# TorchCP Application Scenario Analysis

12112504 Guo Cheng

Explore the experimental code at [this GitHub link](https://github.com/SustechSTA303/STA303-Assignment04/pull/38).

## 1 Description

This report undertakes a comparative analysis of the performance of diverse conformal prediction (CP) scores and predictors within their respective classification settings. We employ TorchCP (Huang et al., 2023), a dynamic toolbox for CP. The score functions subjected to testing include **THR** (Sadinle et al., 2016), **APS** (Romano et al., 2020), **SAPS** (Huang et al., 2023), and **RAPS** (Angelopoulos et al., 2020). Simultaneously, the predictors under scrutiny comprise **SplitPredictor** (Lei et al., 2016), **ClusterPredictor** (Ding et al., 2023), and **ClassWisePredictor** (Shi et al., 2013). In total, we assess 12 pairs of scores and predictors. In our approaches, we concentrate on assessing the performance of these scores and predictors across a variety of datasets.

The testing procedure follows a classification methodology, encompassing the curation of datasets, including STL10 (Coates et al., 2011), CIFAR (Krizhevsky and Hinton, 2009) Stanford dogs (Khosla et al., 2011), FashionMINIST (Xiao et al., 2017), and UTKFace (Zhang et al., 2017). The evaluation process consists of: 1. dataset selection, 2. model training for a restricted number of epochs, and 3. applying the standard conformal prediction process to the pairs of models and datasets, resulting in a comprehensive conformal prediction outcome. In our study, we employed two conformal prediction approaches on the second process: Ordinary Training with conformal prediction and Conformal Training (Stutz et al., 2022) with conformal prediction.

## 2 Detailed Design Patterns of the Project

### 2.1 Dataset Standardization

We standardized the datasets to ensure consistency, with each comprising 10,000 training samples, 5,000 test samples, and 5,000 calibration samples. STL10 is an exception, constrained to 8,000 training samples, and 4,000 each for test and calibration due to limitations.

### 2.2 Selected Model: RESNET18

Our chosen model is **RESNET18**, renowned for its robust performance across various classification tasks. We adapted the model by replacing its final fully connected layer to align with the specific number of categories intended for classification.

### 2.3 Parameters of Score Functions

In the classification model domain, SAPS and RAPS are variations of APS. When the weight of SAPS or the penalty of RAPS is set to 0, they become identical to APS. To assess their effectiveness, we chose the values of 0.2 for weight and 0.001 for penalty, as suggested in their respective papers. Value selections were also validated through tests on the selected datasets below. Our experiments utilized an alpha value of 0.1.

### 2.4 Selected Datasets

The selected datasets encompass a diverse range, including:

1. CIFAR10 and CIFAR100 , known for low-quality images. They also serve as comparisons between classes with fewer labels (10) and with more labels (100).
2. Stanford Dogs and FashionMNIST, featuring specific types of relatively high-quality images.
3. STL10, offering high-quality images with diverse content.
4. UTKFace, containing diverse human face images representing various ethnicities and categorized by age groups.

### 2.5 Experimentation Approaches

The experimentation involved two approaches on the second process:

1. **Ordinary Training**: Employing `CrossEntropyLoss` as the criterion and subsequently executing the standard conformal prediction process.
2. **Conformal Training**: Utilizing each predictor in the loss function as a criterion for training the model, followed by a comparative analysis of their conformal prediction outcomes.

Both approaches were developed through 5 training epochs to signify the last criterion prediction process.

## 3 Result and Discussion

### 3.1 Approach 1 (Ordinary Training)

In Approach 1, the model is trained using a basic loss function, namely `CrossEntropyLoss`.

#### 3.1.1 Overall Outcome

In the later tablula, we use 'cr' to denote `Coverage Rate` and 'as' to represent `Average Size`.

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.898 | **21.9872** | 0.9086 | 27.5674 | 0.8976 | 25.6884 | 0.8996 | 24.3844 |
| Cluster | 0.9014 | **22.9068** | 0.9012 | 26.9682 | 0.8918 | 24.935 | 0.9002 | 24.6876 |
| ClassWise | 0.9006 | **24.2726** | 0.8992 | 28.4864 | 0.9044 | 29.3974 | 0.9012 | 26.6778 |

Table 1: Table with Dataset Information: CIFAR100, Model: ResNet18, Size of Classes: 100

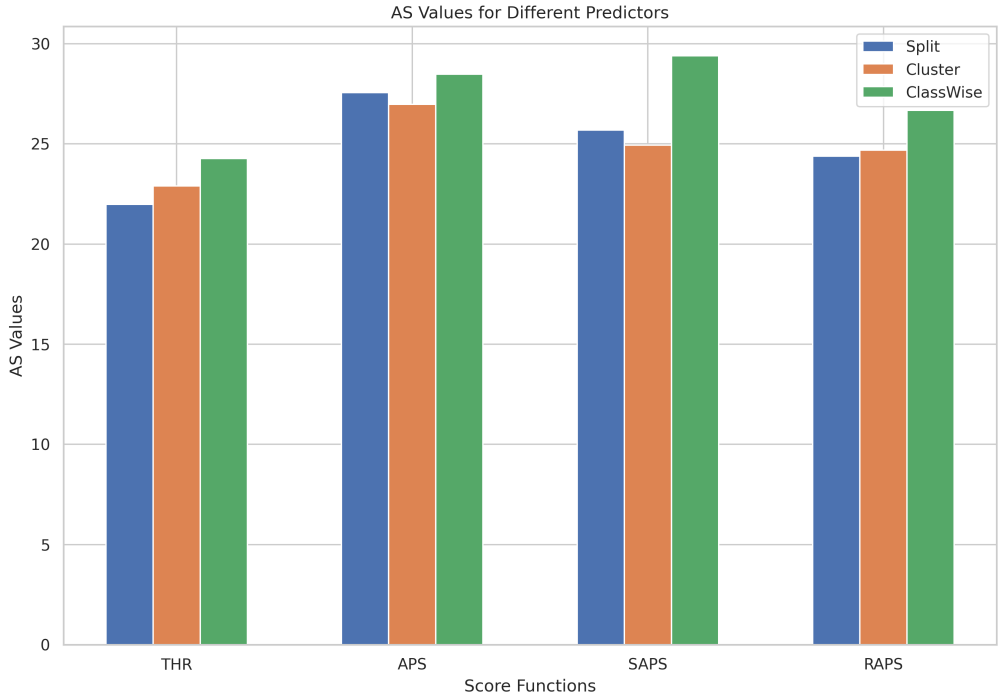


Figure 1: Average Size comparison of predictors and Score Functions on CIFAR100

Taking CIFAR-100 as an example, most other results generally exhibit a similar pattern (refer to Appendix section Approach 1).

According to Figure 1 Regarding the conformal prediction (CP) Score functions, THR demonstrated the best Average Size among all Score functions without compromising the Coverage Rate. SAPS and RAPS, as upgraded APS, exhibited marginal improvements in Average Size.

In terms of CP predictors, the Split and Cluster predictors yield similar results, while the ClassWise predictor generally performs the least favorably across all Score functions.

In summarizing the overall outcome, Approach 1 suggests that THR produces the most favorable results and exhibits stability across the majority of datasets. Additionally, both SAPS and RAPS showcase a notable improvement compared to APS.

However, in my experiments, the UTKFace_age test unveiled a significant and unexpected observation.

### 3.1.2 Interesting Findings

- **Cluster and ClassWise predictors exhibited poor performance on dense label dataset.**

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.8901 | 34.7039 | 0.8952 | 37.4626 | 0.8857 | 39.2993 | 0.8905 | 35.5149 |
| Cluster | **0.8553** | 29.9829 | **0.8686** | 36.6121 | **0.8494** | 35.1618 | **0.8705** | 34.8806 |
| ClassWise | 0.9154 | **64.5615** | 0.9034 | **67.483** | 0.9085 | **73.8958** | 0.9055 | **68.4482** |

Table 2: Table with Dataset Information: UTKFace_age, Model: ResNet18, Size of Classes: 103

UTKFace is a dataset designed for categorizing people's faces into various age groups, making it suitable for dense classification. The challenge lies in precisely aligning a specific single year of age with its corresponding label, especially compared to cases with more diverse labels. Conformal Prediction (CP) fits well in this scenario, as it involves predicting an age set or range, aligning seamlessly with the complexities of this task.

In our UTKFace results, some interesting patterns emerge, adding nuance to our overall observations. The Cluster predictor, while achieving the best Average Size, comes with an inevitable 3% to 5% decline in Coverage Rate—a trade-off that cannot be overlooked. It's crucial to mention that our chosen significance level (alpha) is set at 0.1. Additionally, the ClassWise predictor exhibits notably poor Average Size, exceeding 60, which is significantly worse than the Split predictor.

This indicates that both the Cluster and ClassWise predictors may not perform well in dense classification scenarios, potentially compromising either the Coverage Rate or Average Size.

## 3.2 Approach 2 (Conformal Training)

In Approach 2, the model undergoes Conformal training, employing a classification loss with various Score functions.

### 3.2.1 Overall Outcome

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.88825 | 2.93925 | 0.89 | 3.2495 | 0.9005 | 3.42925 | 0.893 | 3.20175 |
| Cluster | 0.88675 | 2.9225 | 0.89775 | 3.3395 | 0.8955 | 3.3485 | 0.90025 | 3.25875 |
| ClassWise | 0.894 | 3.00575 | 0.90225 | 3.38675 | 0.905 | 3.49 | 0.895 | 3.2295 |

Table 3: Table with Dataset Information: STL10 (Second Approach), Model: ResNet18, Size of Classes: 10

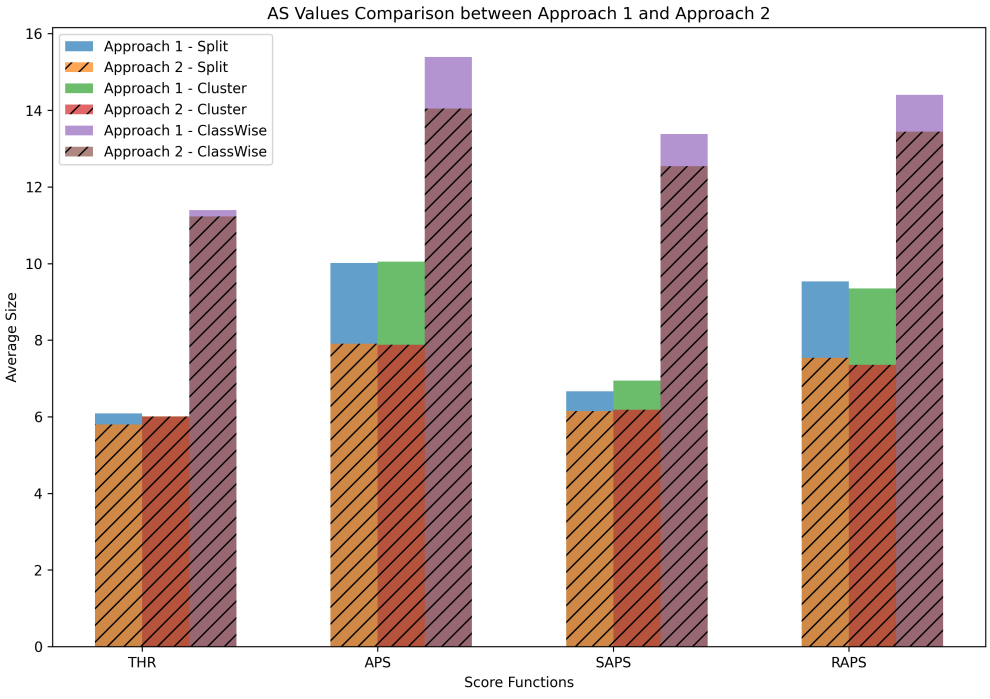


Figure 2: Stanford Dogs

Summarizing the outcomes of Approach 2 (refer to Appendix section Approach 2) poses a challenge. Overall, the performance of APS and THR on the Split and Cluster predictors shows improvement compared to Approach 1, as illustrated in Figure 2. Notably, the increment of APS is more substantial than that of THR, as depicted in the same figure. Although SAPS and RAPS exhibit improvement in

Figure 2, the observed stability issues in their performance often negate the demonstrated improvement of SAPS and RAPS over APS. Further exploration of this topic will be undertaken in the next subsection.

### 3.2.2 Interesting Findings

- **SAPS and RAPS may not match the performance of APS in conformal training.**

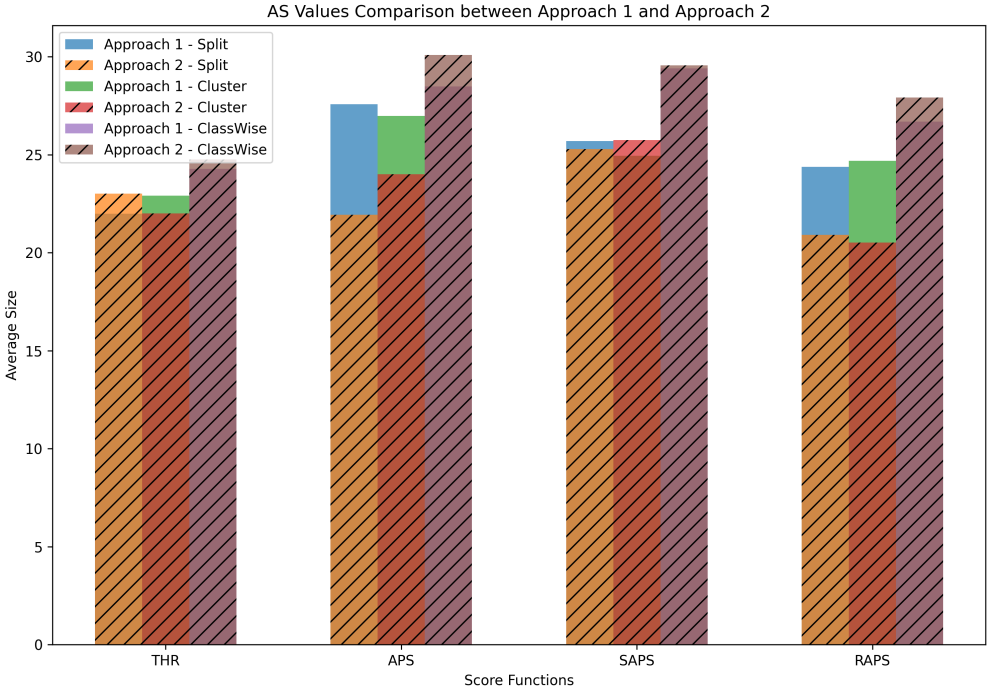


Figure 3: CIFAR100

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.8973 | 35.2982 | 0.9015 | **35.0287** | 0.8931 | **41.1166** | 0.8909 | **37.1881** |
| Cluster | 0.8580 | 29.7870 | 0.8549 | **31.6334** | 0.8595 | **37.9922** | 0.8515 | **32.3046** |
| ClassWise | 0.9068 | 64.6796 | 0.9072 | **68.2434** | 0.8931 | **74.6793** | 0.9017 | **69.7397** |

Table 4: Table with Dataset Information: UTKFace_age, Model: ResNet18, Size of Classes: 103 (Second Approach)

Although SPAS and RAPS are intended to be upgraded versions of the APS algorithm, their performance may not align with their intended goals in Conformal Training tasks. As highlighted in Table 9 and Figure 3, a compare of Ordinary training and Conformal training, SAPS exhibits inferior performance compared to APS on CIFAR100 and UTKFace. At the same time, While RAPS generally enhances its Average Size in Conformal Training of Approach 2, there are instances where its improvement is less than that of APS, leading to situations where its performance falls short of matching APS. This discrepancy is apparent in FashionMNIST Table 12 and UTKFace Table 4, where the performance of SAPS and RAPS fails to catch up with APS. This could be attributed to the selection of weight and penalty for SAPS and RAPS. However, Despite conducting additional experiments across a broader parameter range, this phenomenon persists (see Table 13 and Table 14). These observations directly implies that SAPS and RAPS do not guarantee an improvement over APS in Conformal Training tasks.

# 4 Conclusion

In our experimental observations, we used TorchCP to apply Conformal Prediction (CP) methods to a diverse range of datasets to obtain a comprehensive understanding of their performance.

THR consistently demonstrated superior performance across all approaches compared to other Score Functions within the same predictors. Notably, THR's excellence is independent of the training method, whether it be Ordinary Training or Conformal Training.

Classic Score Function APS exhibited noteworthy improvement when employed in Conformal Training as opposed to Ordinary Training. Its modified counterparts, SPAS and RPAS, showcased enhanced Average Size in the Ordinary Training approach when compared to APS. Interestingly, they became unstable and lost this advantage in the Conformal Training approach in our experiments settings.

Turning our attention to predictors, Cluster and ClassWise predictors exhibited a poor performance when applied to a dense class dataset like UTKFace, specifically in predicting human age. While Cluster predictor struggled to guarantee a satisfactory Coverage Rate, the ClassWise predictor failed to provide an acceptable Average Size.

In summary, it is advisable to avoid using Cluster or ClassWise predictors on dense class datasets. Opting for APS in Conformal training yields improved performance, and considering SAPS or RAPS in Ordinary training instead of APS can be beneficial for achieving better Average Size. However, caution is warranted when employing SAPS or RAPS in Conformal Training, as they may introduce instability. In most cases, using the THR Score Function and Split Predictor ensures a stable outcome.

# 5 Issues of TorchCP

TorchCP serves as a convenient Python toolbox for conducting conformal prediction research on deep learning models. The integration of the latest conformal prediction methods simplifies the usage of the most advanced models in the field of CP, making it a one-tap solution. While under construction, it may have some issues that need to be addressed:

- The SAPS and RAPS require the input of a weight and a penalty parameter, respectively. They do not have initial values, making initialization a necessity. However, the optimal values are not arbitrary. It is advisable to define a range or provide an optimal initial value to facilitate easier implementation of these methods, saving time that would otherwise be spent reading corresponding papers to determine suitable values. The values used in the experiments mentioned in the corresponding papers fall within the range of weight **(0.02, 0.6)** and penalty **(0.001, 0.01)**. However, refraining from initialization can, on the other hand, draw attention to the key parameters of SAPS and RAPS, as an error is raised when they are lacking. To address this, it is advisable to include a notice specifying the optimal range for these two parameters for convenience.

- In the example file `conformal_training.py`, within the main method, there is an unused parameter `'num_trials = 5'`, which appears to be intended for testing purposes. It should be removed.

# References


Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. (2020). Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223.

Ding, T., Angelopoulos, A. N., Bates, S., Jordan, M. I., and Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*.

Huang, J., Xi, H., Zhang, L., Yao, H., Qiu, Y., and Wei, H. (2023). Conformal prediction for deep classifier via label ranking.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2016). Distribution-free predictive inference for regression. *arXiv preprint arXiv:1604.04173*.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc.

Sadinle, M., Lei, J., and Wasserman, L. (2016). Least ambiguous set-valued classifiers with bounded error levels.

Shi, F., Ong, C. S., and Leckie, C. (2013). Applications of class-conditional conformal predictor in multi-class classification. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 235–239.

Stutz, D., Dvijotham, K. D., Cemgil, A. T., and Doucet, A. (2022). Learning optimal conformal classifiers. In *International Conference on Learning Representations*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Zhang, Z., Song, Y., and Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

# A Approach 1

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.907 | 3.0725 | 0.90225 | 3.3225 | 0.91025 | 3.28325 | 0.91025 | 3.315 |
| Cluster | 0.907 | 3.07225 | 0.91125 | 3.36225 | 0.915 | 3.451 | 0.90875 | 3.3505 |
| ClassWise | 0.9125 | 3.095 | 0.9075 | 3.267 | 0.91275 | 3.337 | 0.91475 | 3.35275 |

Table 5: Table with Dataset Information: STL10, Model: ResNet18, Size of Classes: 10

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9094 | 2.556 | 0.9064 | 2.8016 | 0.9128 | 2.8096 | 0.91 | 2.7476 |
| Cluster | 0.9106 | 2.574 | 0.9038 | 2.7534 | 0.9102 | 2.7892 | 0.9038 | 2.7076 |
| ClassWise | 0.9174 | 2.6576 | 0.9026 | 2.7866 | 0.9146 | 2.8816 | 0.9046 | 2.7596 |

Table 6: Table with Dataset Information: CIFAR10, Model: ResNet18, Size of Classes: 10

| | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| Predictor | cr | as | cr | as | cr | as | cr | as |
| Split | 0.8964 | 6.0892 | 0.9032 | 10.0168 | 0.8962 | 6.6666 | 0.9032 | 9.5284 |
| Cluster | 0.8934 | 6.007 | 0.9002 | 10.0512 | 0.8944 | 6.9418 | 0.9042 | 9.3494 |
| ClassWise | 0.8962 | 11.3976 | 0.8986 | 15.3916 | 0.9024 | 13.3782 | 0.9048 | 14.4064 |

Table 7: Table with Dataset Information: Stanford Dogs, Model: ResNet18, Size of Classes: 120

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9008 | 1.0046 | 0.9008 | 1.2418 | 0.9008 | 1.24 | 0.9012 | 1.2442 |
| Cluster | 0.8992 | 0.9992 | 0.901 | 1.2478 | 0.8894 | 1.216 | 0.9034 | 1.2342 |
| ClassWise | 0.903 | 1.0858 | 0.9052 | 1.2794 | 0.8868 | 1.2268 | 0.8968 | 1.2516 |

Table 8: Table with Dataset Information: FashionMNIST, Model: ResNet18, Size of Classes: 10

# B Approach 2

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9096 | 23.009 | 0.9004 | 21.9244 | 0.8998 | 25.2818 | 0.9006 | 20.9052 |
| Cluster | 0.898 | 21.9954 | 0.9094 | 23.9964 | 0.9114 | 30.0862 | 0.9016 | 20.516 |
| ClassWise | 0.904 | 24.76 | 0.9026 | 29.5632 | 0.9056 | 27.919 | 0.9016 | 20.516 |

Table 9: Table with Dataset Information: CIFAR100, Model: ResNet18, Size of Classes: 100 (Second Approach)

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9048 | 2.4992 | 0.9012 | 2.7776 | 0.904 | 2.7122 | 0.8936 | 2.6474 |
| Cluster | 0.904 | 2.591 | 0.8942 | 2.6396 | 0.909 | 2.7614 | 0.8952 | 2.7272 |
| ClassWise | 0.9034 | 2.5746 | 0.9022 | 2.8156 | 0.9004 | 2.8152 | 0.9068 | 2.8106 |

Table 10: Table with Dataset Information: CIFAR10, Model: ResNet18, Size of Classes: 10 (Second Approach)

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.8944 | 5.8014 | 0.901 | 7.9042 | 0.899 | 6.1452 | 0.8968 | 7.541 |
| Cluster | 0.895 | 6.009 | 0.8934 | 7.8804 | 0.8912 | 6.1794 | 0.8968 | 7.3554 |
| ClassWise | 0.9018 | 11.2272 | 0.8996 | 14.0406 | 0.9012 | 12.5366 | 0.9014 | 13.4384 |

Table 11: Table with Dataset Information: Stanford Dogs, Model: ResNet18, Size of Classes: 120 (Second Approach)

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.902 | 1.0052 | 0.9156 | 1.2806 | 0.8976 | 1.2124 | 0.9052 | 1.2292 |
| Cluster | 0.9058 | 1.0084 | 0.9014 | 1.094 | 0.8954 | 1.2 | 0.903 | 1.2416 |
| ClassWise | 0.9014 | 1.094 | 0.9064 | 1.264 | 0.8994 | 1.2438 | 0.8998 | 1.2798 |

Table 12: Table with Dataset Information: FashionMNIST, Model: ResNet18, Size of Classes: 10 (Second Approach)

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9097 | 36.3978 | 0.9129 | 35.2052 | 0.9104 | 41.9772 | 0.9074 | 42.3501 |
| Cluster | 0.8711 | 31.7954 | 0.8724 | 32.4271 | 0.8618 | 39.9637 | 0.8608 | 34.0160 |
| ClassWise | 0.9019 | 65.7022 | 0.9055 | 68.3438 | 0.9049 | 71.3568 | 0.9074 | 69.9857 |

Table 13: Table with Dataset Information: UTKFace_age, Model: ResNet18, Size of Classes: 103 (Second Approach), **Especially, SAPS weight: 0.02, PAPS penalty: 0.003**

| Predictor | THR | | APS | | SAPS | | RAPS | |
|---|---|---|---|---|---|---|---|---|
| | cr | as | cr | as | cr | as | cr | as |
| Split | 0.9144 | 38.0542 | 0.9087 | 35.1304 | 0.9129 | 42.5838 | 0.9123 | 43.4503 |
| Cluster | 0.8791 | 32.3733 | 0.8839 | 35.8477 | 0.8661 | 36.9692 | 0.8785 | 39.1036 |
| ClassWise | 0.9057 | 64.8511 | 0.8945 | 67.7872 | 0.8996 | 74.2605 | 0.8960 | 73.4486 |

Table 14: Table with Dataset Information: UTKFace_age, Model: ResNet18, Size of Classes: 103 (Second Approach), **Especially, SAPS weight: 0.6, PAPS penalty: 0.01**