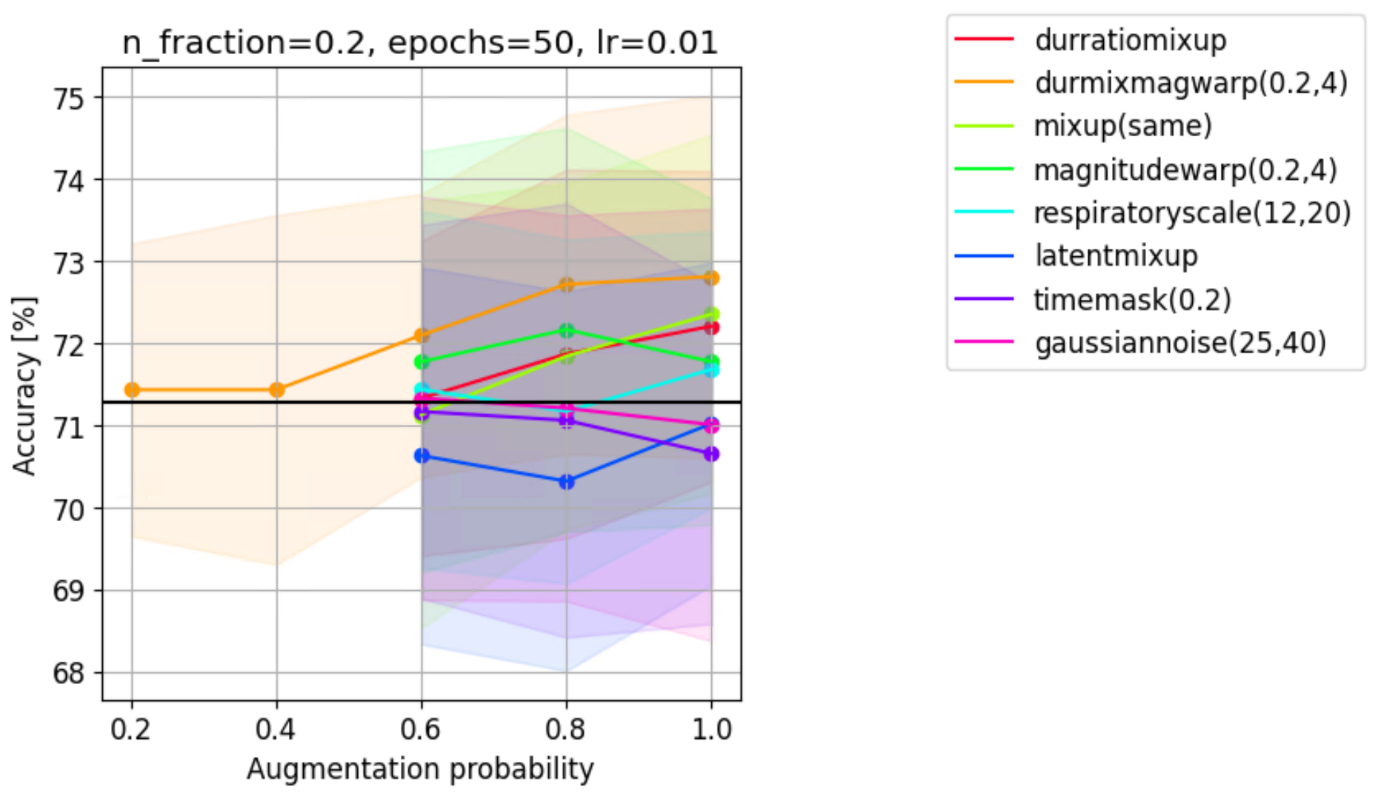


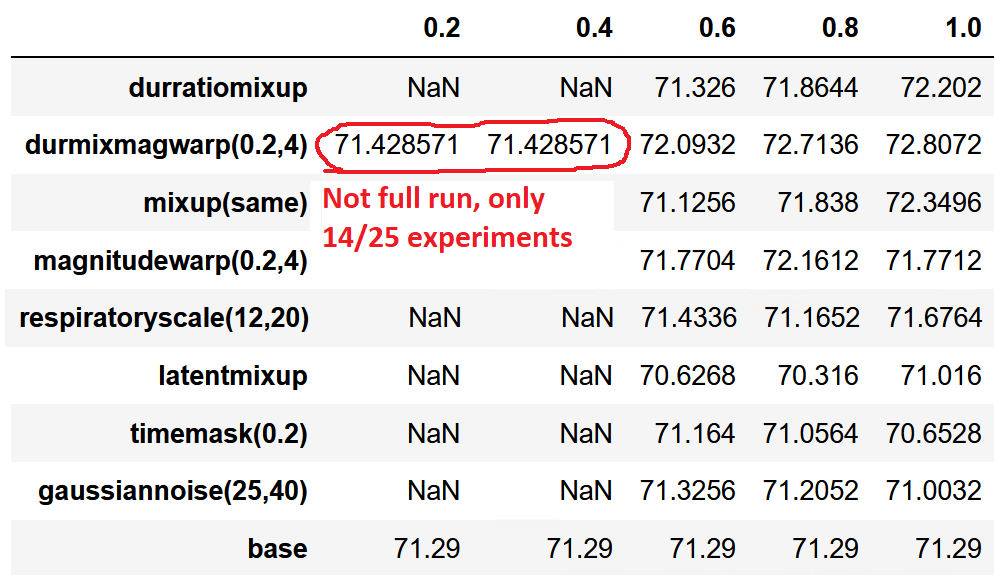


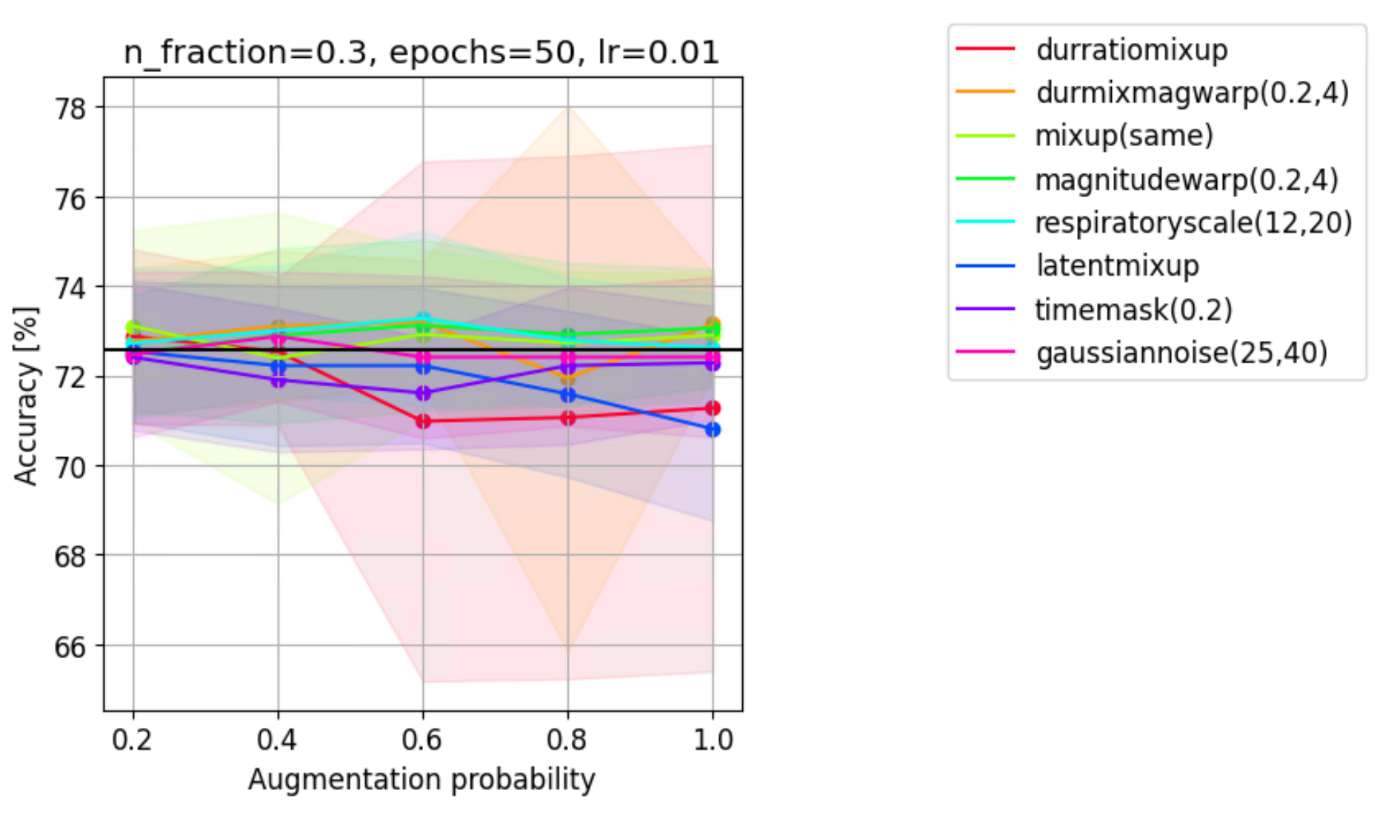


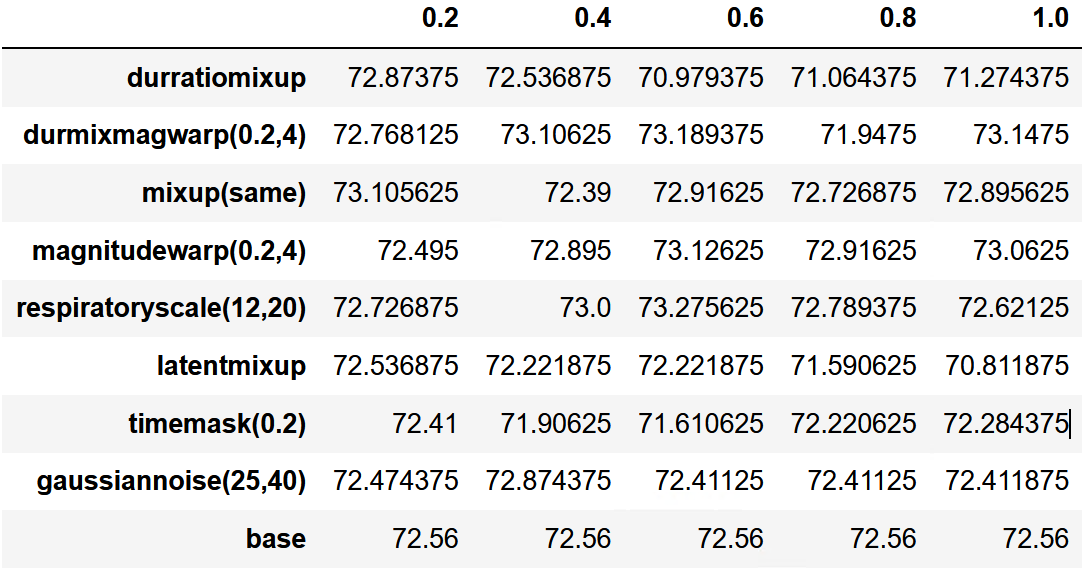












Average across aug\_proba = (0.6, 0.8, 1.0)

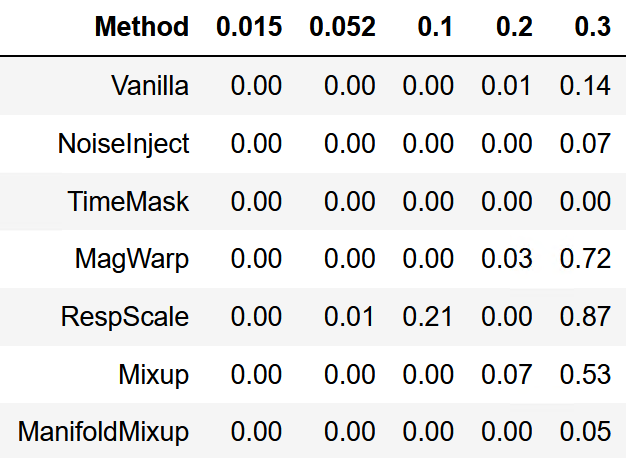
* durmixmagwarp(0.2,4) = 72.761458
* Magnutudewarp(0.2,4) = 73.035
* Mixup = 72.84625
* Respiratoryscale(12,20)= 72.895416

->our method 4th due to unluckly outlier - keep in mind that still better than baseline (72.56) and best aug\_proba=1.0 and second best at aug\_proba=0.6

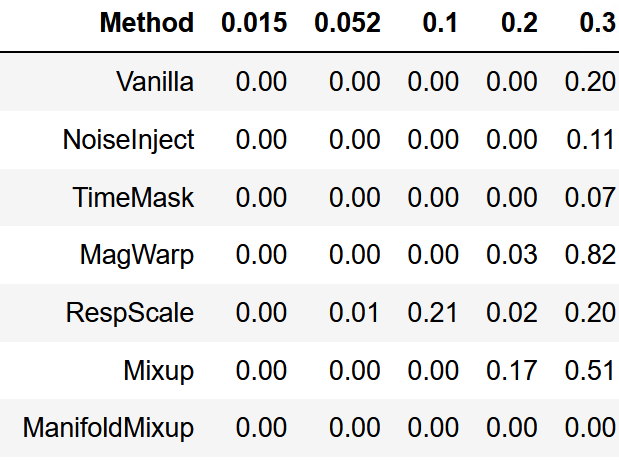
Default mode: aug\_proba aways 1.0 -> our method, durmixmagwarp(0.2,4), always best

P-values between PCGmix+ and other methods

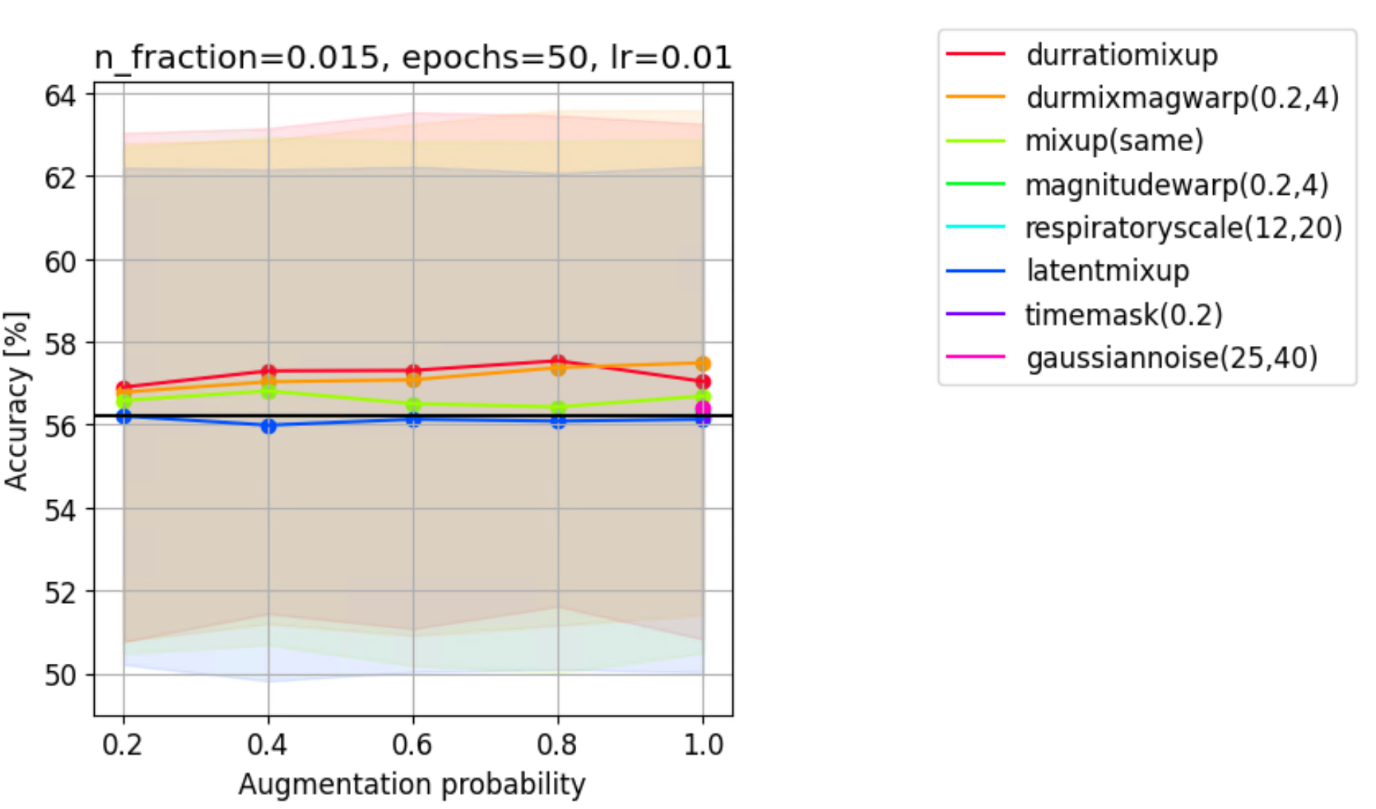
Potes final (selected) aug probas:

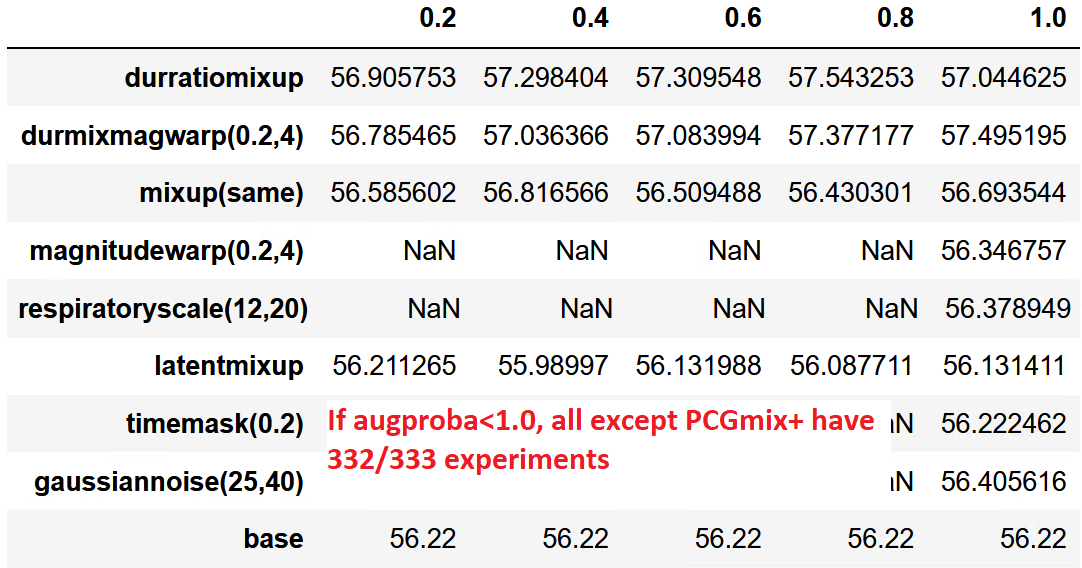


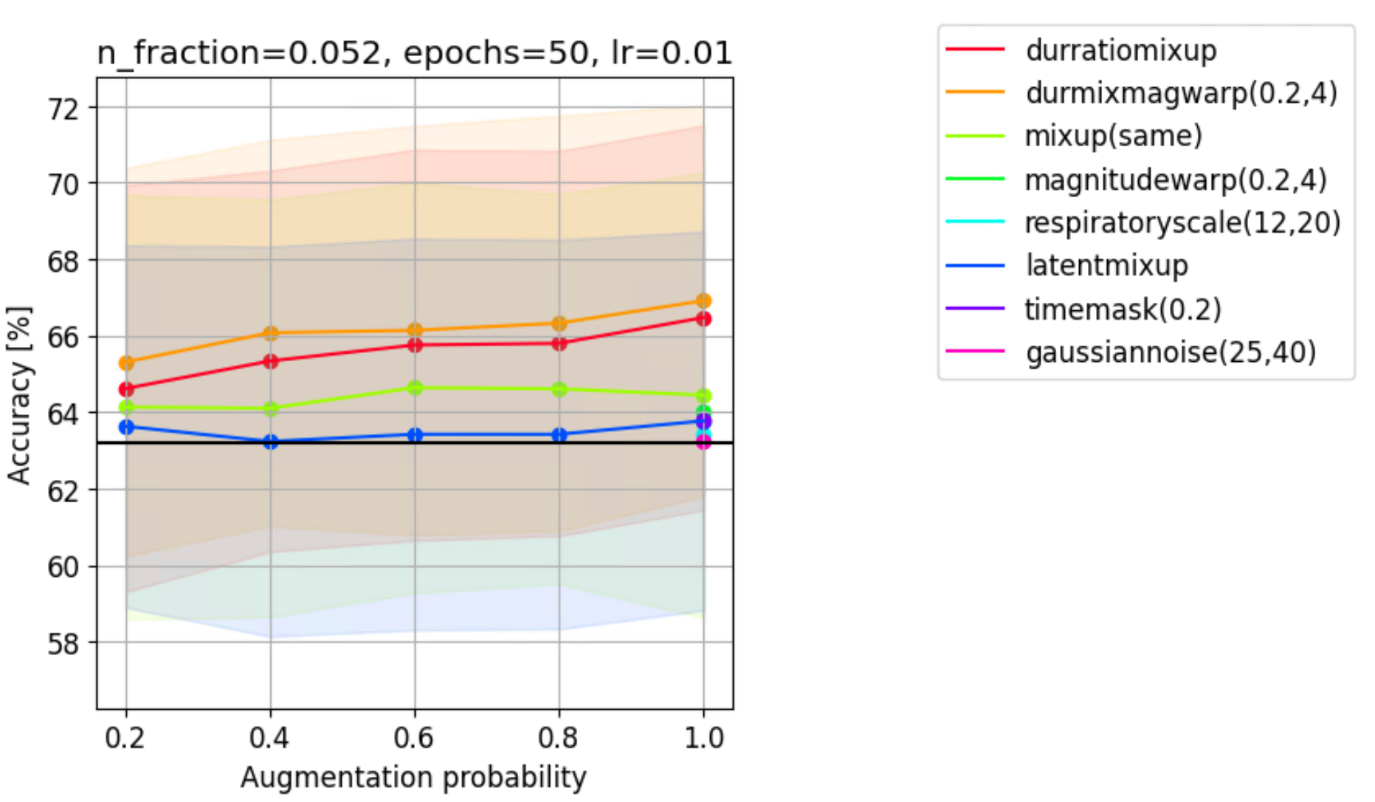
Potes default (always 1.0) aug probas:

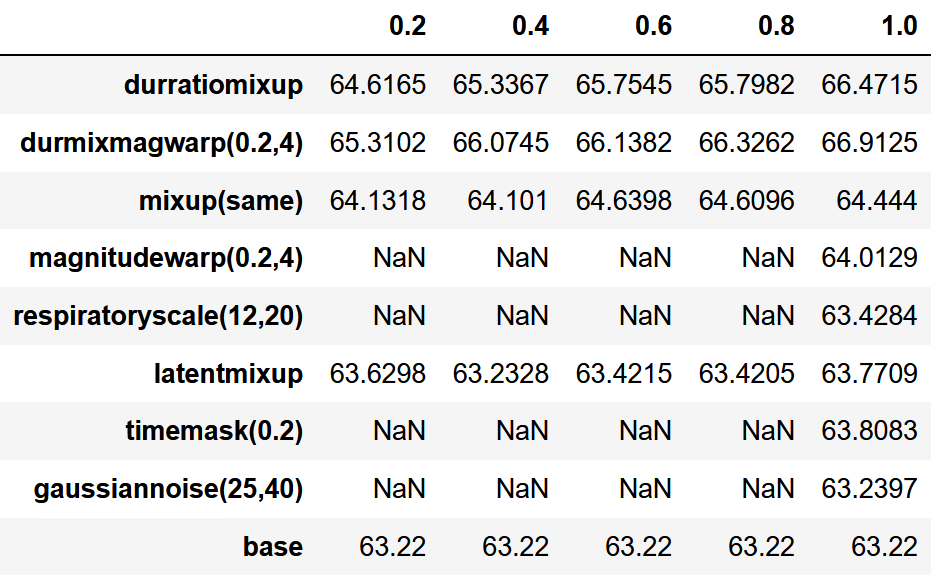


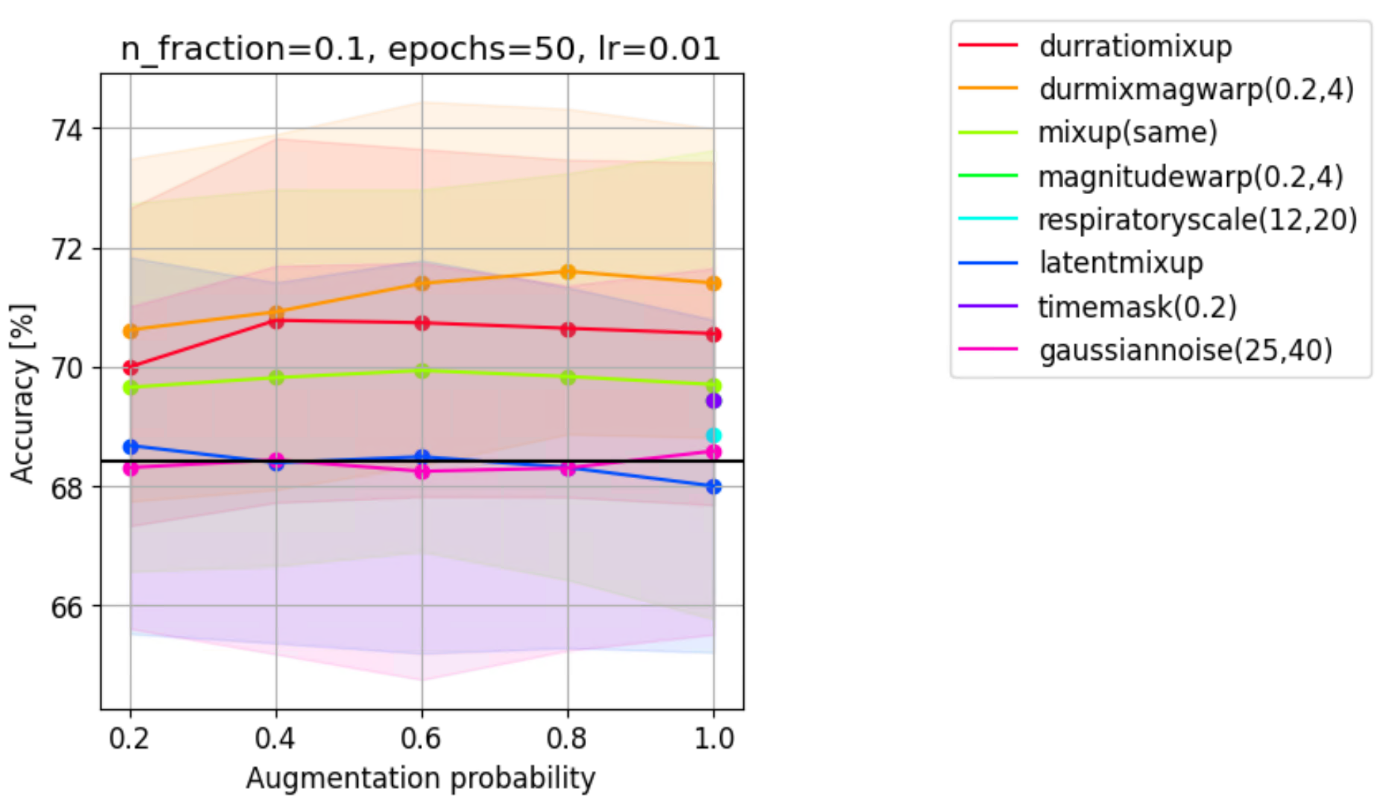
Resnet

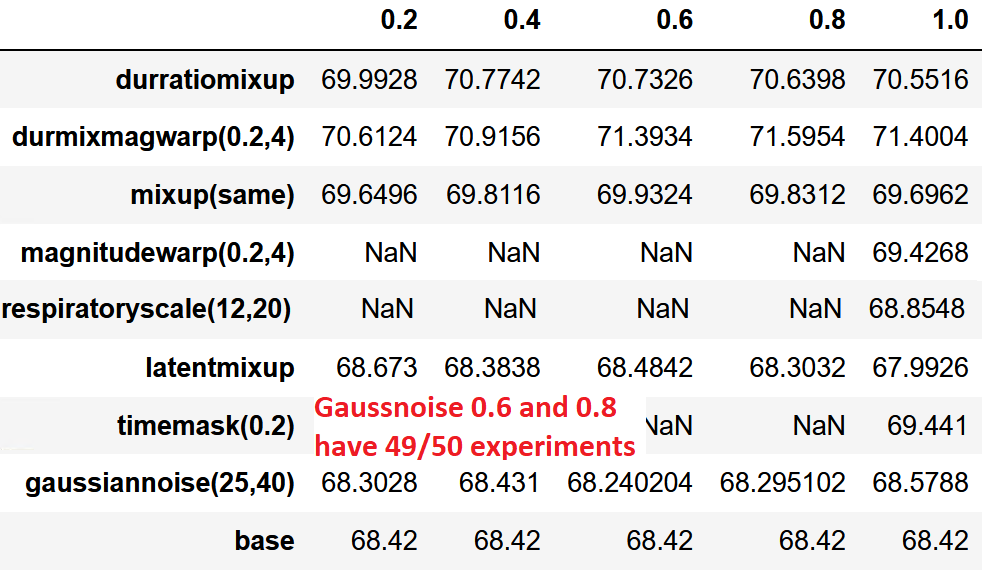


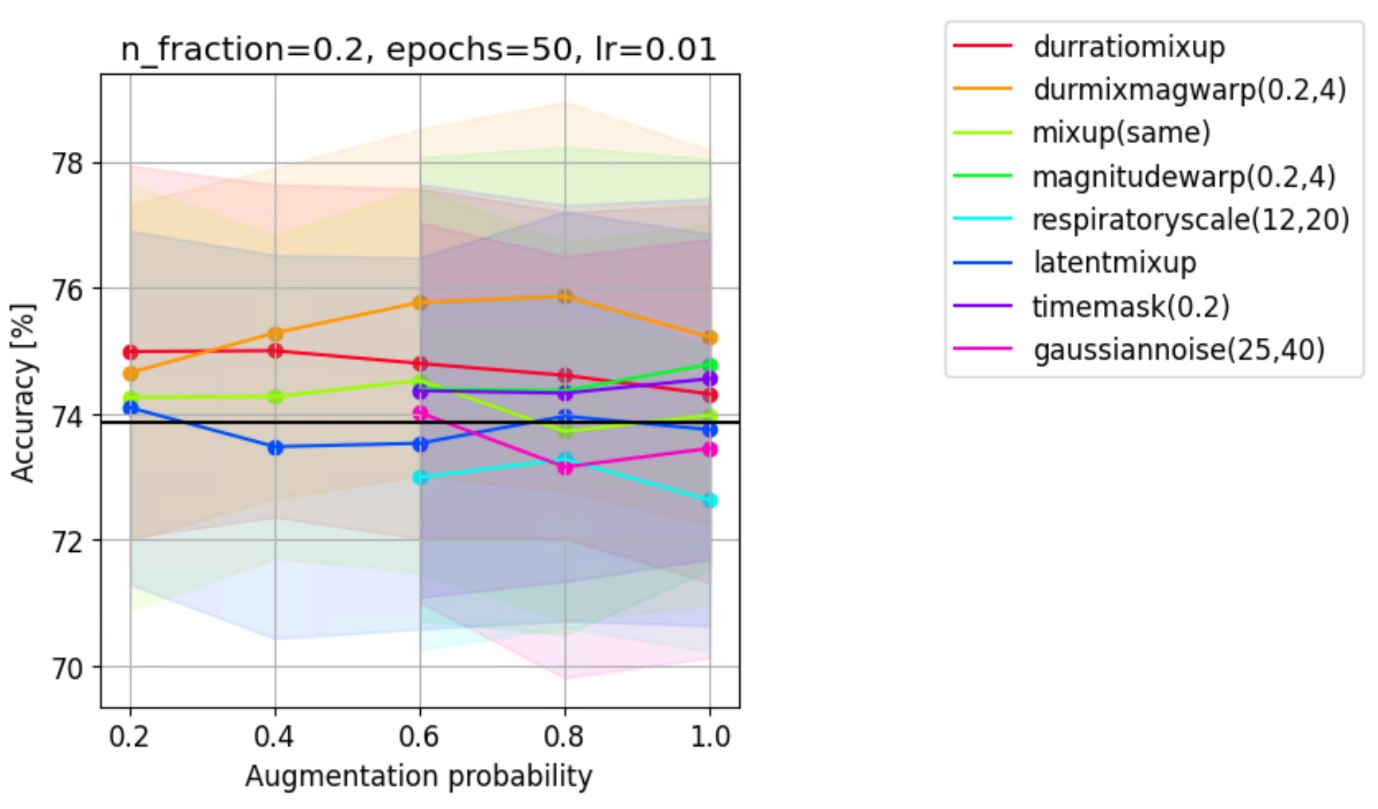


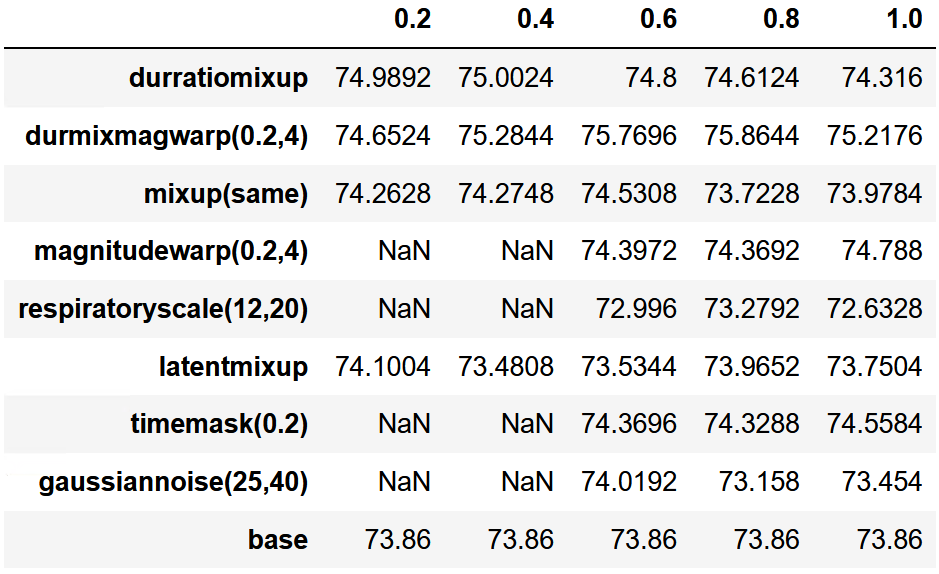




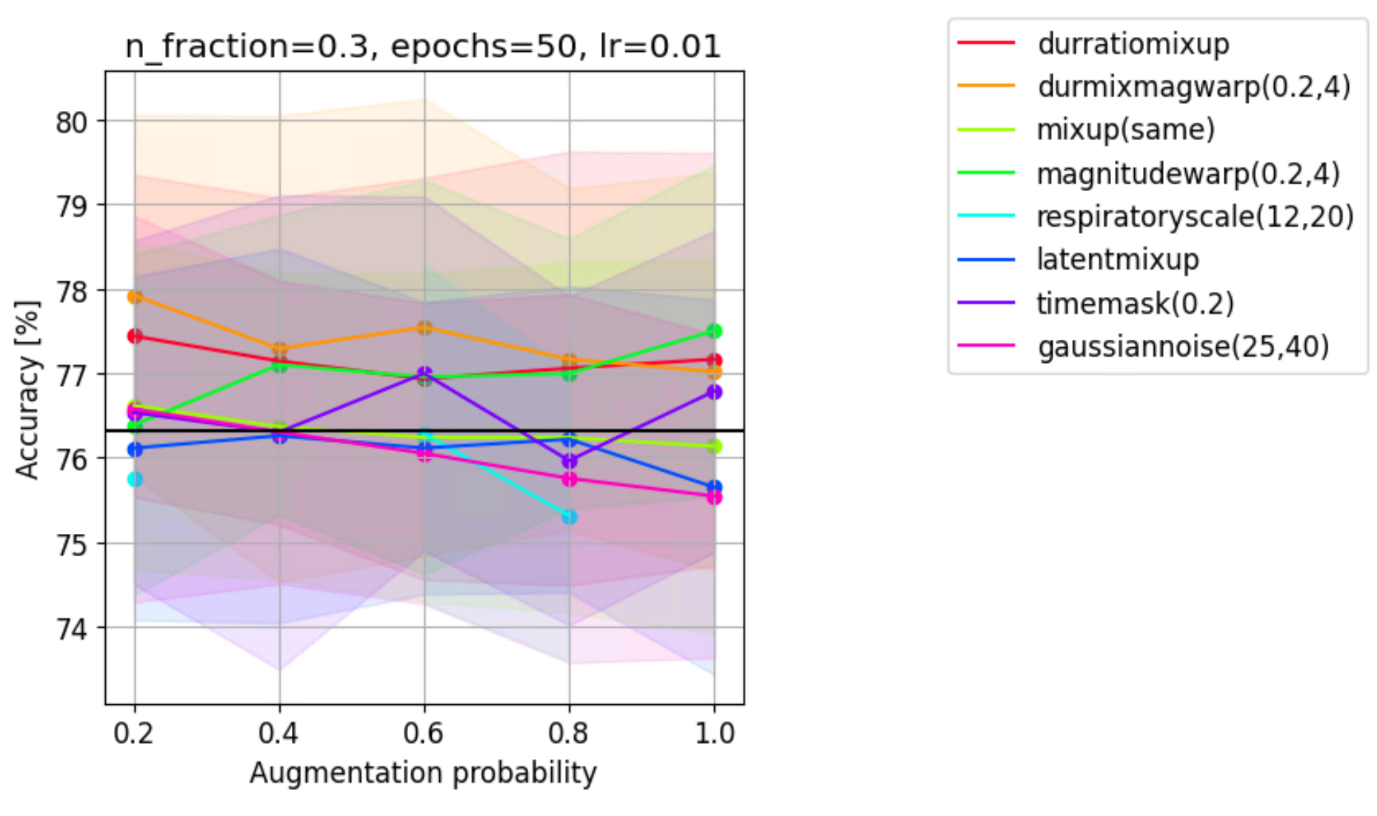


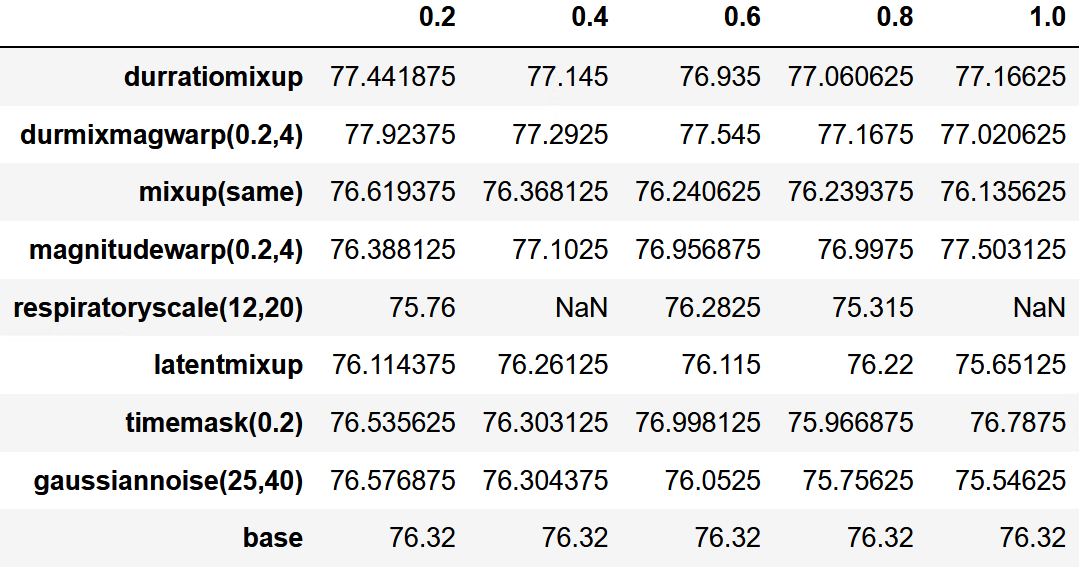






Aug\_proba=1.0 is 0.65% acc worse than aug\_proba=0.8





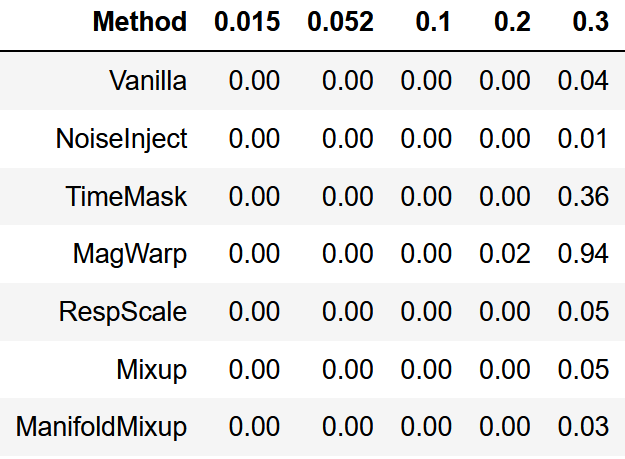
Average across aug\_proba = (0.6, 0.8, 1.0) (note that at aug\_proba=0.2, ours is »by far« the best overall:

* durmixmagwarp(0.2,4) = 77.244375
* Magnutudewarp(0.2,4) = 77.1525

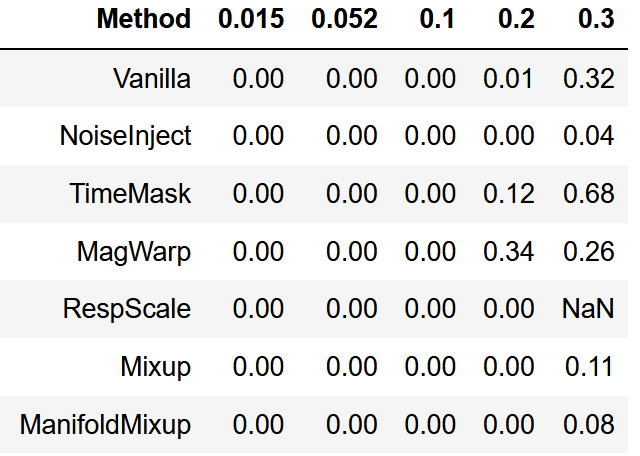
->Ours wins

Default mode: aug\_proba aways 1.0 -> our method, durmixmagwarp(0.2,4), always best, except nfrac=0.3 where it is 0.5% acc worse than magnitudewarp(0.2,4)

Resnet final aug probas:



Resnet default (1.0) aug probas:



(note that for ResScape we didn't run experiments at aug\_proba=1.0, but that method is worse than baseline)

VERDICT

Up until n\_frac=0.3, our method is clearly by far the best at all times, no matter the aug\_proba. At n\_frac=0.3, our method still potentially the best, although we do not claim that anymore. It's most important that it's still much better than the baseline. At that point, it looks like magnitudewarp(0.2,4) is comparable to our method.

Statistical significance: again, up until n\_frac=0.3, our method statistically significantly better than the baseline.

* For potes: using final (selected) or default aug\_probas results in the same story up until n\_frac=0.3 (RespScale at n\_frac=0.1 and mixup at n\_frac=0.2 have p-value above 0.05), whereas at 0.3 (where we dont claim superiority over other methods anyway) TimeMask's p-value gets ovr 0.05, whereas ManifoldMixup's deases from 0.05 to 0.00.
* For resnet: when using default aug\_proba compared to final (selected), TimeMask and MagWarp get their p-values above 0.05 at n\_frac=0.2, and at n\_frac=0.3, baseline and ManifoldMixup p-values get above 0.05.

Detailed aug\_proba analysis showed that our method is top for up to (including) n\_frac=0.2. At n\_frac=0.3, it's still great (although not as superior) and still always better than the baseline.