|Summary|

**Background**

In our experiment, we tested and compared nine algorithms, categorized into three features (without labels, with labels, and with weighted labels), to find which ones would provide the best classification for social media data such as tweets from Hurricanes Sandy and Maria. These features were each divided further into three subclasses to represent the targets (ground truth, majority voting with labels, and majority voting with weighted labels) and thus yielding the nine total cases. We recorded the accuracy and AUC (Area Under the Curve) while comparing with the ground truth and majority voting; this gives us four datasets per tested algorithm [with the exception of two cases (2) and (3)]. Note that “majority voting” can be abbreviated as MV and that “ground truth” can be abbreviated as GT.

* Algorithm **(1)**: [Feature] **Weighted Labels**  [Target] **Ground Truth**
* Algorithm **(2)**: [Feature] **Labels** [Target] **Ground Truth**
* Algorithm **(3)**: [Feature] **Without Labels** [Target] **Ground Truth**
* Algorithm **(4)**: [Feature] **Weighted Labels** [Target] **Majority Voting (Weighted Labels)**
* Algorithm **(5)**: [Feature] **Labels** [Target] **Majority Voting (Weighted Labels)**
* Algorithm **(6)**: [Feature] **Without Labels** [Target] **Majority Voting (Weighted Labels)**
* Algorithm **(7)**: [Feature] **Weighted Labels** [Target] **Majority Voting (Labels)**
* Algorithm **(8)**: [Feature] **Labels**  [Target] **Majority Voting (Labels)**
* Algorithm **(9)**: [Feature] **Without Labels** [Target] **Majority Voting (Labels)**

**Results**

1) *Effect of Labeling*: (1), (2) v. (3) | (4), (5) v. (6) | (7), (8) v. (9)

After the executions, we first sorted the produced data into its three distinct targets (ground truth, majority voting with labels, and majority voting with weighted labels as seen above) to clearly observe the effect of adding labels and weighted labels. In each set, algorithms that were modified with either type of labels as features [(1), (2); (4), (5); (7), (8)] performed much better in test trials than those that had no labels as features [(3), (6), (9)] in all instances. Although in some cases algorithms with no labels [(6), (9)] as the feature proved to be better in all training trials than the other algorithms in their sets when compared to the original ground truth, the discrepancies are only marginal (such differences tended to be ≈0.02, as shown in both Figs. 1 and 2). Thus, these exceptions should hold minimal leverage when determining which algorithm is better for training. Note that bolded values represent the highest value in the corresponding place in the tables. Hence, adding some type of label to the feature is needed to increase the accuracy and AUC. Because of this, we have ruled that (3), (6), and (9) are less ideal for maximizing accuracy and AUC when compared to the other six algorithms.

2) *Labels v. Weighted Labels*: (1) v. (2) | (4) v. (5) | (7) v. (8)

**\*Set 1: (1), (2), (3)** | **Set 2: (4), (5), (6)** | **Set 3: (7), (8), (9)\***

After taking out (3), (6), (9), we now compare the differences of using labels and weighted labels as the feature as according to their original set shown above.

In Set 1, (1) runs the same quality of performance compared with (2) in both training datasets measuring accuracy and AUC [(2) does not have comparisons with majority voting] when it is tested against the ground truth. In other words, both algorithms have five cases better than the other in terms of both accuracy and AUC. As seen in Fig. 3, for accuracy and Sandy during training, (1) is better when the number of labelers, *x* ∈ {7, 9}; Maria when *x* ∈ {3, 5, 11}. For AUC and Sandy during training, (1) is better when *x* ∈ {7, 9}; Maria when *x* ∈ {3, 5, 11}. Note that *x* represents the number of labelers and that bolded values represent the higher value in the corresponding place in the tables. Moreover, keeping the same format as above, (1) executes slightly better than (2) in the test trials. (1) works better since it is better in six cases for both accuracy and AUC compared with (2)’s four better cases in each. This can be seen in Fig. 4, for accuracy and Sandy, (1) is better when *x* ∈ {3, 5, 9, 11}; Maria when *x* ∈ {7, 9}. For AUC and Sandy during training, (1) is better when *x* ∈ {3, 5, 9, 11}; Maria when *x* ∈ {7, 9}. Statistically, (1) has an edge over (2), but since the difference is so slight, we cannot distinguish the better algorithm for higher accuracy and AUC values. With this being said, in Set 1, there is no dominant advantage in selecting labels over weighted labels and vice versa.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(4)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89412 | 0.90902 | 0.91259 | 0.91375 | 0.91658 | 0.88676 | 0.90386 | 0.9078 | 0.90877 | 0.91168 |
| Sandy | 0.84283 | 0.86732 | 0.87724 | 0.8835 | 0.89271 | 0.83362 | 0.85989 | 0.87084 | 0.87716 | 0.8873 |
| **(5)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89966 | 0.90885 | 0.90979 | 0.91521 | 0.91654 | 0.89496 | 0.90374 | 0.90466 | 0.91024 | 0.91161 |
| Sandy | 0.86248 | 0.86949 | 0.88477 | 0.88622 | 0.89234 | 0.85547 | 0.86273 | 0.87901 | 0.8803 | 0.88697 |
| **(6)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.9174** | **0.92255** | **0.92436** | **0.92521** | **0.92895** | **0.91862** | **0.92533** | **0.92668** | **0.92753** | **0.93117** |
| Sandy | **0.88364** | **0.8905** | **0.89285** | **0.90216** | **0.90508** | **0.8809** | **0.88778** | **0.8901** | **0.90002** | **0.90296** |

Fig. 1. a. Accuracy and AUC comparisons for Set 2 training trials compared to ground truth.

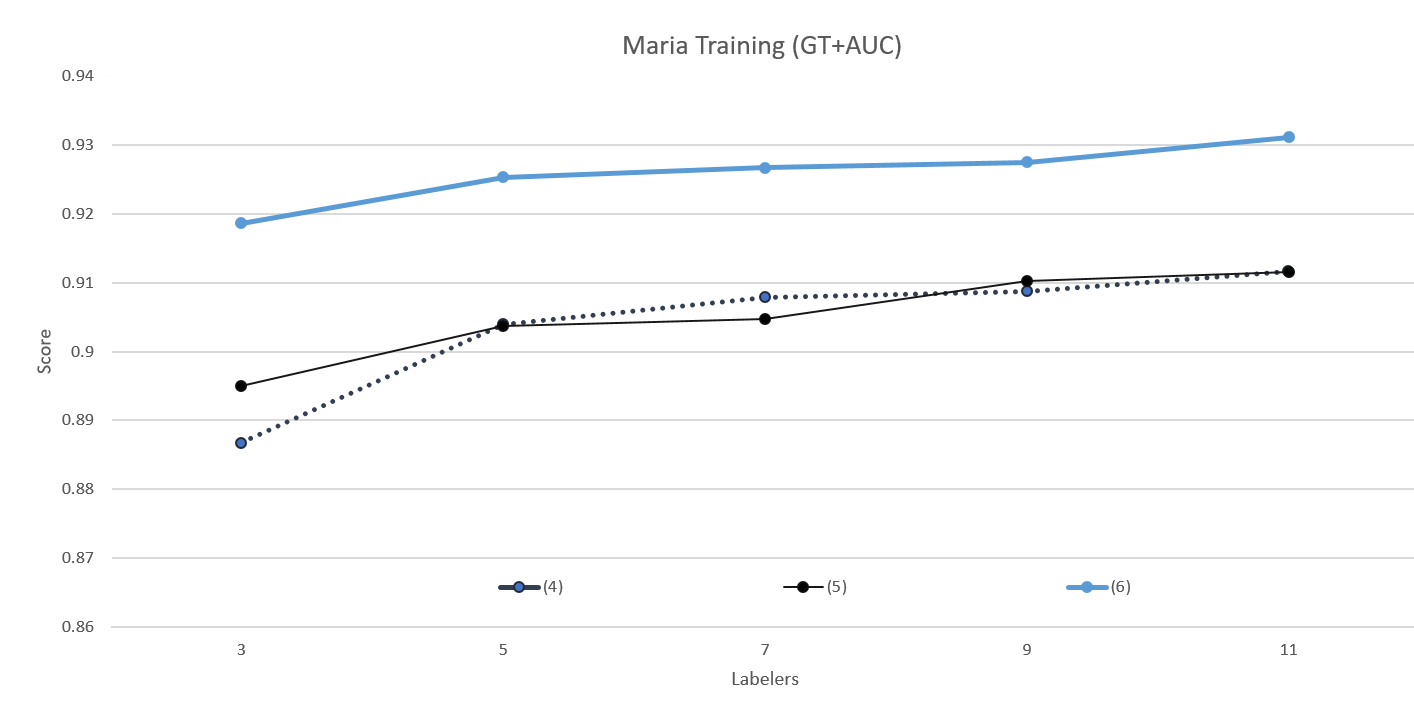


Fig. 1. b. AUC comparisons for Set 2 Maria training trials compared to ground truth.

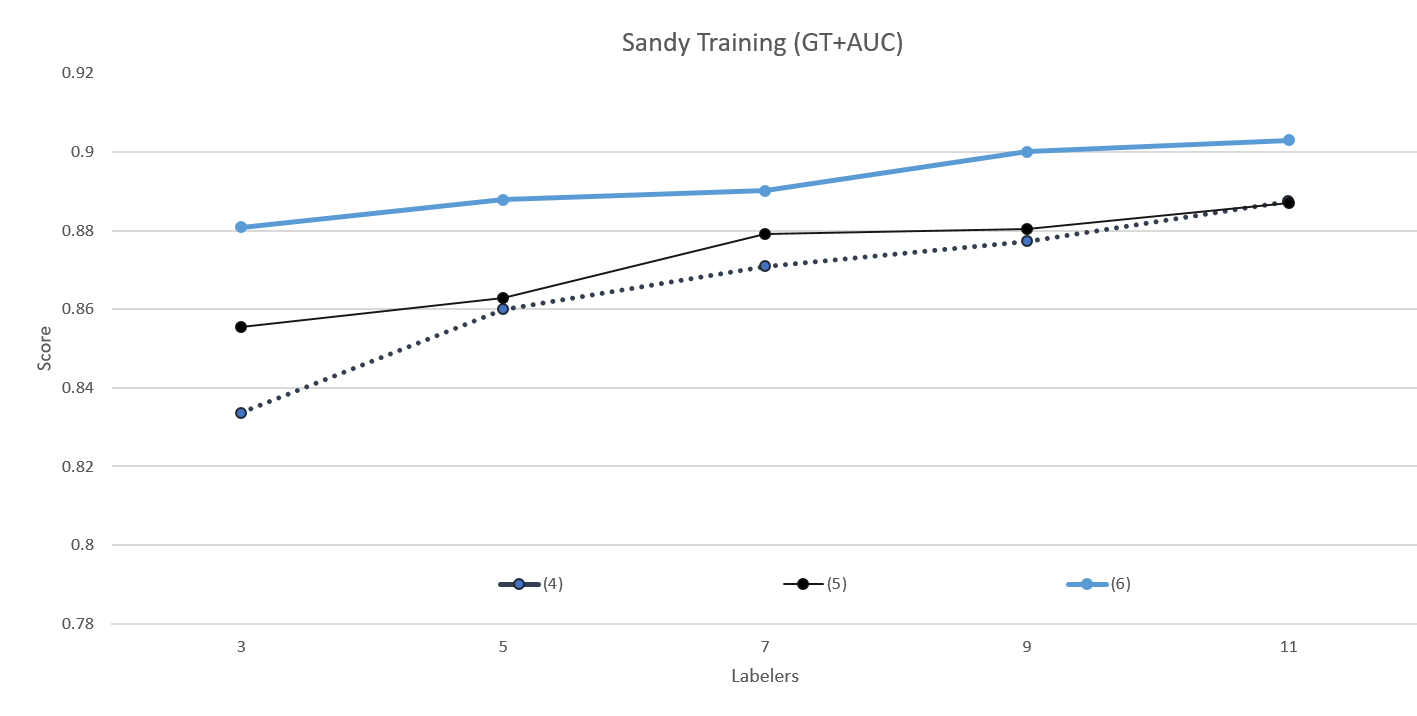


Fig. 1. c. AUC comparisons for Set 2 Sandy training trials compared to ground truth.

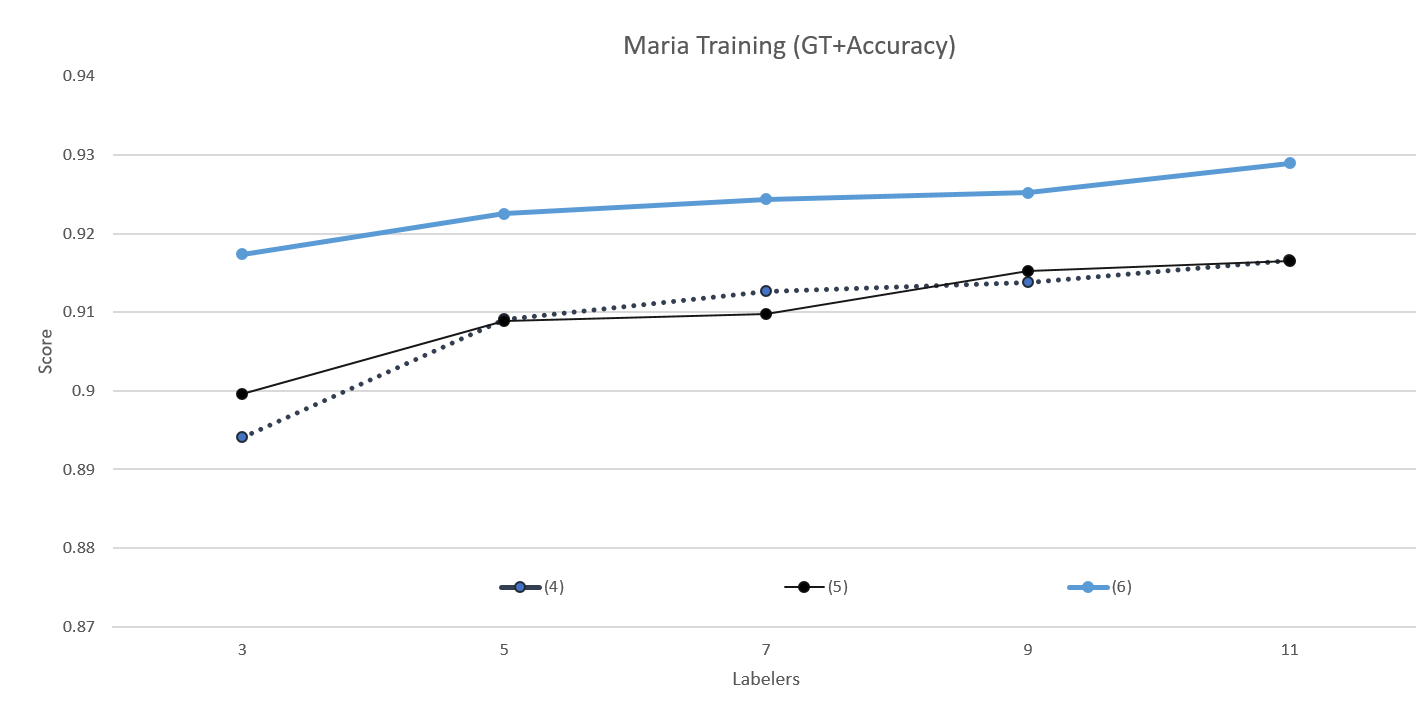


Fig. 1. d. Accuracy comparisons for Set 2 Maria training trials compared to ground truth.

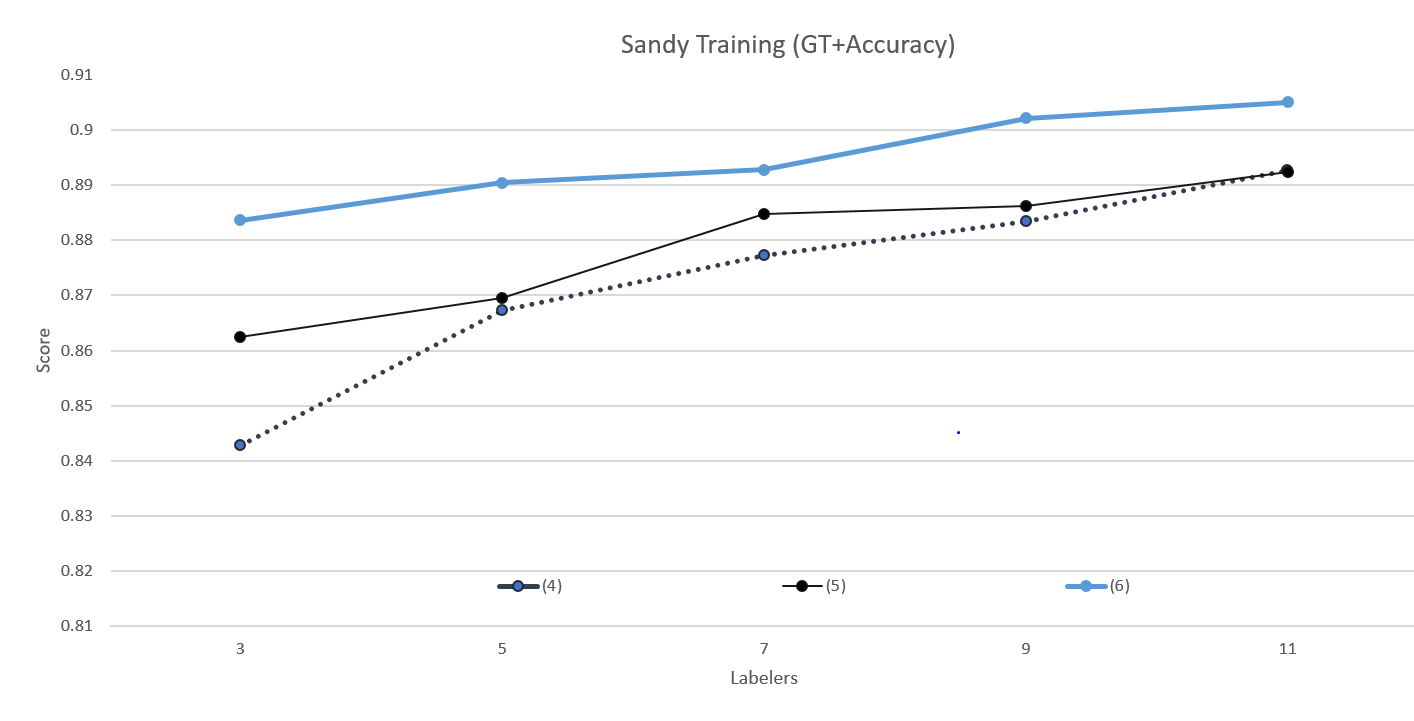


Fig. 1. e. Accuracy comparisons for Set 2 Sandy training trials compared to ground truth.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(7)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89674 | 0.90786 | 0.91349 | 0.91555 | 0.9171 | 0.89029 | 0.90267 | 0.90818 | 0.91087 | 0.9123 |
| Sandy | 0.85792 | 0.86676 | 0.87626 | 0.88764 | 0.89262 | 0.85068 | 0.85942 | 0.86968 | 0.88177 | 0.88729 |
| **(8)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89781 | 0.90966 | 0.91138 | 0.91551 | 0.9168 | 0.89202 | 0.90449 | 0.90596 | 0.91072 | 0.91181 |
| Sandy | 0.85651 | 0.86653 | 0.87729 | 0.88256 | 0.89149 | 0.84871 | 0.85904 | 0.87079 | 0.87631 | 0.88605 |
| **(9)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.91637** | **0.92208** | **0.92436** | **0.92564** | **0.92865** | **0.91826** | **0.92492** | **0.92701** | **0.92812** | **0.93089** |
| Sandy | **0.87677** | **0.87579** | **0.89281** | **0.89314** | **0.90367** | **0.8724** | **0.87107** | **0.88987** | **0.88993** | **0.90122** |

Fig. 2. a. Accuracy and AUC comparisons for Set 3 training trials compared to ground truth.

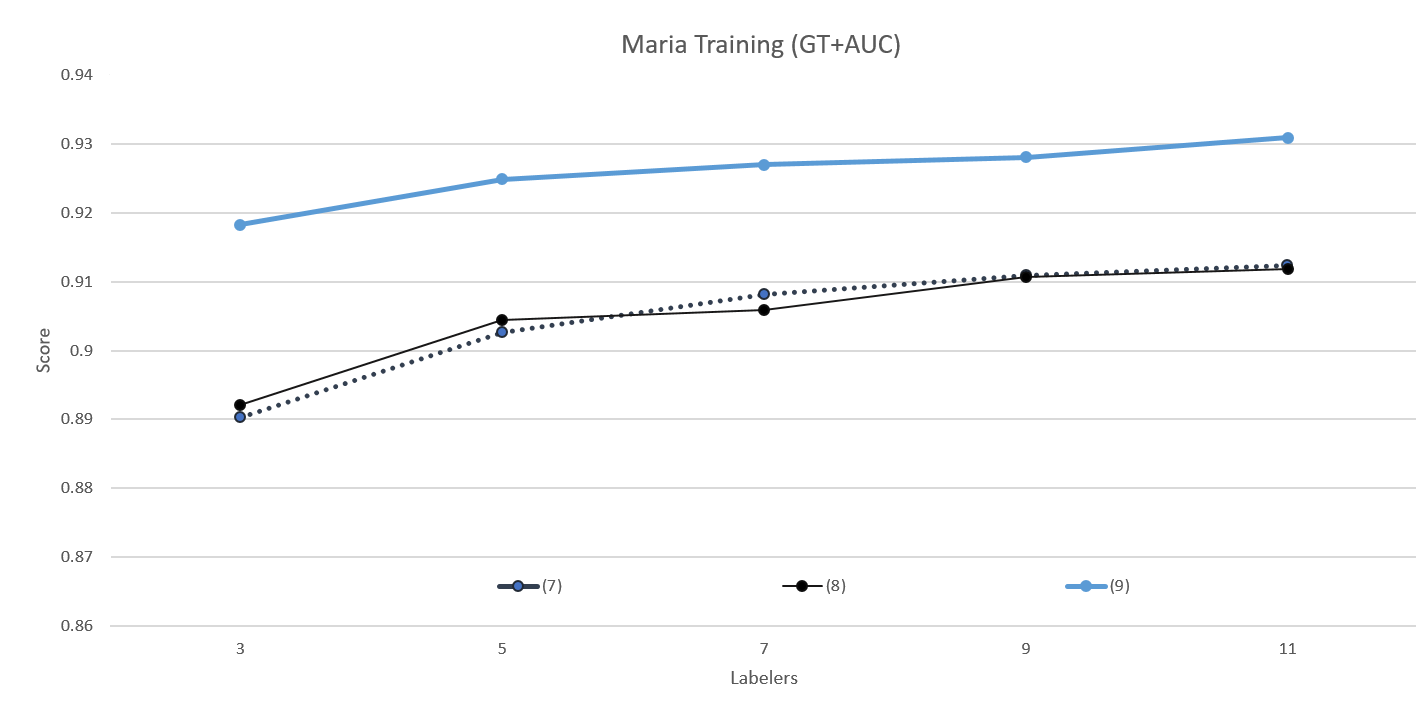


Fig. 2. b. AUC comparisons for Set 3 Maria training trials compared to ground truth.

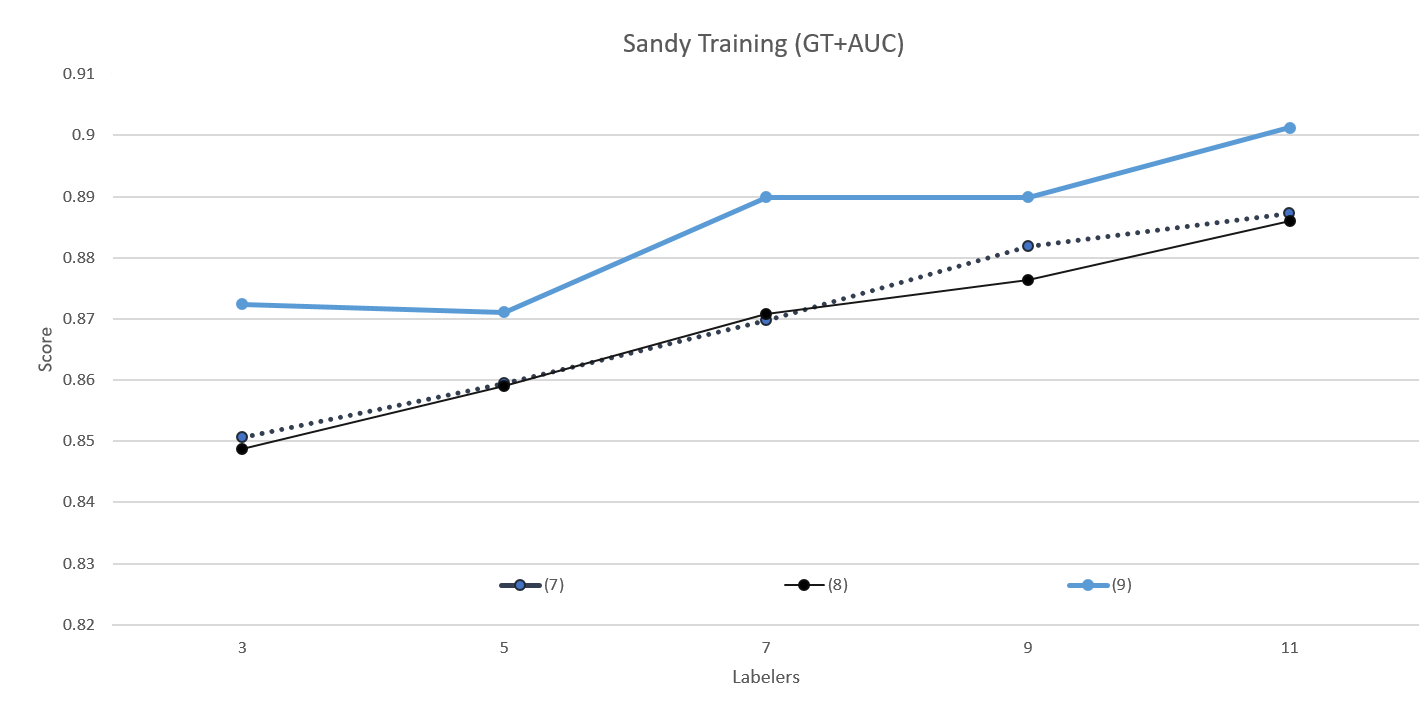


Fig. 2. c. AUC comparisons for Set 3 Sandy training trials compared to ground truth.

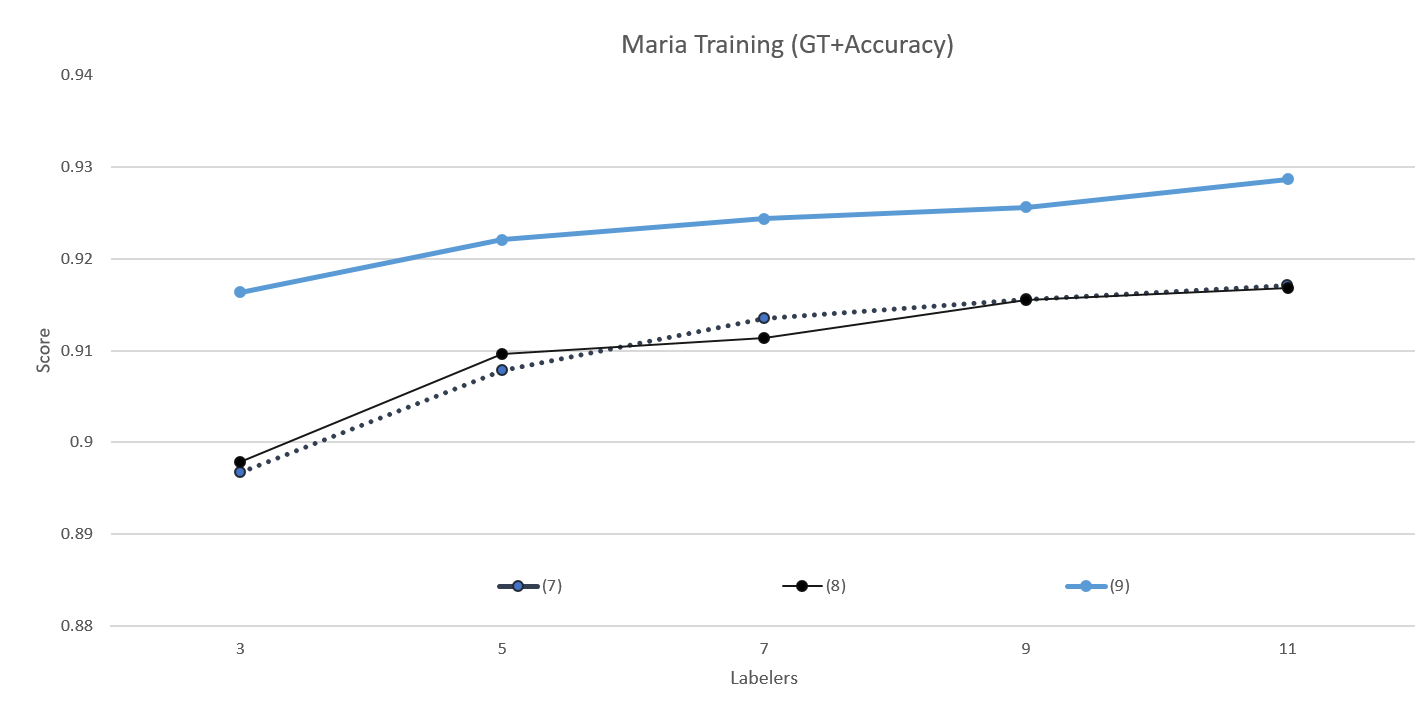


Fig. 2. d. Accuracy comparisons for Set 3 Maria training trials compared to ground truth.

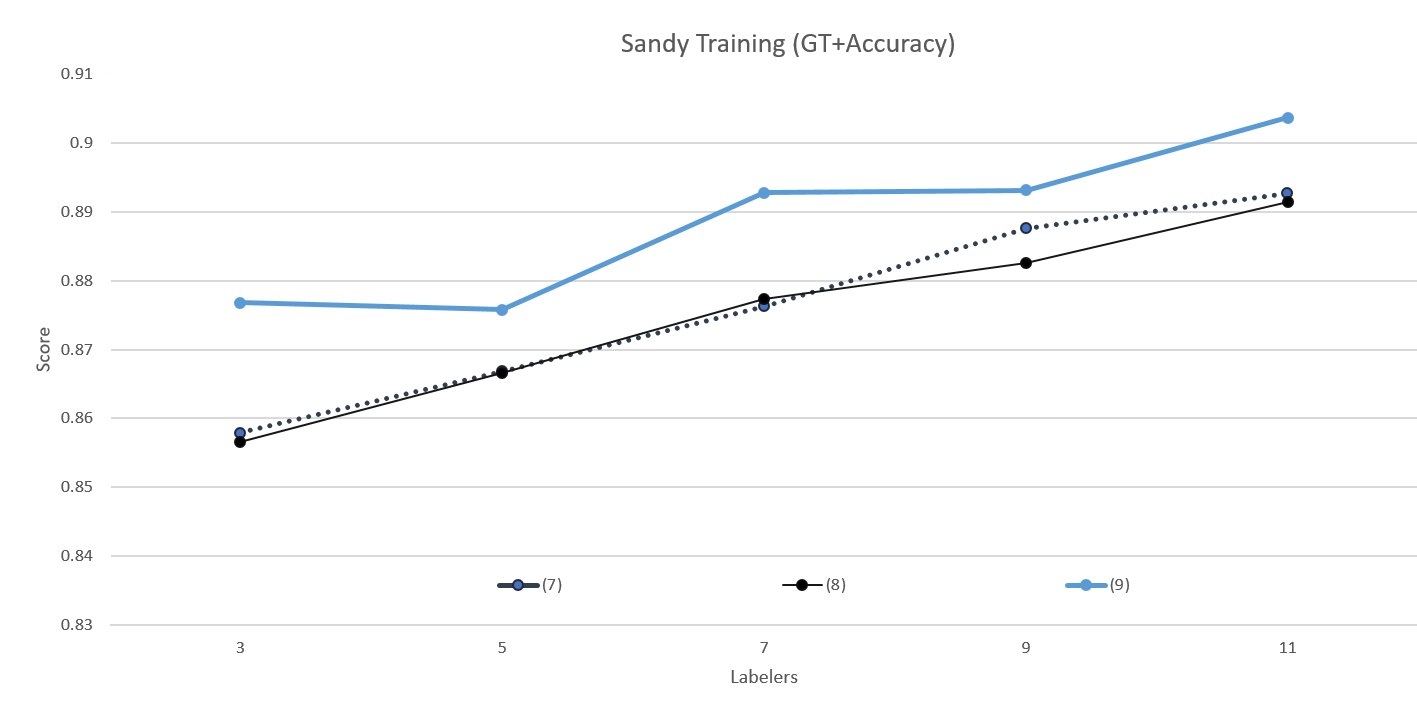


Fig. 2. e. Accuracy comparisons for Set 3 Sandy training trials compared to ground truth.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(1)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.98458** | **0.98677** | 0.98694 | 0.9881 | **0.98935** | **0.98649** | **0.98841** | 0.98856 | 0.98958 | **0.99067** |
| Sandy | 0.97988 | 0.98237 | **0.98392** | **0.985** | 0.9843 | 0.98128 | 0.9836 | **0.98504** | **0.98605** | 0.98539 |
| **(2)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.98368 | 0.98621 | **0.98724** | **0.98857** | 0.98913 | 0.9857 | 0.98792 | **0.98883** | **0.98999** | 0.99048 |
| Sandy | **0.98082** | **0.98256** | 0.98387 | 0.98481 | **0.98463** | **0.98215** | **0.98377** | 0.985 | 0.98587 | **0.9857** |

Fig. 3. a. (1) and (2) accuracy and AUC comparisons for Set 1 training trials compared to ground truth.

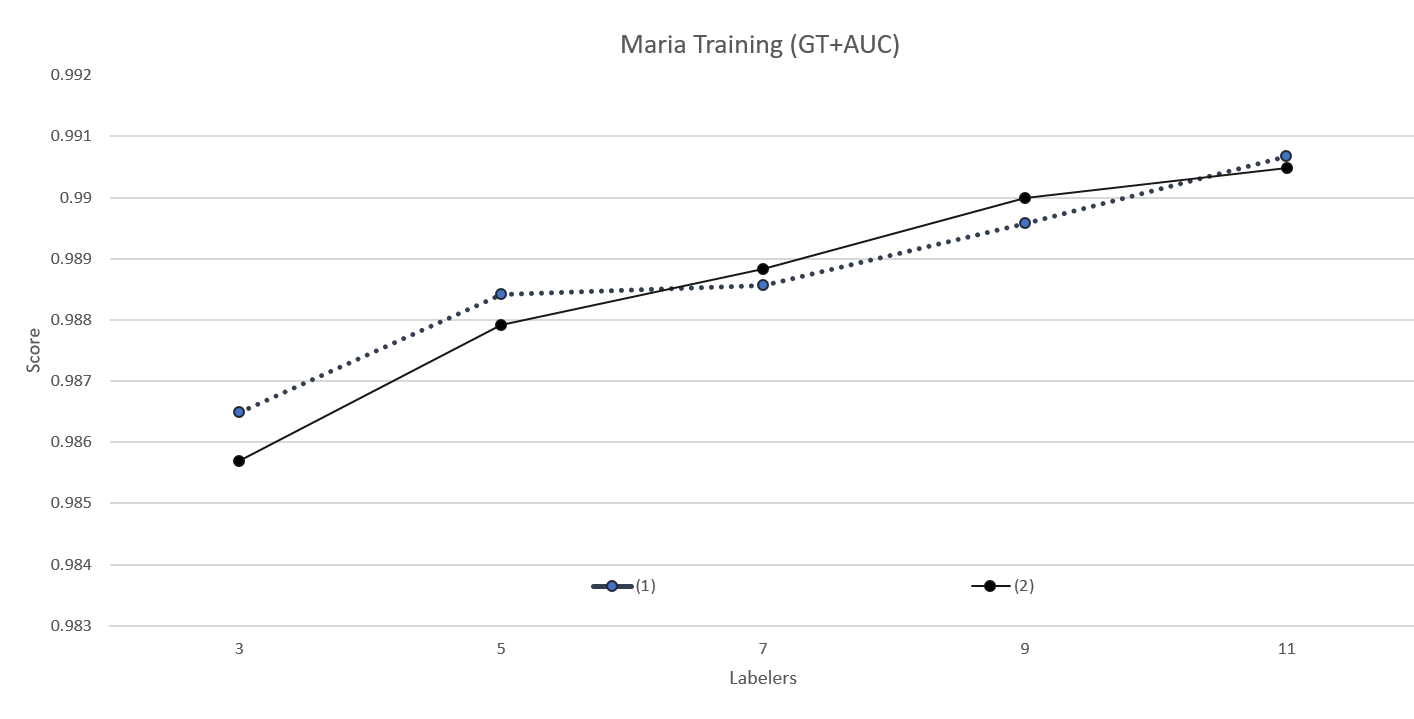


Fig. 3. b. (1) and (2) AUC comparisons for Set 1 Maria training trials compared to ground truth.

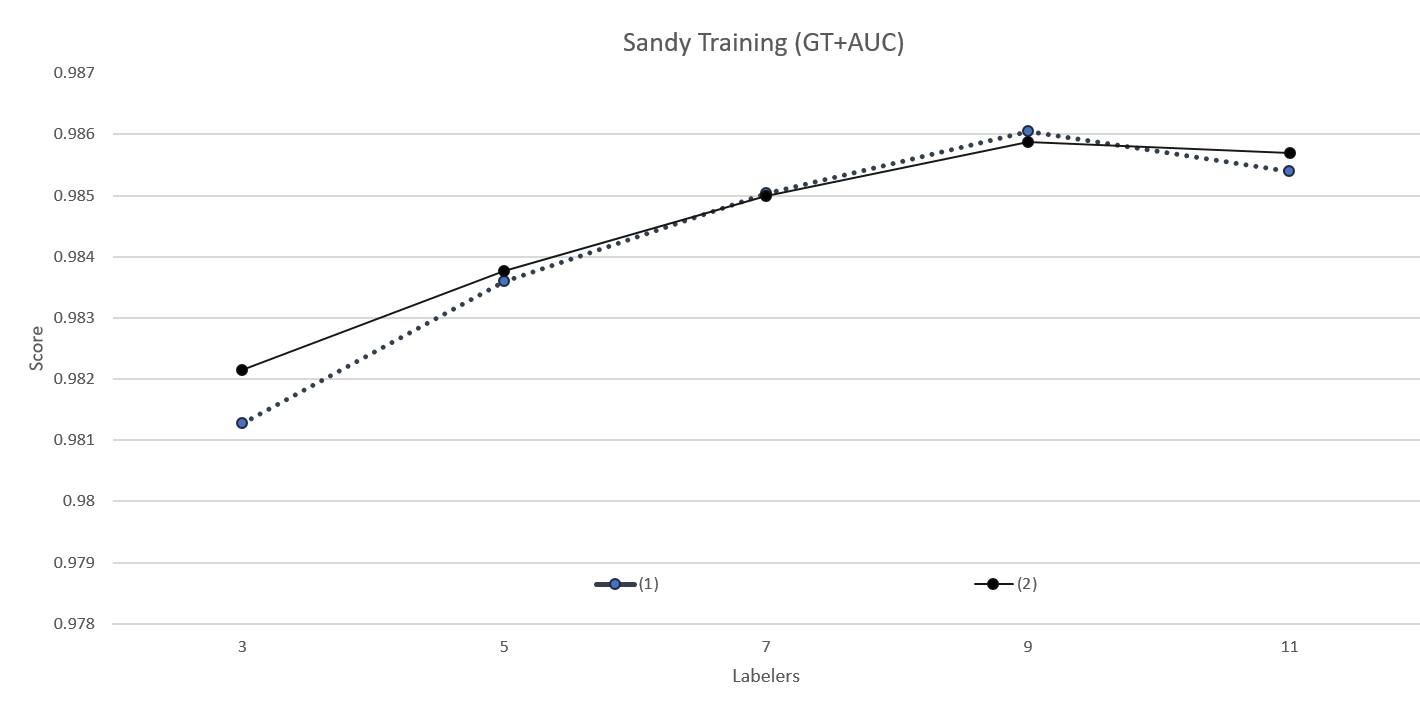


Fig. 3. c. (1) and (2) AUC comparisons for Set 1 Sandy training trials compared to ground truth.

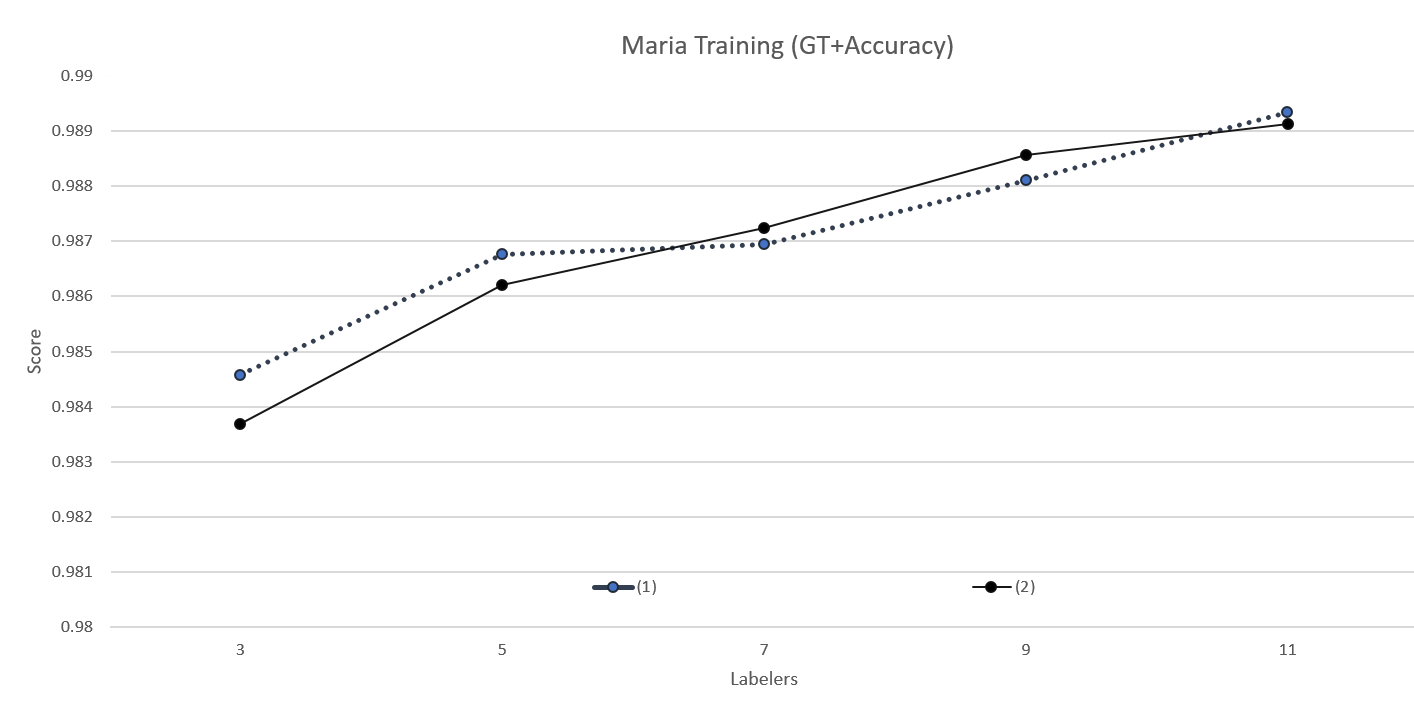


Fig. 3. d. (1) and (2) accuracy comparisons for Set 1 Maria training trials compared to ground truth.

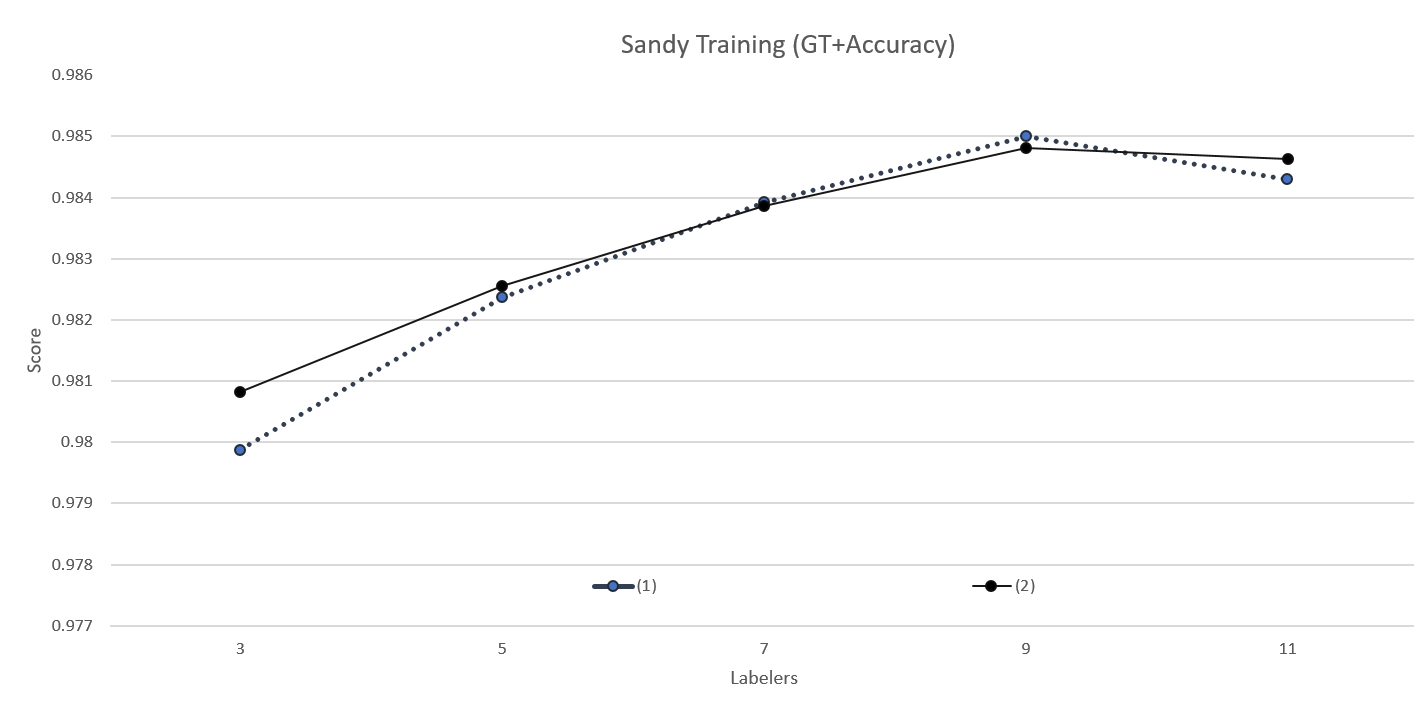


Fig. 3. e. (1) and (2) accuracy comparisons for Set 1 Sandy training trials compared to ground truth.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(1)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.8701 | 0.8732 | **0.87938** | **0.87938** | 0.87887 | 0.8795 | 0.88228 | **0.88917** | **0.88966** | 0.88888 |
| Sandy | **0.89738** | **0.89738** | 0.89588 | **0.90169** | **0.89869** | **0.90245** | **0.9019** | 0.90079 | **0.90514** | **0.90213** |
| **(2)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.87165** | **0.87595** | 0.87663 | 0.87801 | **0.87973** | **0.88073** | **0.88494** | 0.8864 | 0.88814 | **0.89001** |
| Sandy | 0.89232 | 0.89288 | **0.90075** | 0.89963 | 0.89813 | 0.89651 | 0.89711 | **0.90518** | 0.90312 | 0.9017 |

Fig. 4. a. (1) and (2) accuracy and AUC comparisons for Set 1 testing trials compared to ground truth.

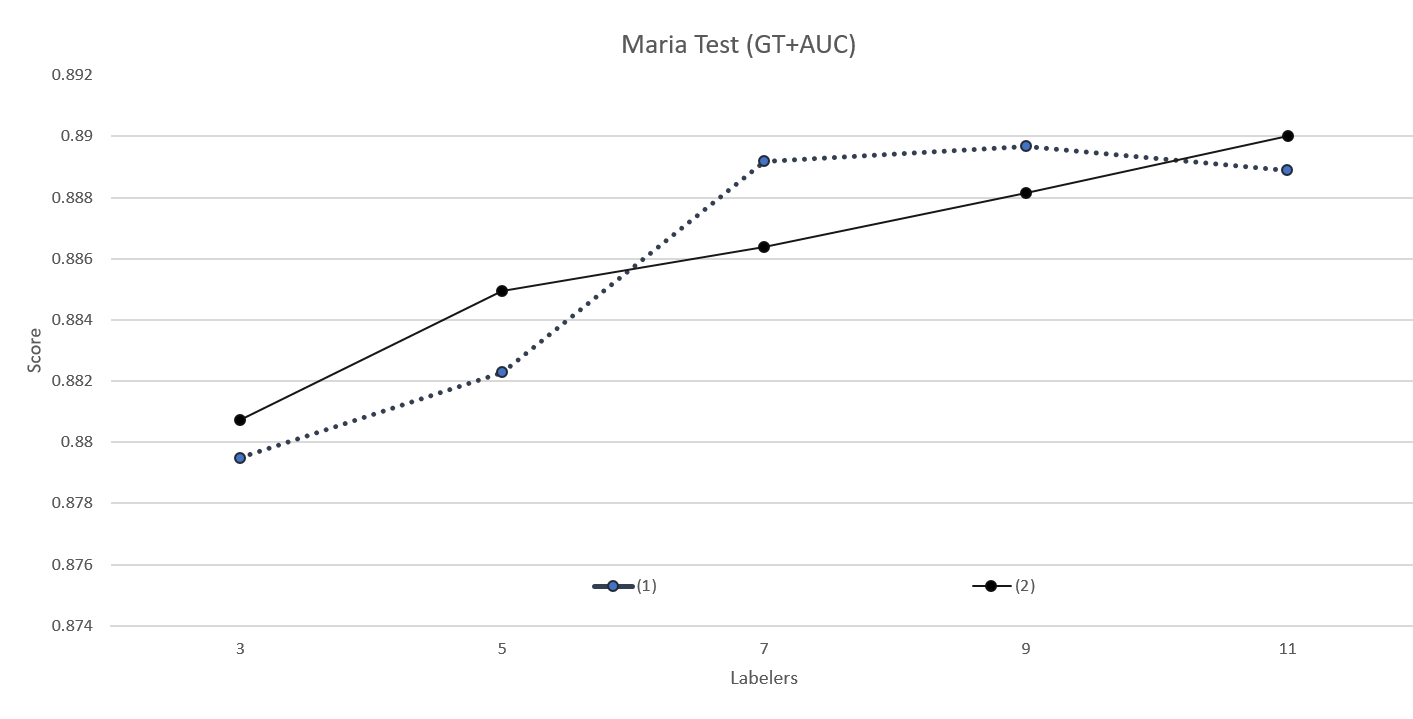


Fig 4. b. (1) and (2) AUC comparisons for Set 1 Maria testing trials compared to ground truth.

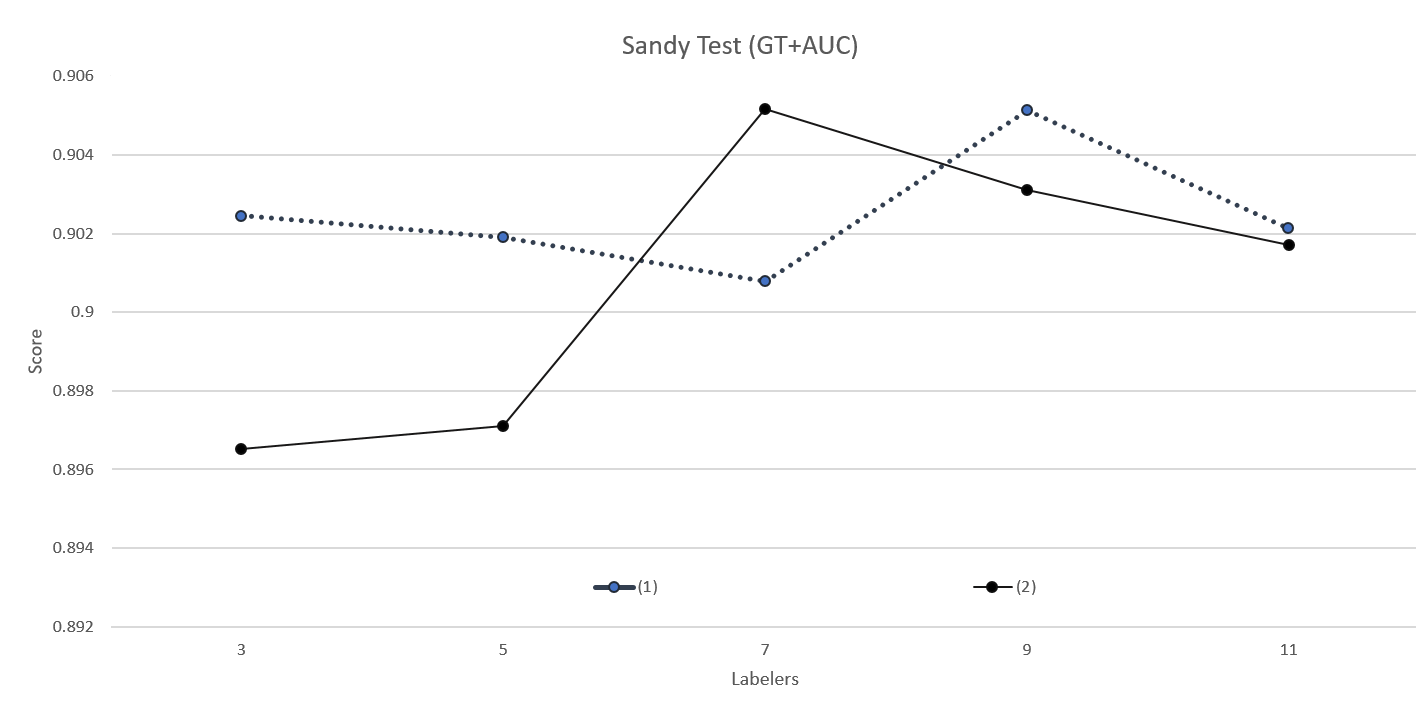


Fig 4. c. (1) and (2) AUC comparisons for Set 1 Sandy testing trials compared to ground truth.

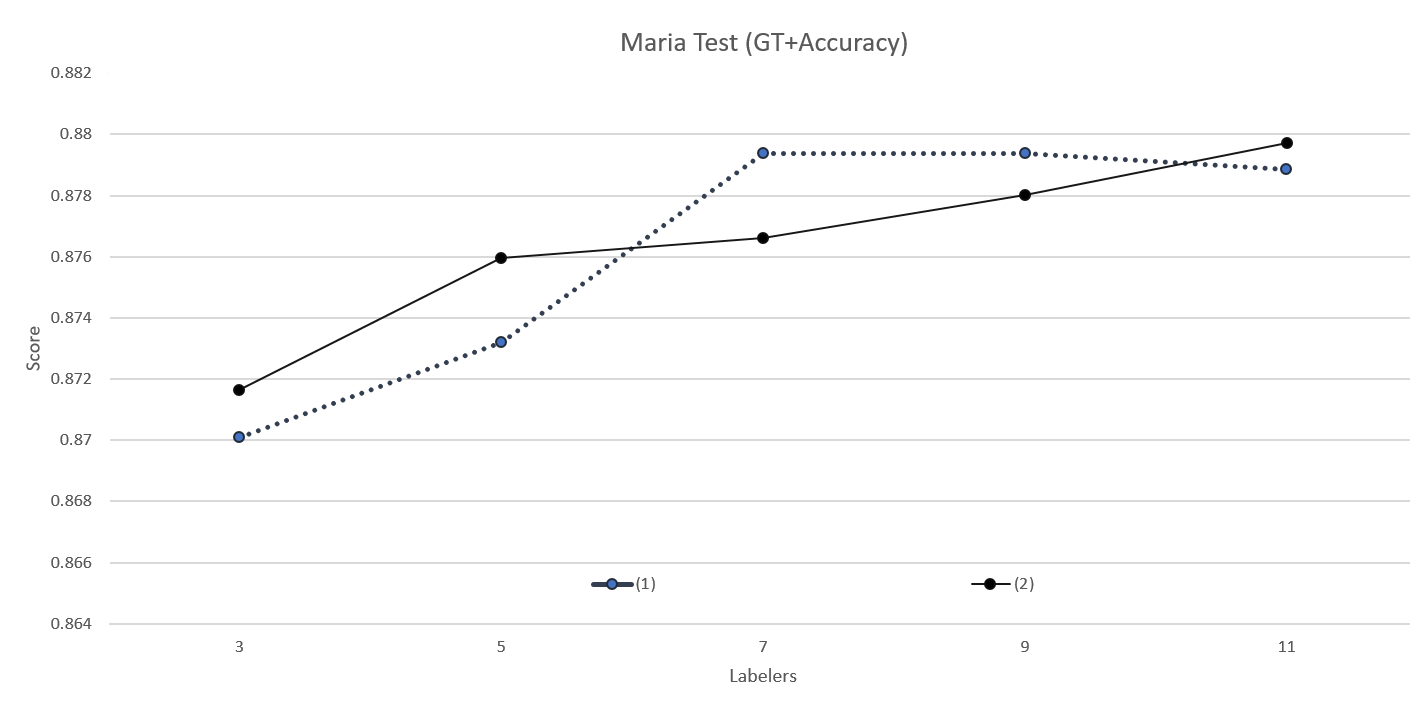


Fig 4. d. (1) and (2) accuracy comparisons for Set 1 Maria testing trials compared to ground truth.

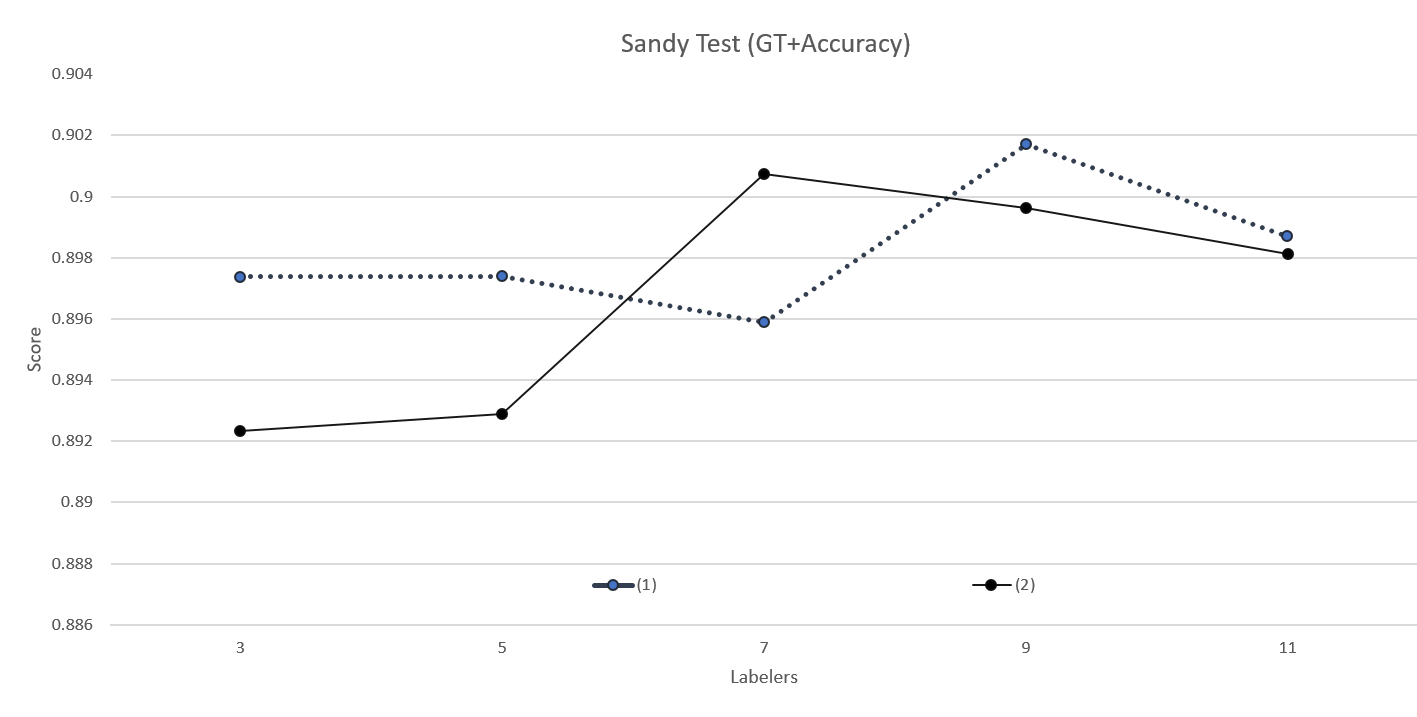


Fig 4. e. (1) and (2) accuracy comparisons for Set 1 Sandy testing trials compared to ground truth.

In Set 2, (4) runs better for both AUC and accuracy than (5) in the cases when compared to the majority voting. It can be seen in Fig. 5 that (4) performs better in five cases in both accuracy and AUC (there are three ties for accuracy and two for AUC) compared to (5)’s two better for accuracy and three better for AUC. Note that the bolded terms represent the higher value in the corresponding points in the tables. However, (5) runs slightly better in training for both AUC and accuracy when compared to the ground truth as it is better in six cases for each compared to (4)’s four better cases in each as seen in Fig. 6. Overall for training, (4) has a slight edge over (5) with 18 total instances when it is better, compared to (5)’s 17. Consequently, because of the small difference between (4) and (5), it is not clear whether adding weighted labels or labels is better for training. Furthermore in the test cases, (4) performs worse than (5). When compared to the ground truth, (4) performs better in eight cases compared to (5)’s twelve better cases; when compared to majority voting, (4) also performs worse with eight cases better to (5)’s twelve. Overall in Fig. 7, it can be seen that (4) has 16 better cases and (5) has 24 better cases – (5) has 50% more percentage values higher than (4). Ultimately, from observation of both training and testing datasets in Set 2, (5) works better than (4) and thus suggests that using labels as the feature is a better choice for increasing the accuracy and AUC of classification than using weighted labels.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(4)** |  |  |  |  |  |  |  |  |  |  |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 1 | 0.99974 | 0.99979 | **0.99991** | 0.99996 | 1 | 0.99978 | 0.99982 | **0.99993** | 0.99996 |
| Sandy | 0.99991 | **0.99986** | **0.99953** | **0.99944** | **0.99953** | 0.99992 | **0.99988** | **0.99961** | **0.99952** | **0.99961** |
| **(5)** |  |  |  |  |  |  |  |  |  |  |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 1 | **1** | 0.99979 | 0.99983 | **1** | 1 | **1** | 0.99982 | 0.99986 | **1** |
| Sandy | 0.99991 | 0.99962 | 0.9992 | 0.99925 | 0.99906 | **0.99993** | 0.99968 | 0.99933 | 0.99937 | 0.99921 |

Fig. 5. a. (4) and (5) accuracy and AUC comparisons for Set 2 training trials compared to majority voting.

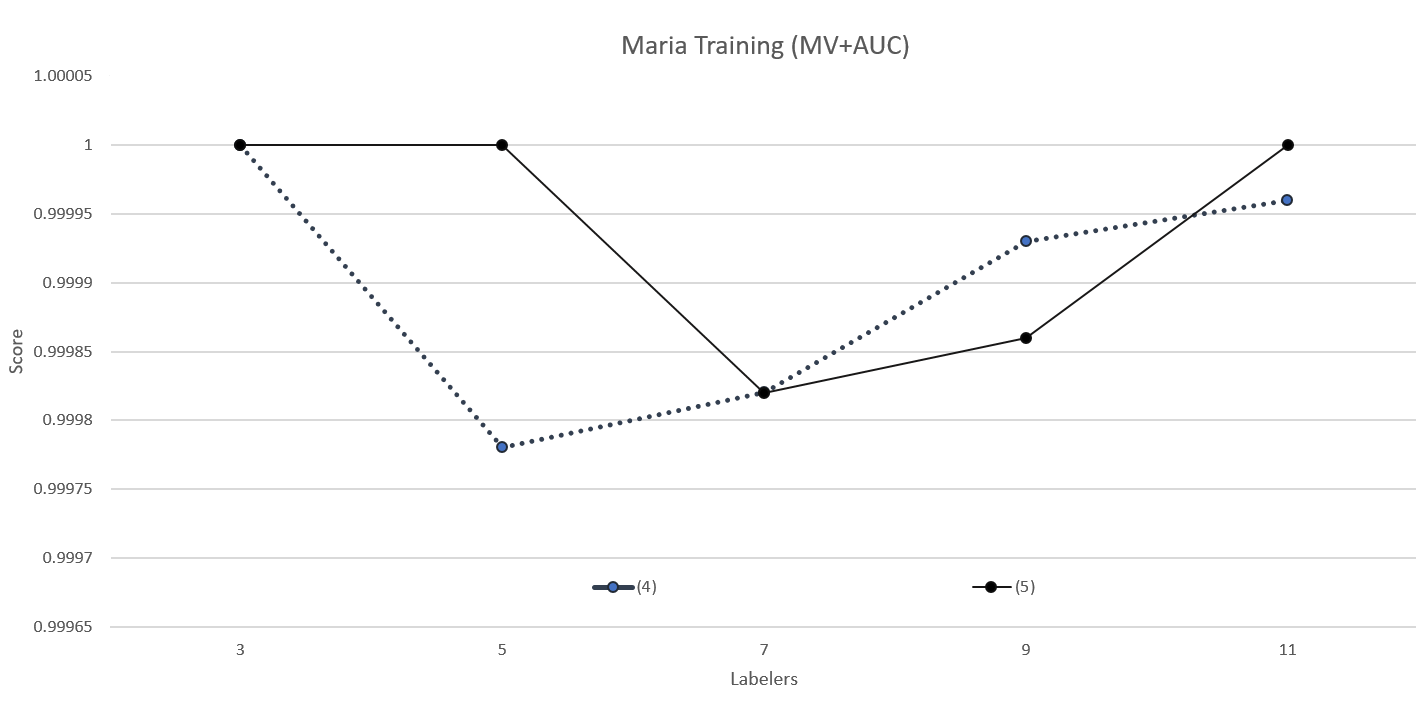


Fig. 5. b. (4) and (5) AUC comparisons for Set 2 Maria training trials compared to majority voting.

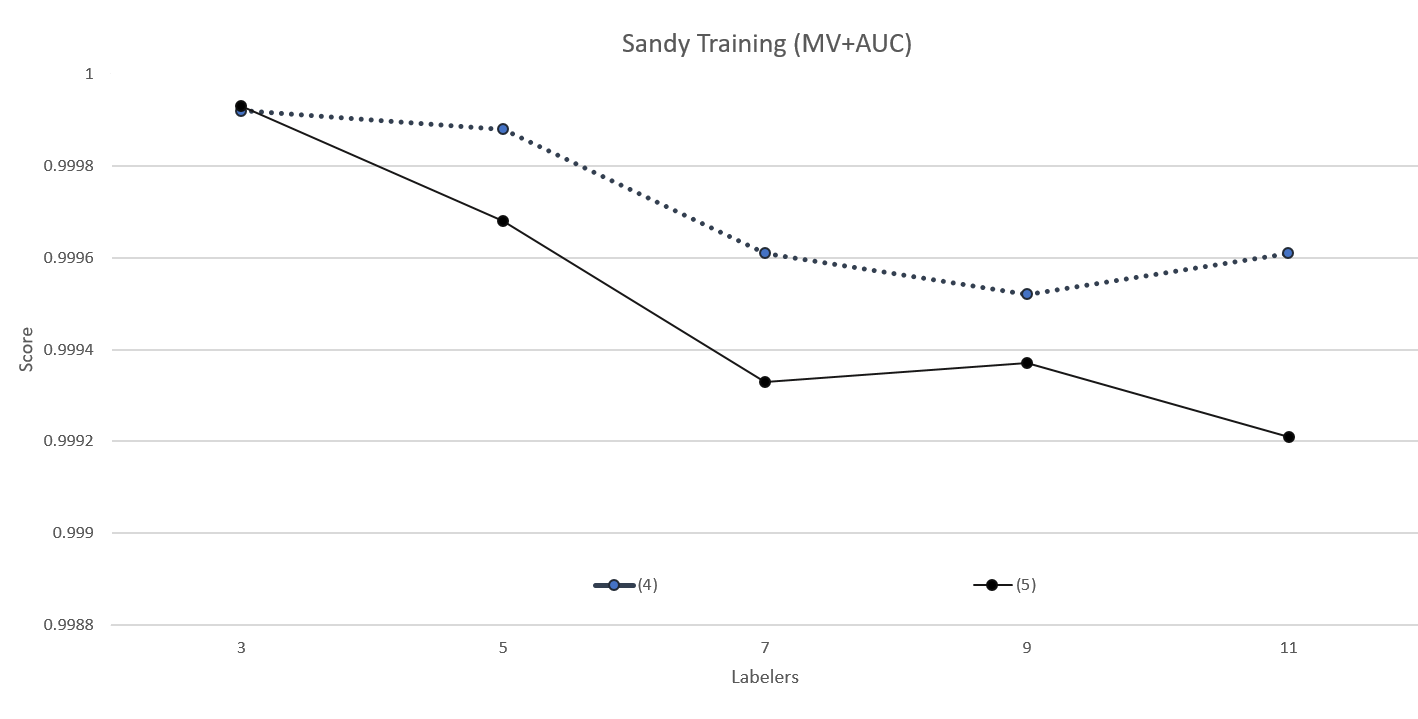


Fig. 5. c. (4) and (5) AUC comparisons for Set 2 Sandy training trials compared to majority voting.

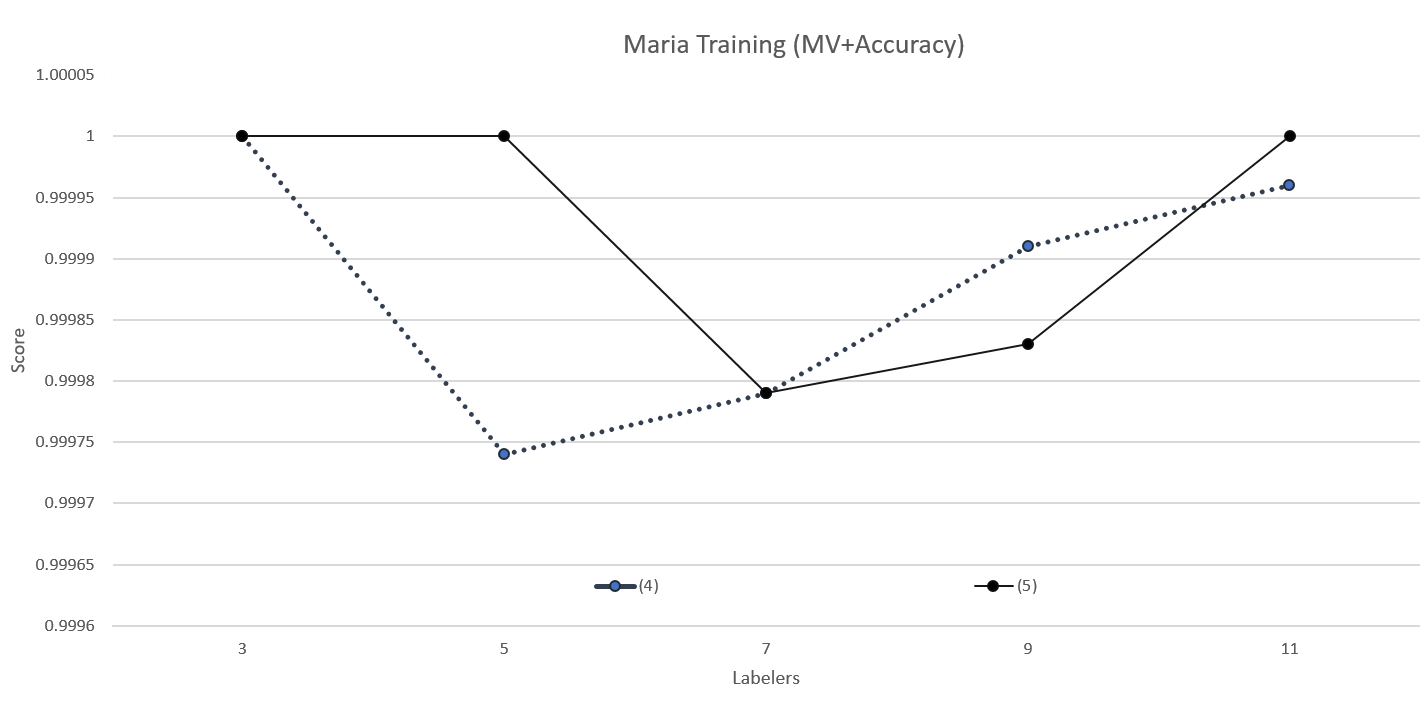


Fig. 5. d. (4) and (5) accuracy comparisons for Set 2 Maria training trials compared to majority voting.

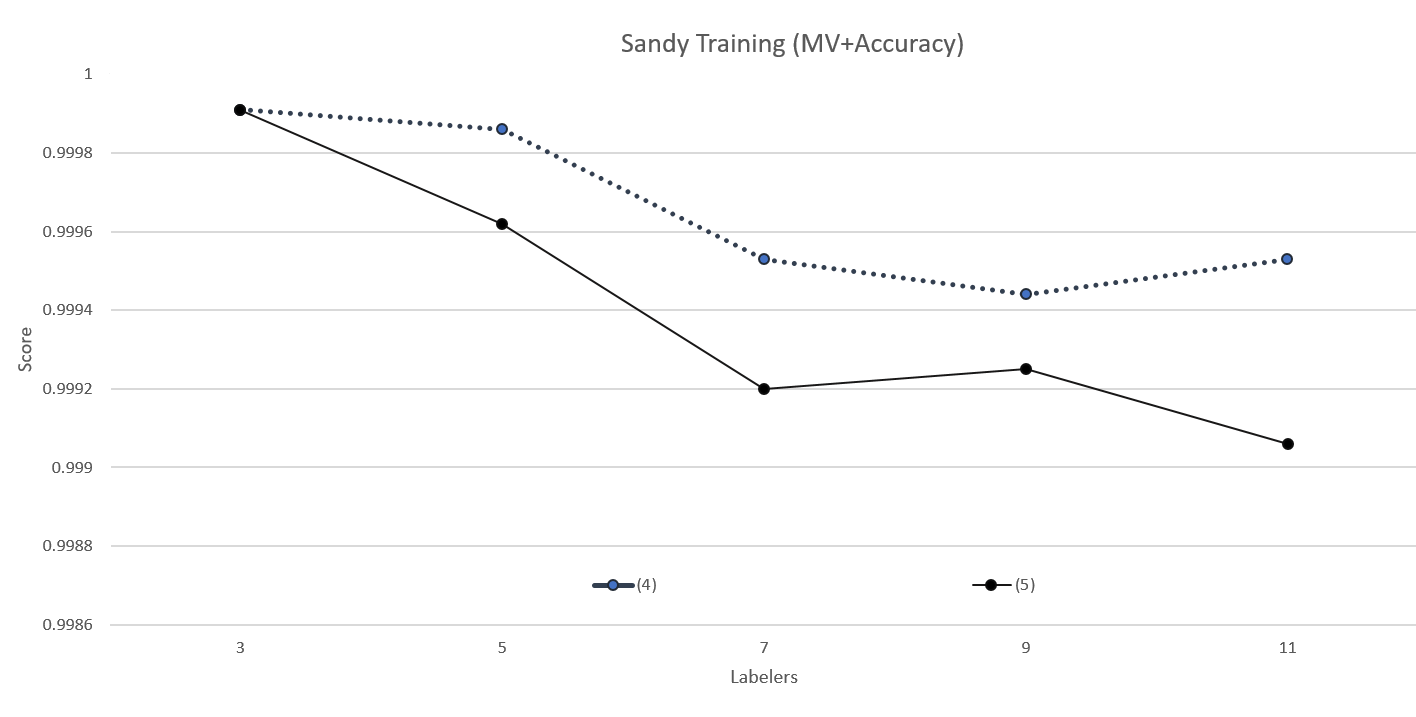


Fig. 5. e. (4) and (5) accuracy comparisons for Set 2 Sandy training trials compared to majority voting.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(4)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89412 | **0.90902** | **0.91259** | 0.91375 | **0.91658** | 0.88676 | **0.90386** | **0.9078** | 0.90877 | **0.91168** |
| Sandy | 0.84283 | 0.86732 | 0.87724 | 0.8835 | **0.89271** | 0.83362 | 0.85989 | 0.87084 | 0.87716 | **0.8873** |
| **(5)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.89966** | 0.90885 | 0.90979 | **0.91521** | 0.91654 | **0.89496** | 0.90374 | 0.90466 | **0.91024** | 0.91161 |
| Sandy | **0.86248** | **0.86949** | **0.88477** | **0.88622** | 0.89234 | **0.85547** | **0.86273** | **0.87901** | **0.8803** | 0.88697 |

Fig. 6. a. (4) and (5) accuracy and AUC comparisons for Set 2 training trials compared to ground truth.

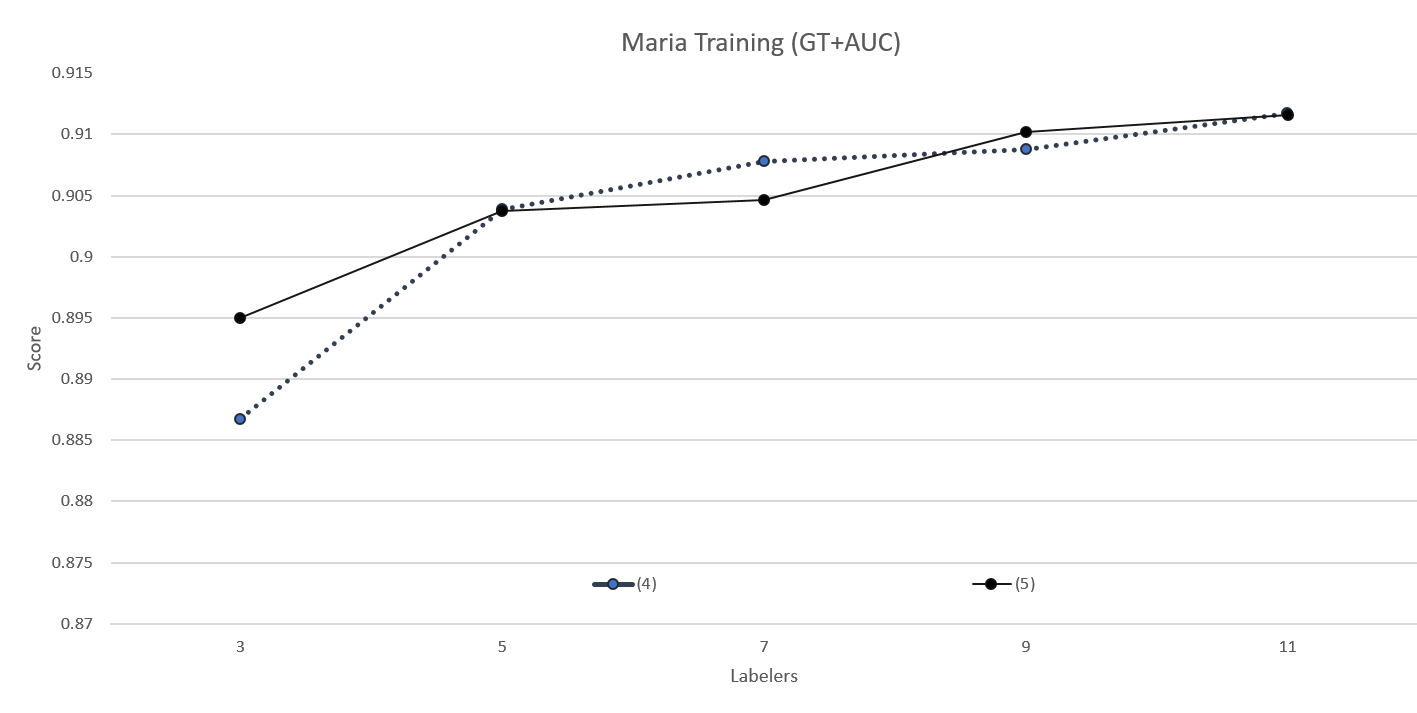


Fig. 6. b. (4) and (5) AUC comparisons for Set 2 Maria training trials compared to ground truth.

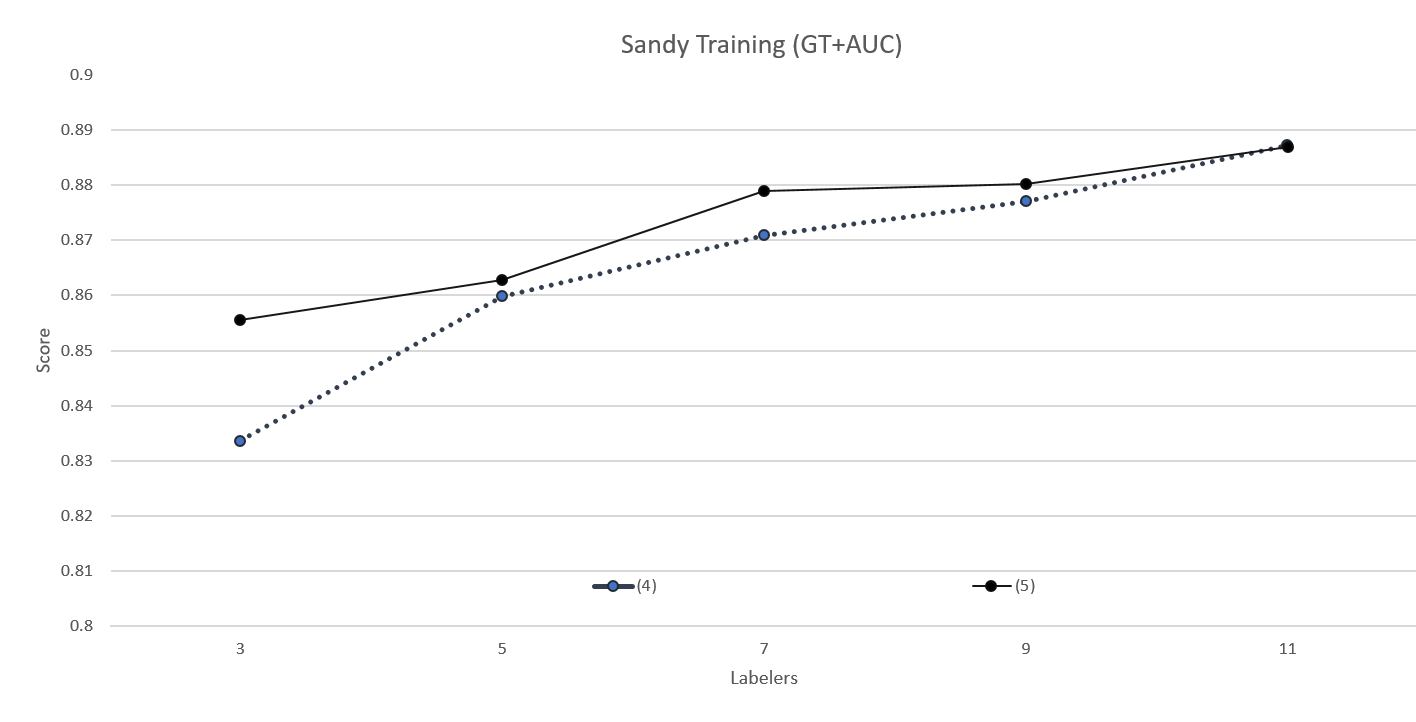


Fig. 6. c. (4) and (5) AUC comparisons for Set 2 Sandy training trials compared to ground truth.

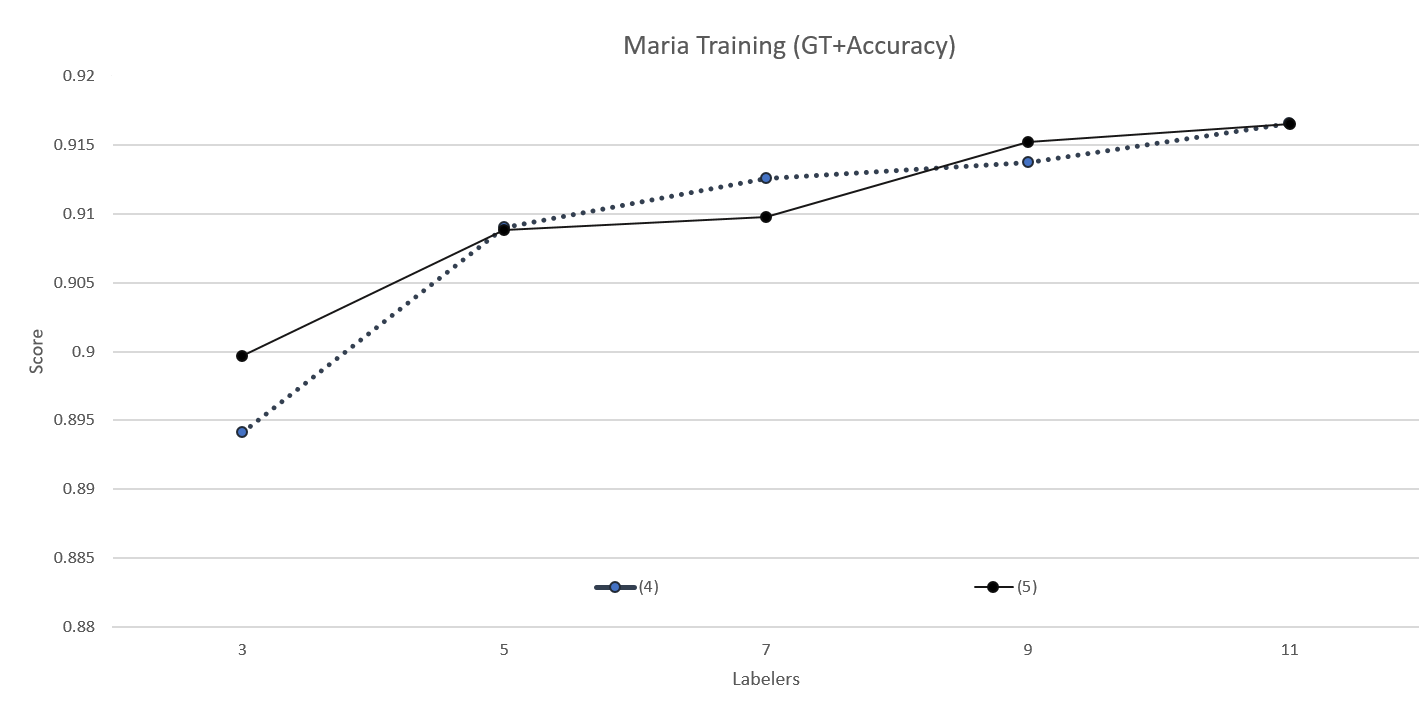


Fig. 6. d. (4) and (5) accuracy comparisons for Set 2 Maria training trials compared to ground truth.

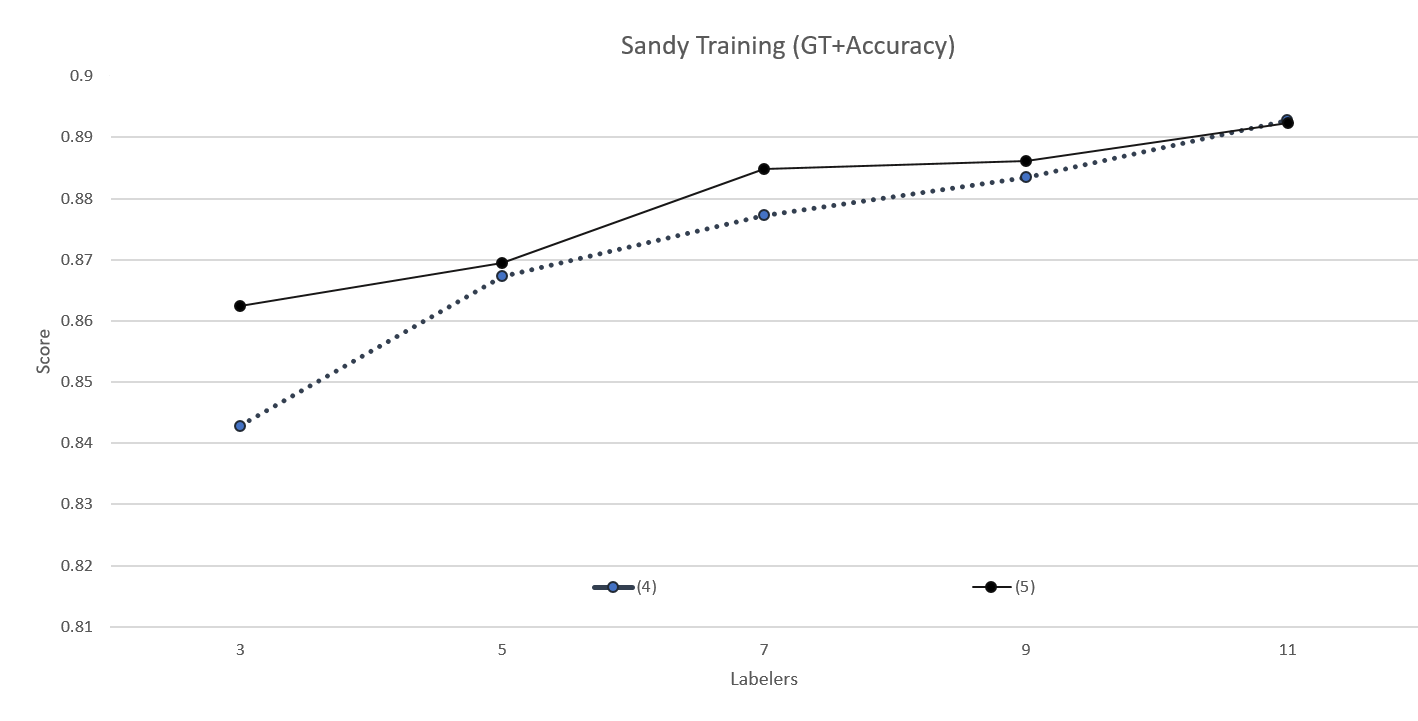


Fig. 6. e. (4) and (5) accuracy comparisons for Set 2 Sandy training trials compared to ground truth.

In Set 3, (7) works much better than (8) in training cases. When compared to the ground truth, (7) has eight better accuracy cases and seven better AUC cases compared to (8), which has two and three respectively. When compared to the majority voting, (7) has five better accuracy and AUC cases while (8) has three each (there were two ties). In sum for training, (7) had 25 better cases compared to (8)’s 11, which can be seen in Fig. 8. Therefore, (7) works much better than (8) in terms of training as it had about 127% higher percentage values expressed in accuracy and AUC. Although (7) does work better for training, (8) is better for testing. Compared to the ground truth, (7) has three better cases for accuracy and four better for AUC; (8) has seven and six, respectively. Compared to the majority voting, both (7) and (8) are better in five cases than the other for accuracy and AUC. Overall from Fig. 9, it can be seen that (7) has 17 better cases and (8) has 23 better cases; (8) has about 35.3% more values that are higher than (7). Since the degree of differences between the training trials and testing trials are both great, it is unclear which algorithm is better.

In Set 1, there was no clear distinction whether having labels or having weighted labels as the feature would increase the accuracy and AUC; (1) and (2) were very similar. In Set 2, it can be seen that using labels as the feature rather weighted labels would increase the accuracy and AUC values; (5) had more values that were higher than (4) so (5) is the clear winner out of the two. Also, in Set 3, there was no clear winner; (8) works better than (7) in testing, but lacked in training. Overall, there is no clear distinction to which feature is the best for increasing AUC and accuracy.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(4)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.86581** | **0.87285** | 0.88196 | **0.8866** | 0.88643 | 0.8656 | **0.87796** | 0.88759 | **0.89264** | 0.89287 |
| Sandy | 0.84644 | 0.87285 | 0.88895 | **0.89195** | **0.89925** | 0.82799 | 0.86041 | 0.88109 | 0.88374 | **0.89416** |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.99433** | 0.97766 | **0.97612** | **0.97371** | 0.97079 | **0.9949** | 0.97959 | **0.97824** | **0.97588** | 0.97269 |
| Sandy | 0.98839 | **0.97191** | 0.96423 | 0.9588 | 0.95655 | 0.99085 | **0.97844** | 0.97238 | 0.96818 | 0.96594 |
| **(5)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.86409 | 0.87251 | **0.88419** | 0.88471 | **0.88814** | **0.8665** | 0.8769 | **0.88952** | 0.89084 | **0.89414** |
| Sandy | **0.86367** | **0.87959** | **0.89157** | 0.8912 | 0.89813 | **0.85079** | **0.86931** | **0.88432** | **0.88388** | 0.89331 |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.9921 | **0.97887** | 0.97595 | 0.97148 | **0.97096** | 0.99277 | **0.98073** | 0.97808 | 0.97387 | **0.97304** |
| Sandy | **0.99176** | 0.96985 | **0.96498** | **0.96105** | **0.95787** | **0.9935** | 0.97684 | **0.97288** | **0.96997** | **0.96699** |

Fig. 7. a. (4) and (5) accuracy and AUC comparisons for Set 2 testing trials compared to ground truth and majority voting.

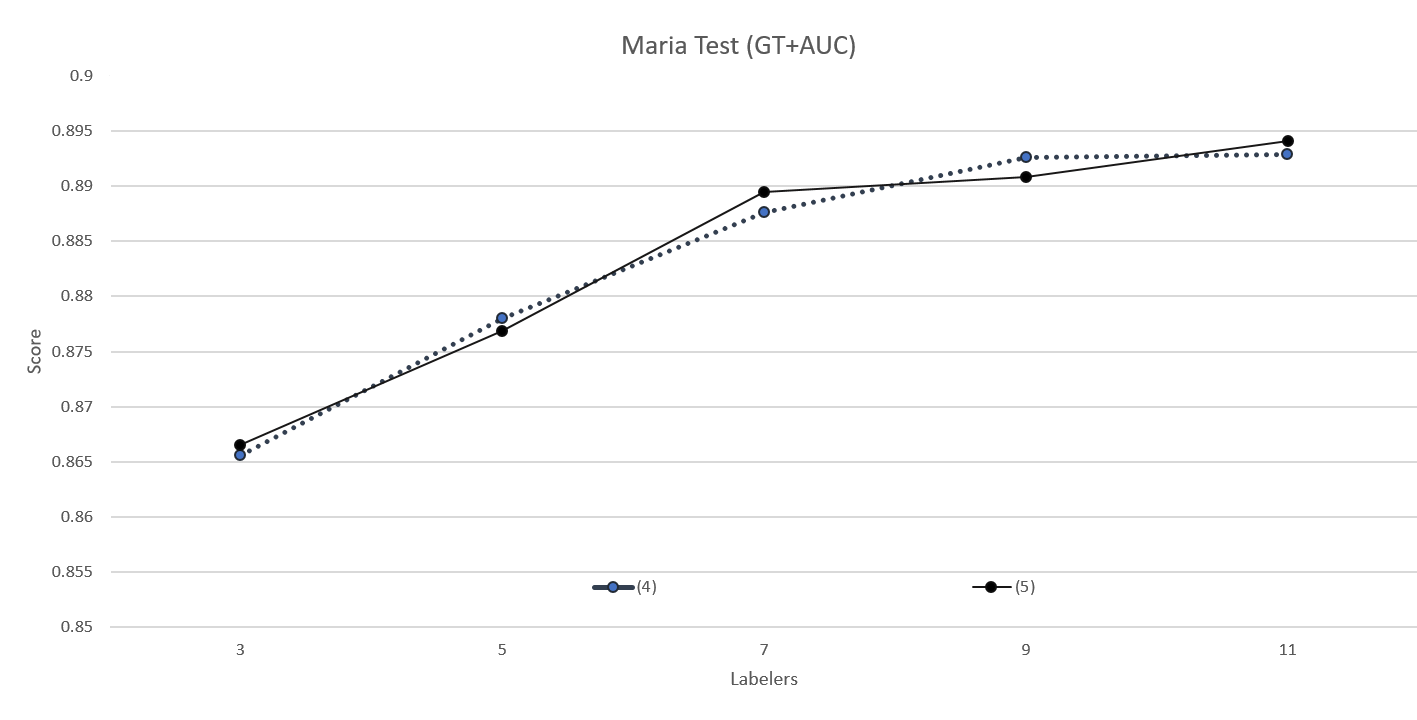


Fig. 7. b. (4) and (5) AUC comparisons for Set 2 Maria testing trials compared to ground truth.

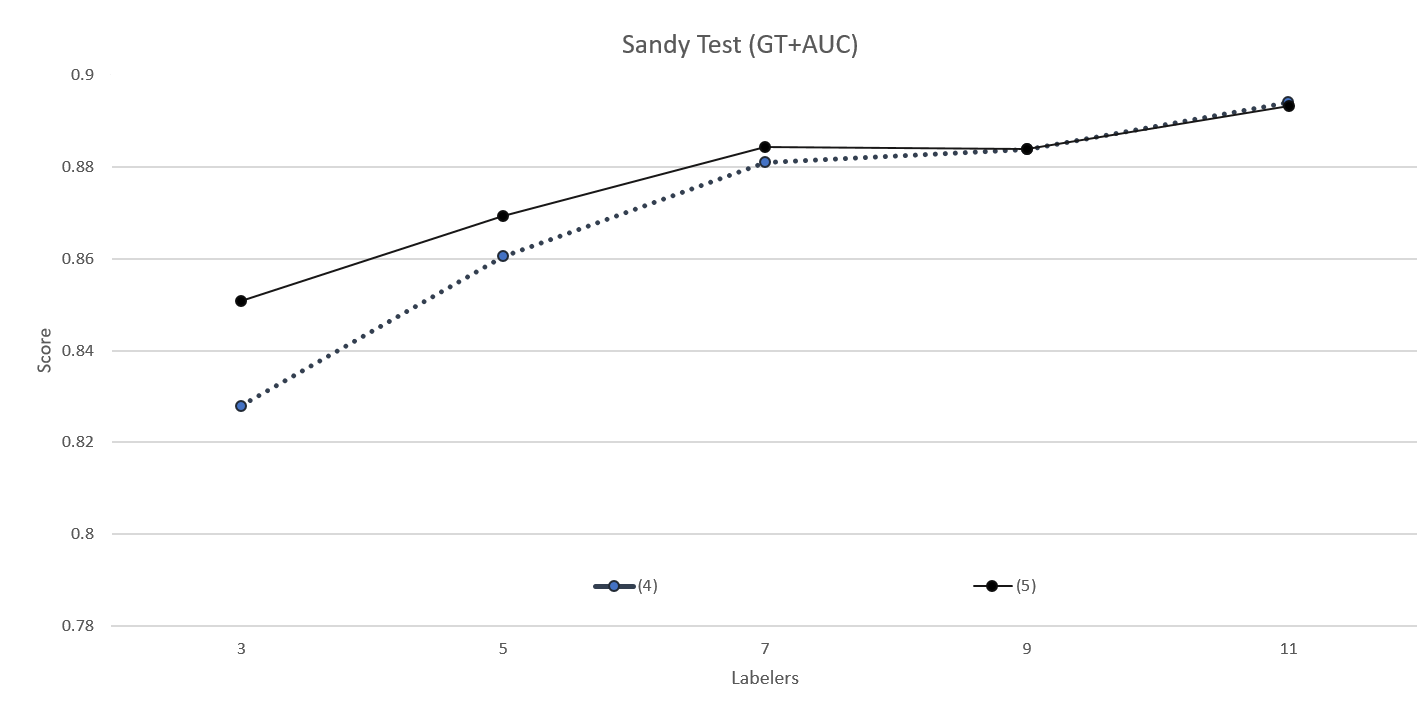


Fig. 7. c. (4) and (5) AUC comparisons for Set 2 Sandy testing trials compared to ground truth.

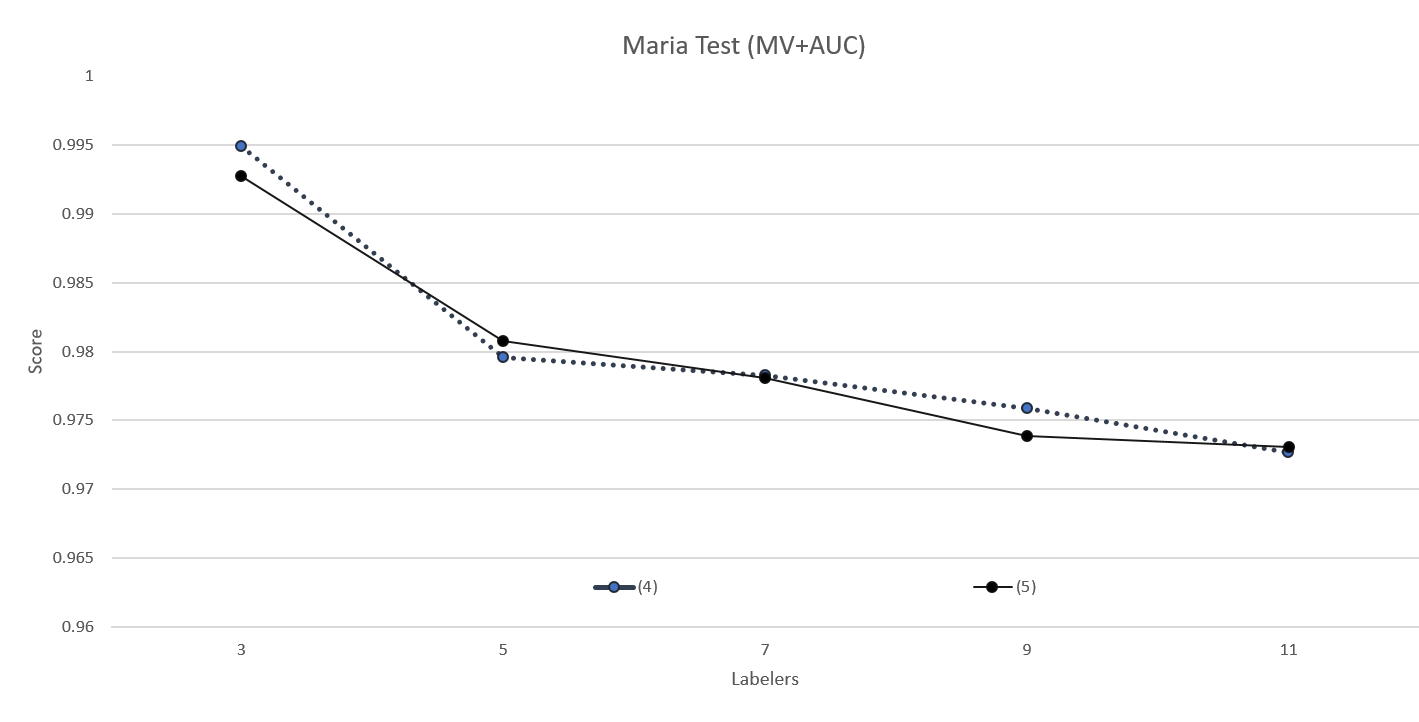


Fig. 7. d. (4) and (5) AUC comparisons for Set 2 Maria testing trials compared to majority voting.

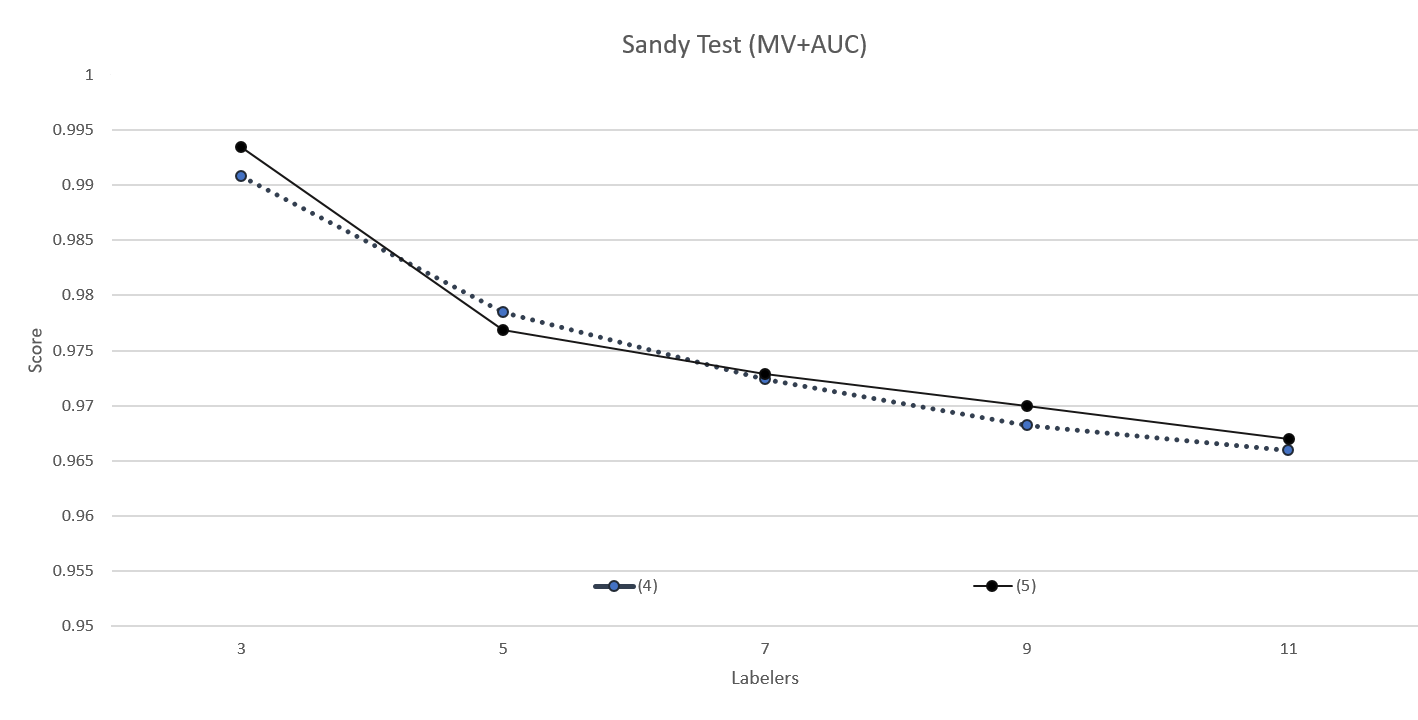


Fig. 7. e. (4) and (5) AUC comparisons for Set 2 Sandy testing trials compared to majority voting.

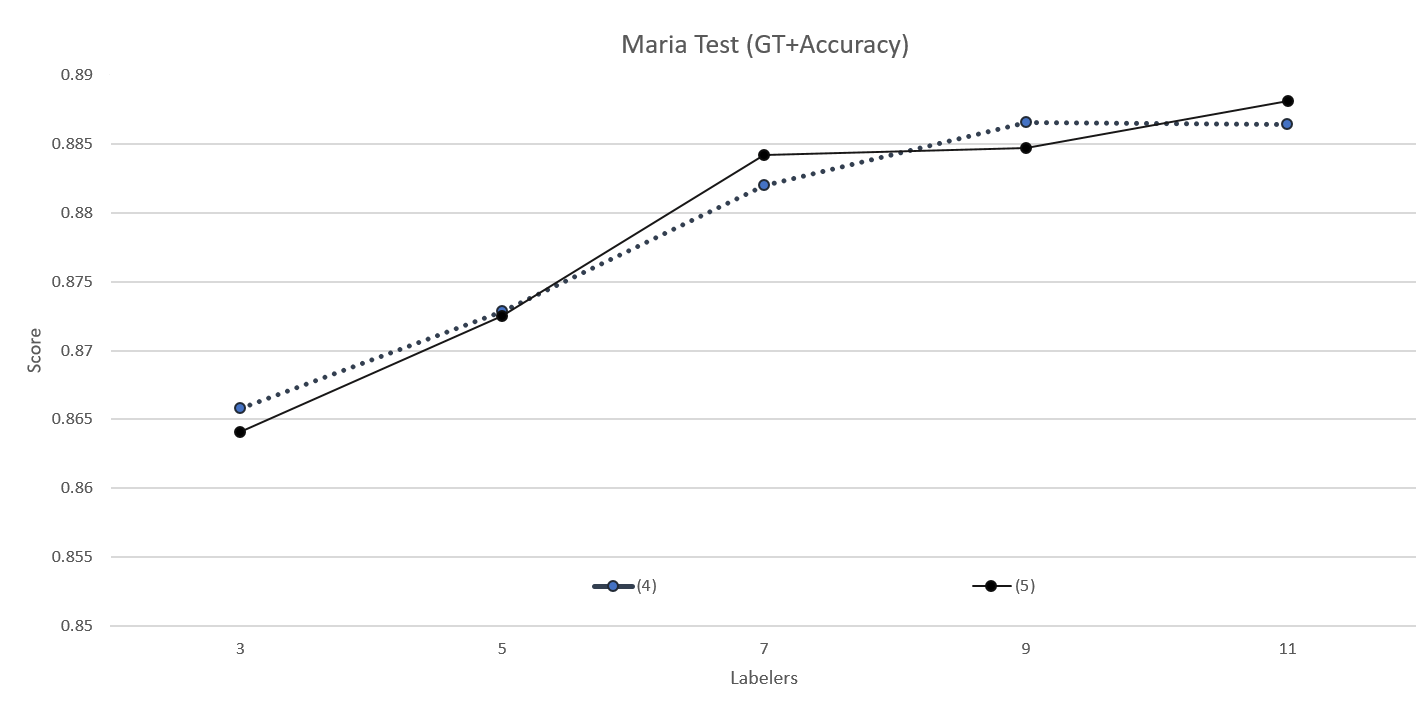


Fig. 7. f. (4) and (5) accuracy comparisons for Set 2 Maria testing trials compared to ground truth.

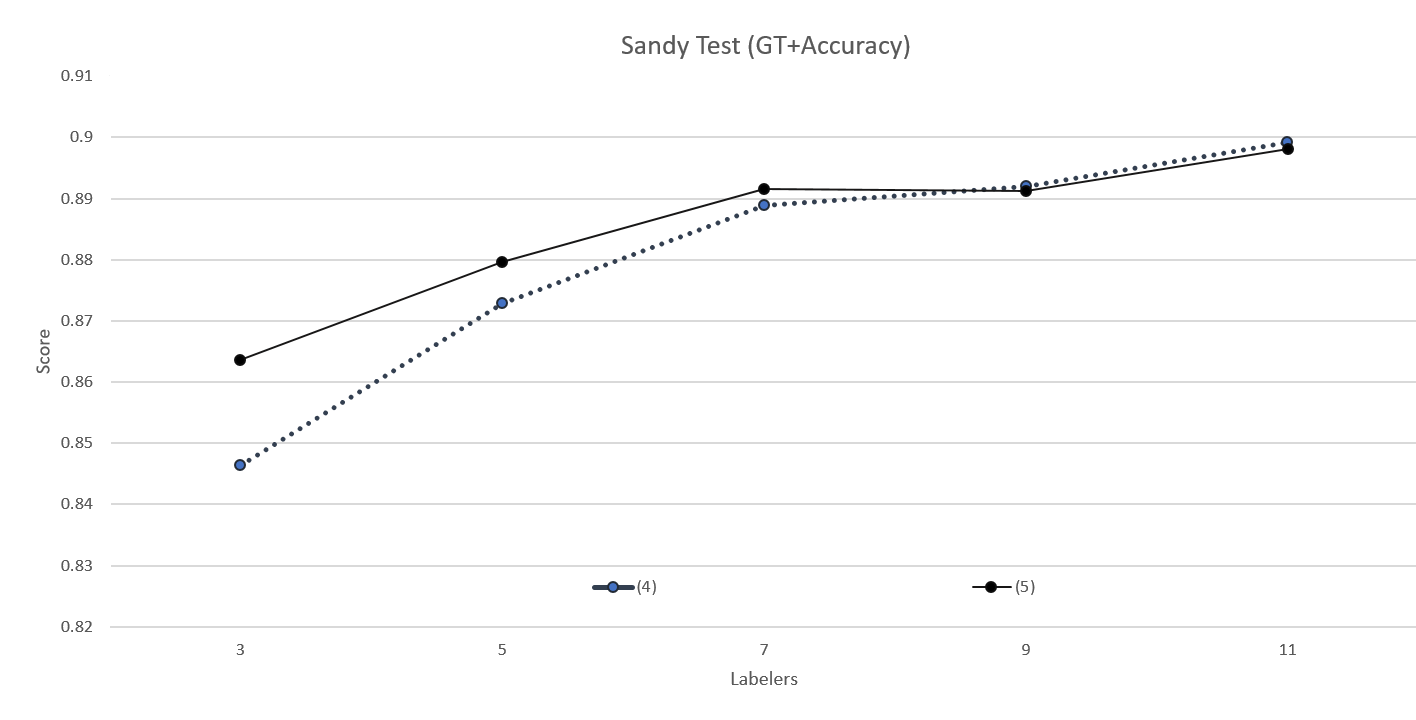


Fig. 7. g. (4) and (5) accuracy comparisons for Set 2 Sandy testing trials compared to ground truth.

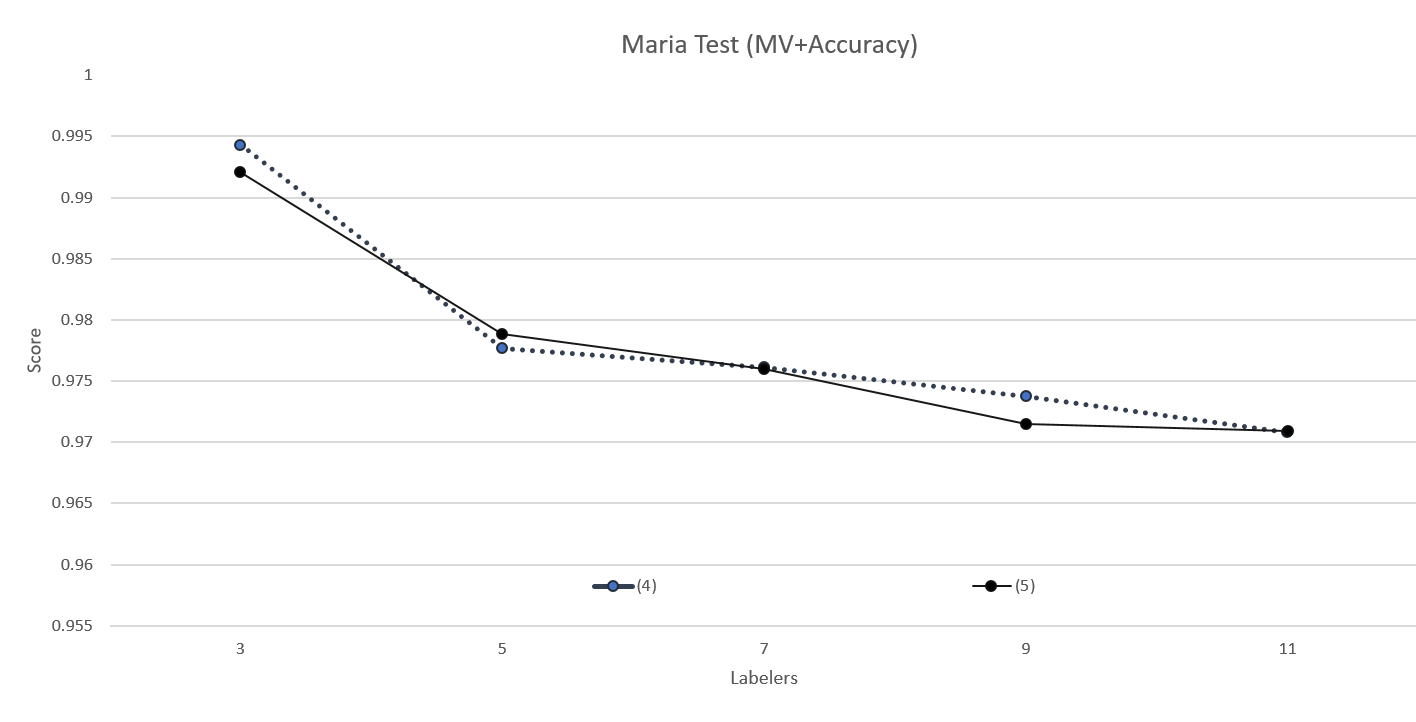


Fig. 7. h. (4) and (5) accuracy comparisons for Set 2 Maria testing trials compared to majority voting.

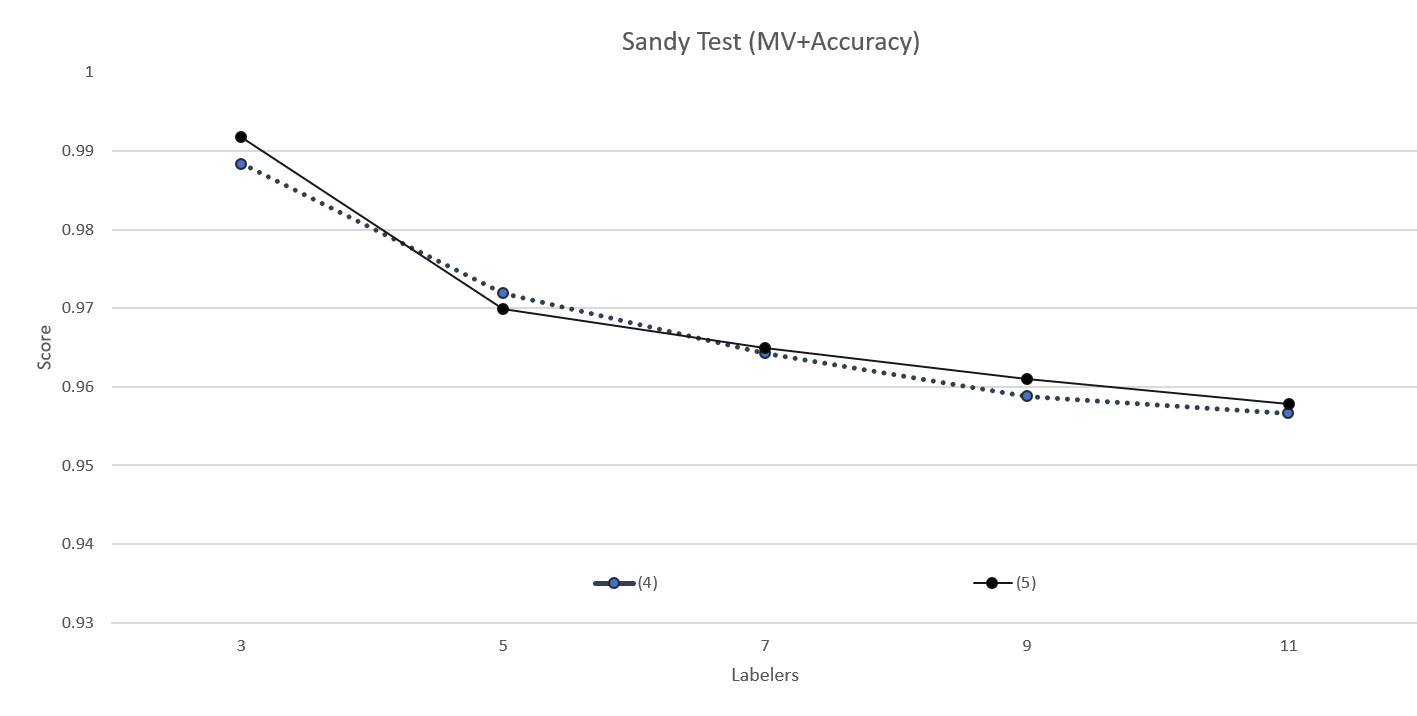


Fig. 7. i. (4) and (5) accuracy comparisons for Set 2 Sandy testing trials compared to majority voting.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(7)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.89674 | **0.90786** | **0.91349** | **0.91555** | **0.9171** | 0.89029 | 0.90267 | **0.90818** | **0.91087** | **0.9123** |
| Sandy | **0.85792** | **0.86676** | 0.87626 | **0.88764** | **0.89262** | **0.85068** | **0.85942** | 0.86968 | **0.88177** | **0.88729** |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **1** | 0.99991 | **0.99996** | 0.99991 | 0.99991 | **1** | 0.99992 | **0.99996** | 0.99993 | 0.99993 |
| Sandy | **0.99991** | 0.99972 | **0.99958** | **0.99934** | 0.99911 | **0.99992** | 0.99977 | **0.99964** | **0.99945** | 0.99925 |
| **(8)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.89781** | 0.90966 | 0.91138 | 0.91551 | 0.9168 | **0.89202** | **0.90449** | 0.90596 | 0.91072 | 0.91181 |
| Sandy | 0.85651 | 0.86653 | **0.87729** | 0.88256 | 0.89149 | 0.84871 | 0.85904 | **0.87079** | 0.87631 | 0.88605 |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Training** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.99996 | **0.99996** | 0.99991 | 0.99991 | 0.99991 | 0.99996 | **0.99996** | 0.99993 | 0.99993 | 0.99993 |
| Sandy | 0.99981 | **0.99976** | 0.99944 | 0.99925 | **0.99929** | 0.99986 | **0.99982** | 0.99955 | 0.99936 | **0.9994** |

Fig. 8. a. (7) and (8) accuracy and AUC comparisons for Set 3 training trials compared to ground truth and majority voting.

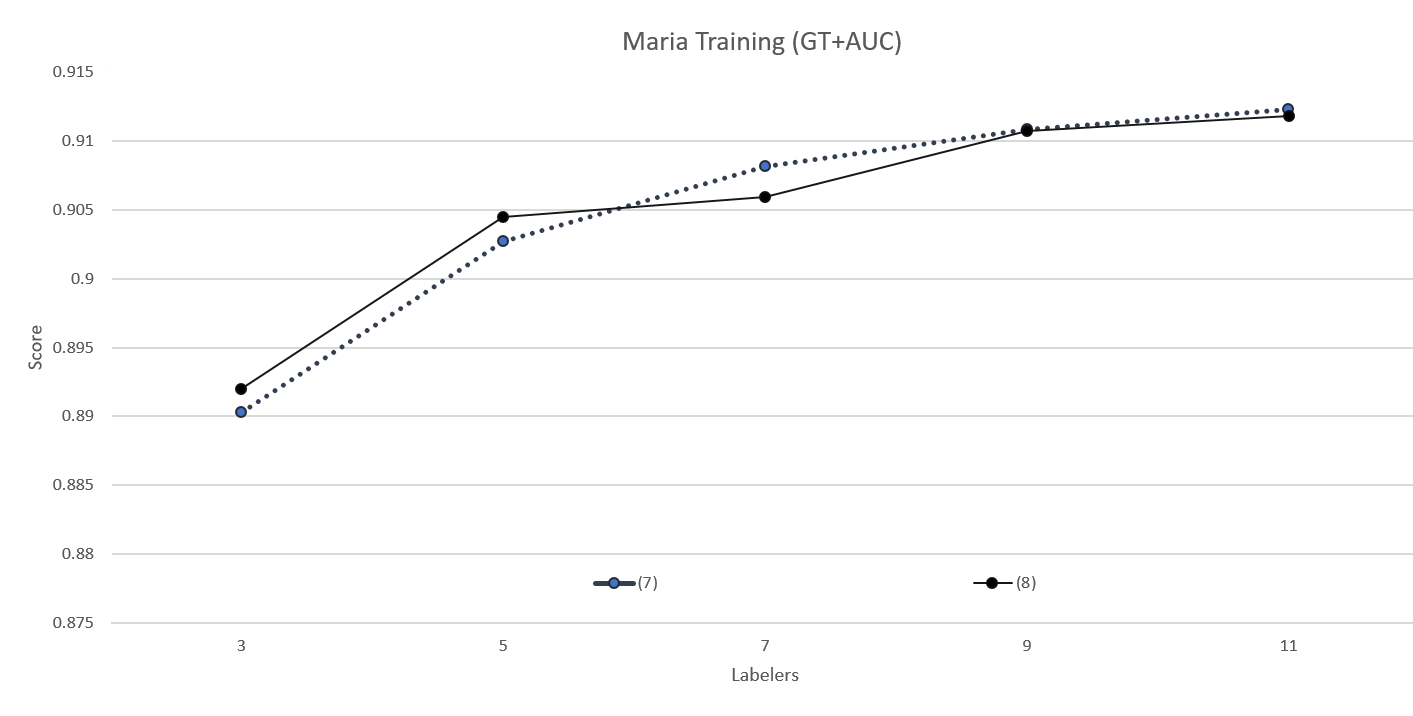


Fig. 8. b. (7) and (8) AUC comparisons for Set 3 Maria training trials compared to ground truth.

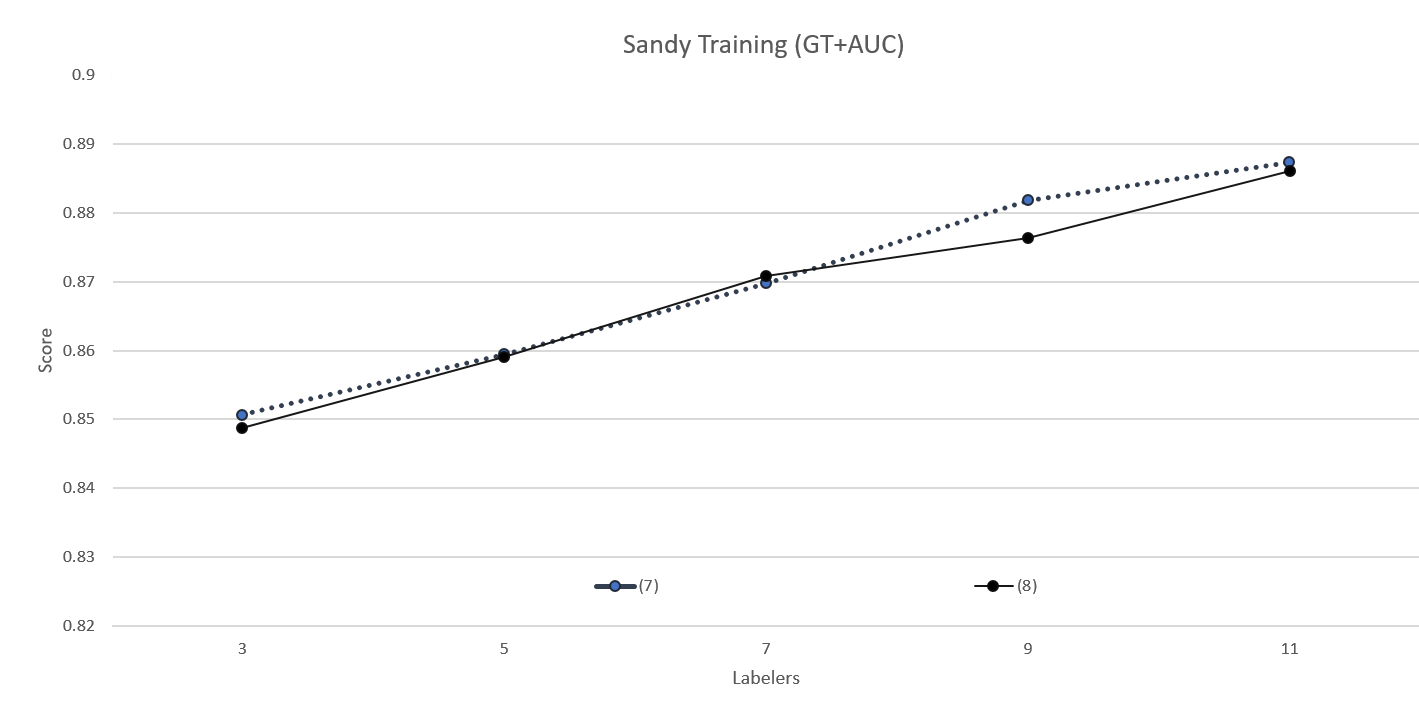


Fig. 8. c. (7) and (8) AUC comparisons for Set 3 Sandy training trials compared to ground truth.

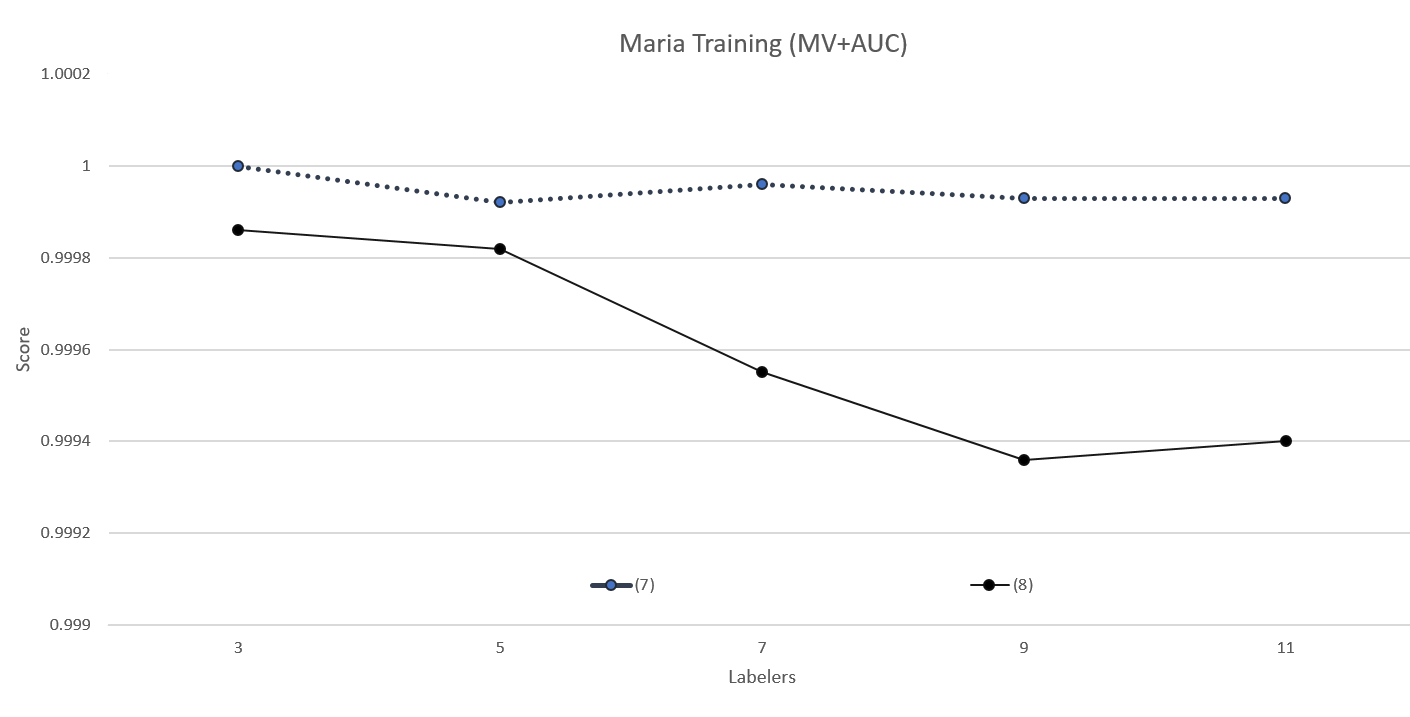


Fig. 8. d. (7) and (8) AUC comparisons for Set 3 Maria training trials compared to majority voting.

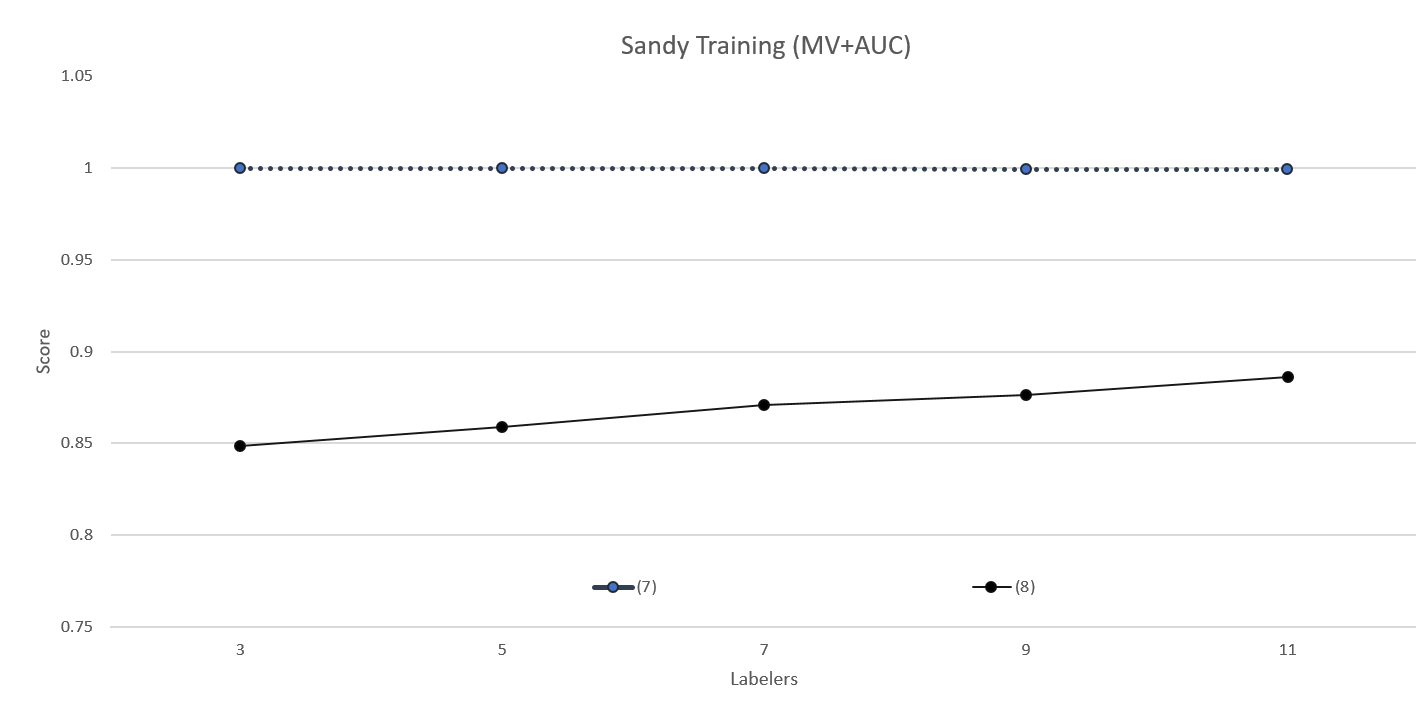


Fig. 8. e. (7) and (8) AUC comparisons for Set 3 Sandy training trials compared to majority voting.

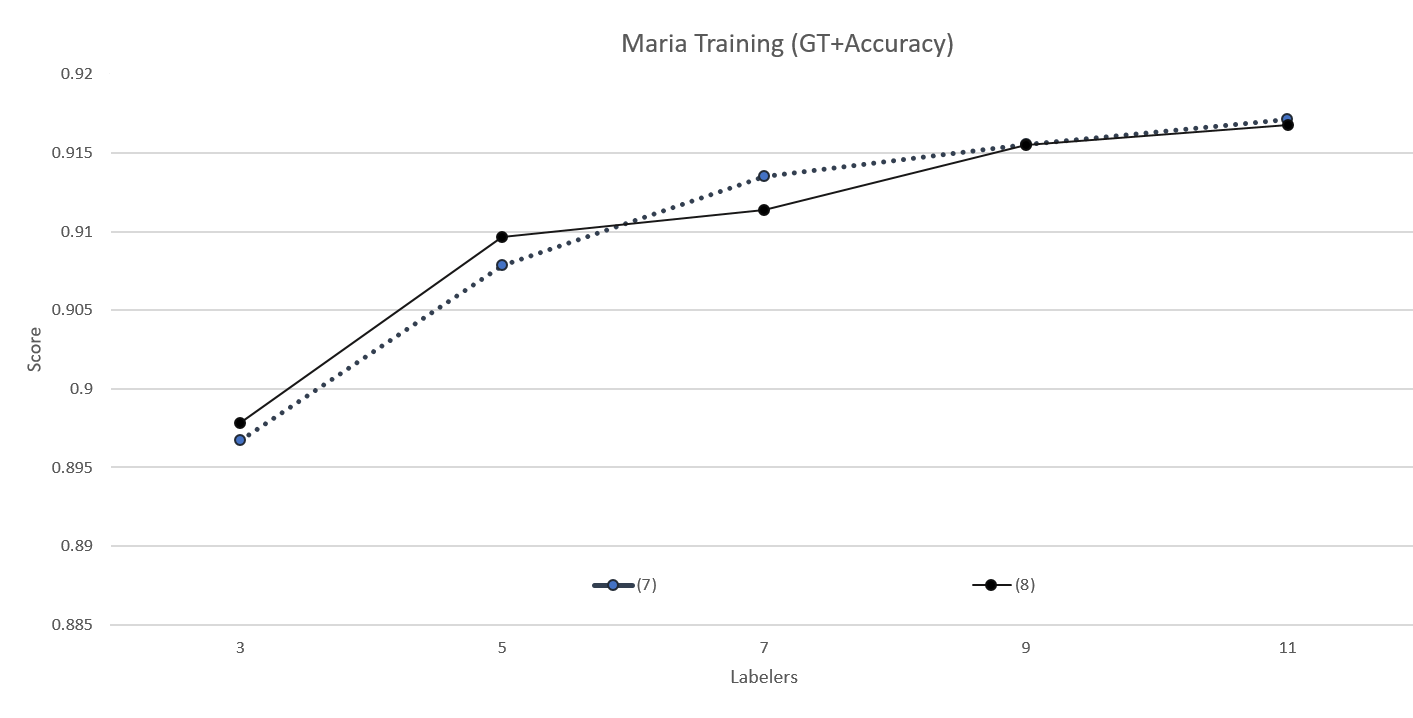


Fig. 8. f. (7) and (8) accuracy comparisons for Set 3 Maria training trials compared to ground truth.

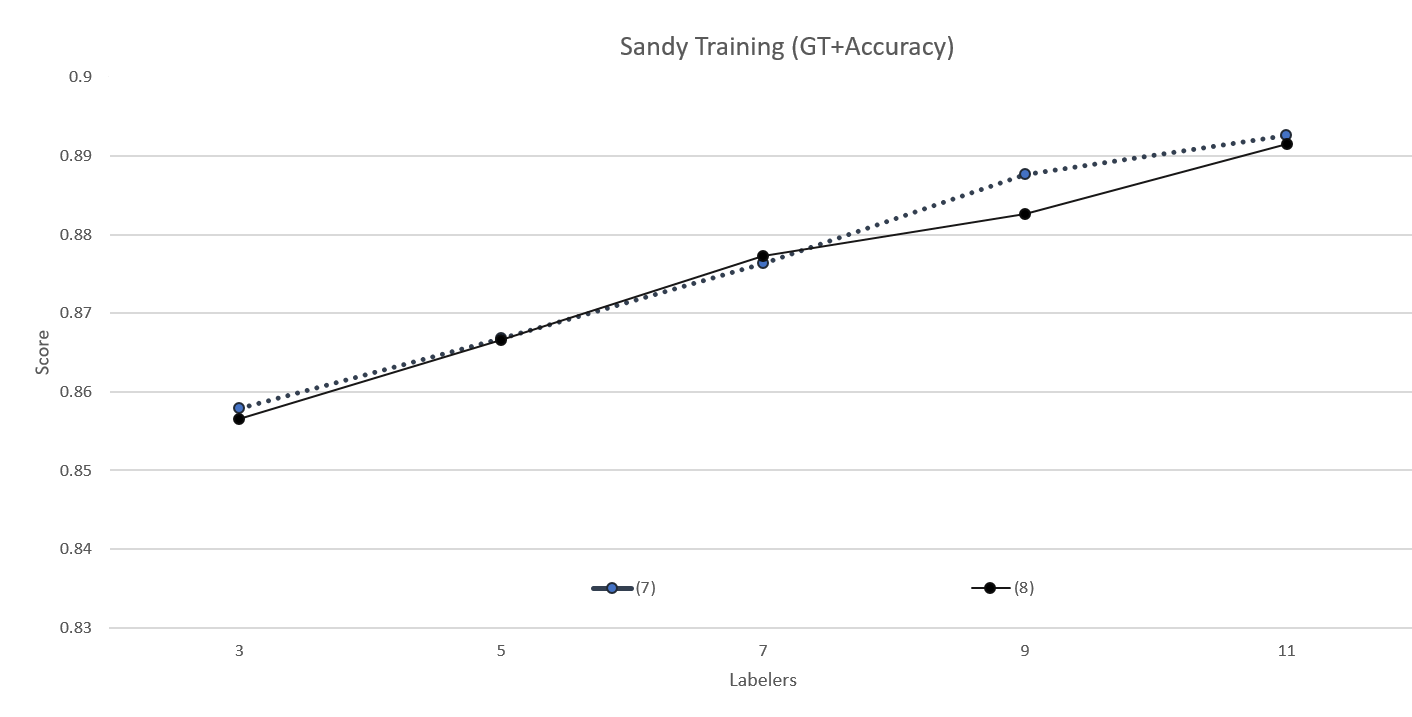


Fig. 8. g. (7) and (8) accuracy comparisons for Set 3 Sandy training trials compared to ground truth.

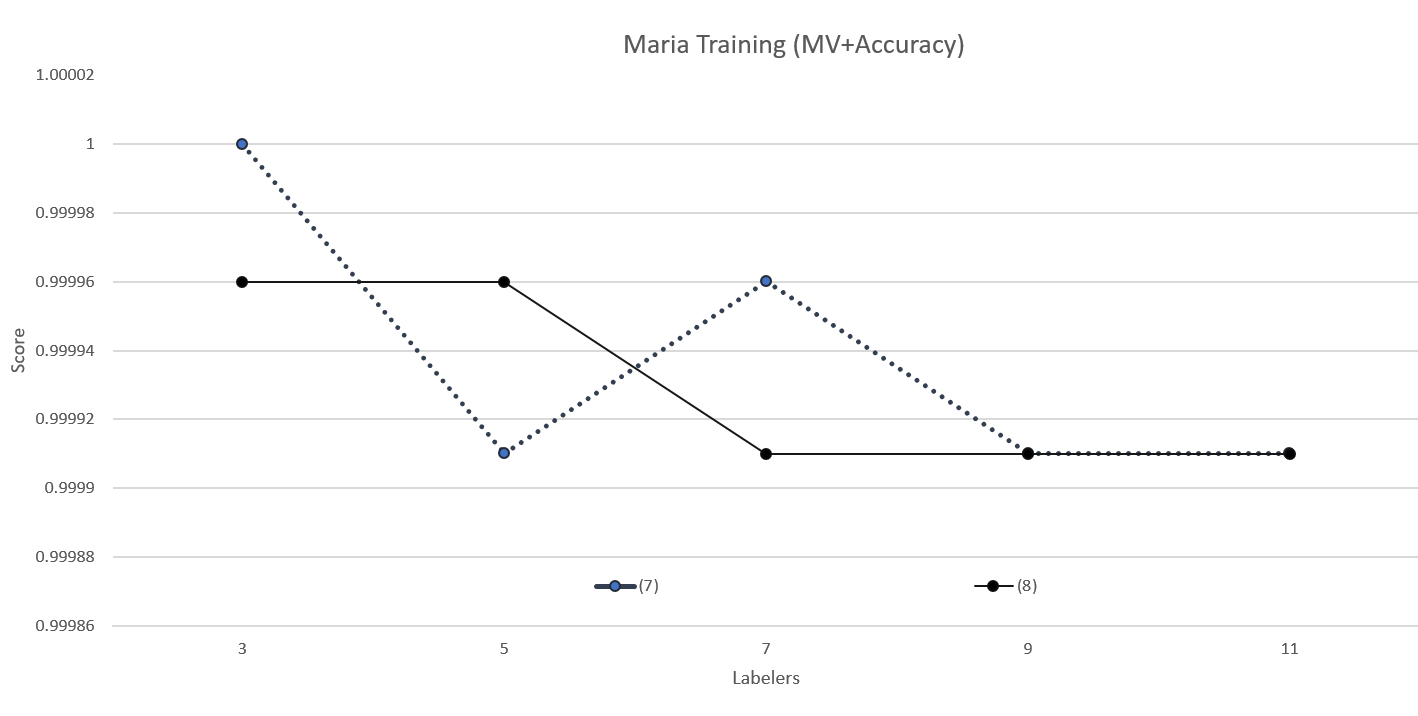


Fig. 8. h. (7) and (8) accuracy comparisons for Set 3 Maria training trials compared to majority voting.

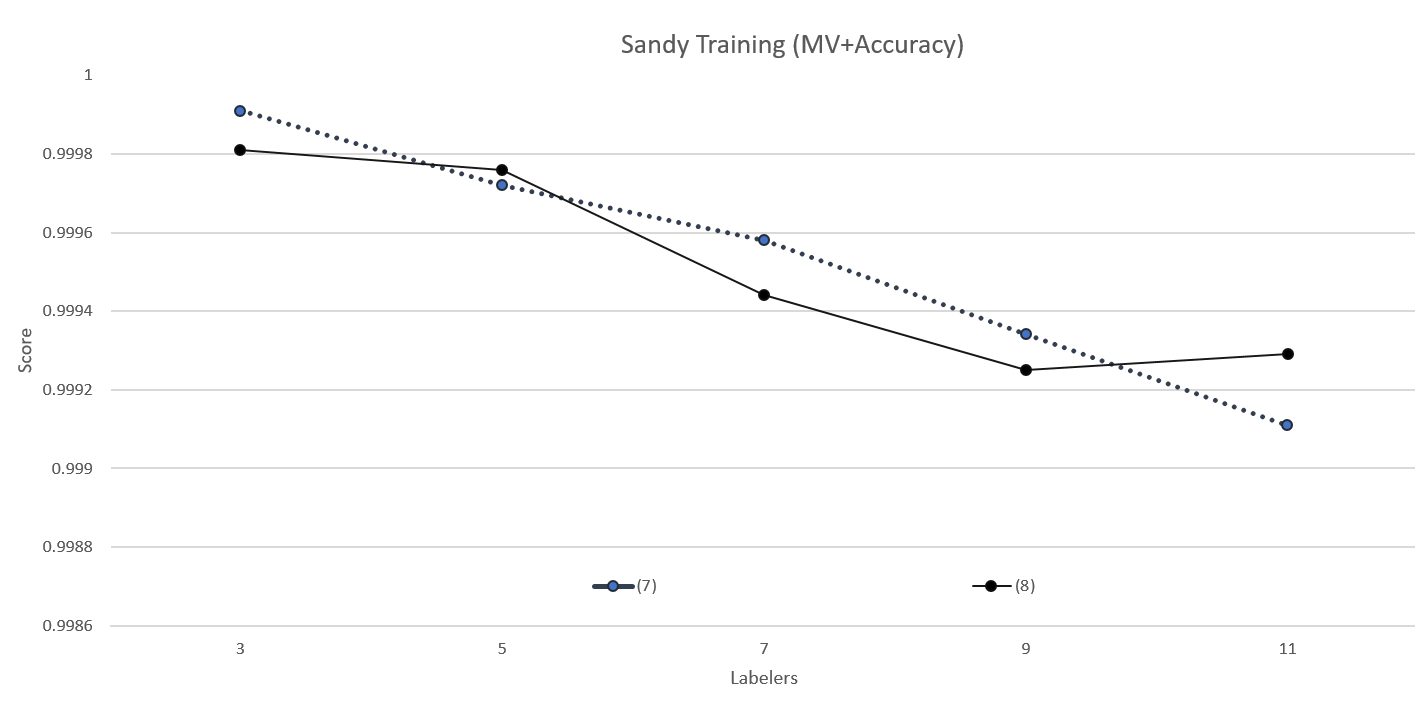


Fig. 8. i. (7) and (8) accuracy comparisons for Set 3 Sandy training trials compared to majority voting.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(7)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.86031 | 0.87457 | 0.87646 | 0.88213 | **0.8866** | 0.86105 | 0.8789 | 0.88103 | 0.88779 | **0.89264** |
| Sandy | 0.86236 | 0.87378 | 0.88839 | **0.89157** | **0.8985** | **0.84857** | 0.86182 | 0.87921 | **0.88493** | **0.89309** |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.9945** | 0.97732 | **0.97285** | 0.97251 | 0.97096 | **0.99517** | 0.97939 | **0.97545** | 0.9747 | 0.97286 |
| Sandy | **0.99045** | **0.97097** | **0.96442** | 0.9603 | 0.95974 | **0.99264** | **0.97814** | **0.97265** | 0.96927 | 0.96846 |
| **(8)** |  |  |  |  |  |  |  |  |  |  |
|  | **Ground Truth + Accuracy** | | |  |  | **Ground Truth + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | **0.86615** | **0.87646** | **0.87749** | **0.88436** | 0.88557 | **0.8678** | **0.88097** | **0.88274** | **0.88994** | 0.89153 |
| Sandy | **0.85787** | **0.87753** | **0.88914** | 0.88839 | 0.89644 | 0.84299 | **0.86542** | **0.88173** | 0.88048 | 0.89088 |
|  | **Majority Voting + Accuracy** | | |  |  | **Majority Voting + AUC** | | |  |  |
| **Labelers** | **3** | **5** | **7** | **9** | **11** | **3** | **5** | **7** | **9** | **11** |
| **Test** |  |  |  |  |  |  |  |  |  |  |
| Maria | 0.99381 | **0.97818** | 0.97234 | **0.97474** | **0.97131** | 0.99425 | **0.98011** | 0.97478 | **0.97659** | **0.97364** |
| Sandy | 0.98464 | 0.97041 | 0.95581 | **0.96199** | **0.96049** | 0.98796 | 0.9777 | 0.96633 | **0.97077** | **0.96912** |

Fig. 9. a. (7) and (8) accuracy and AUC comparisons for Set 3 testing trials compared to ground truth and majority voting.

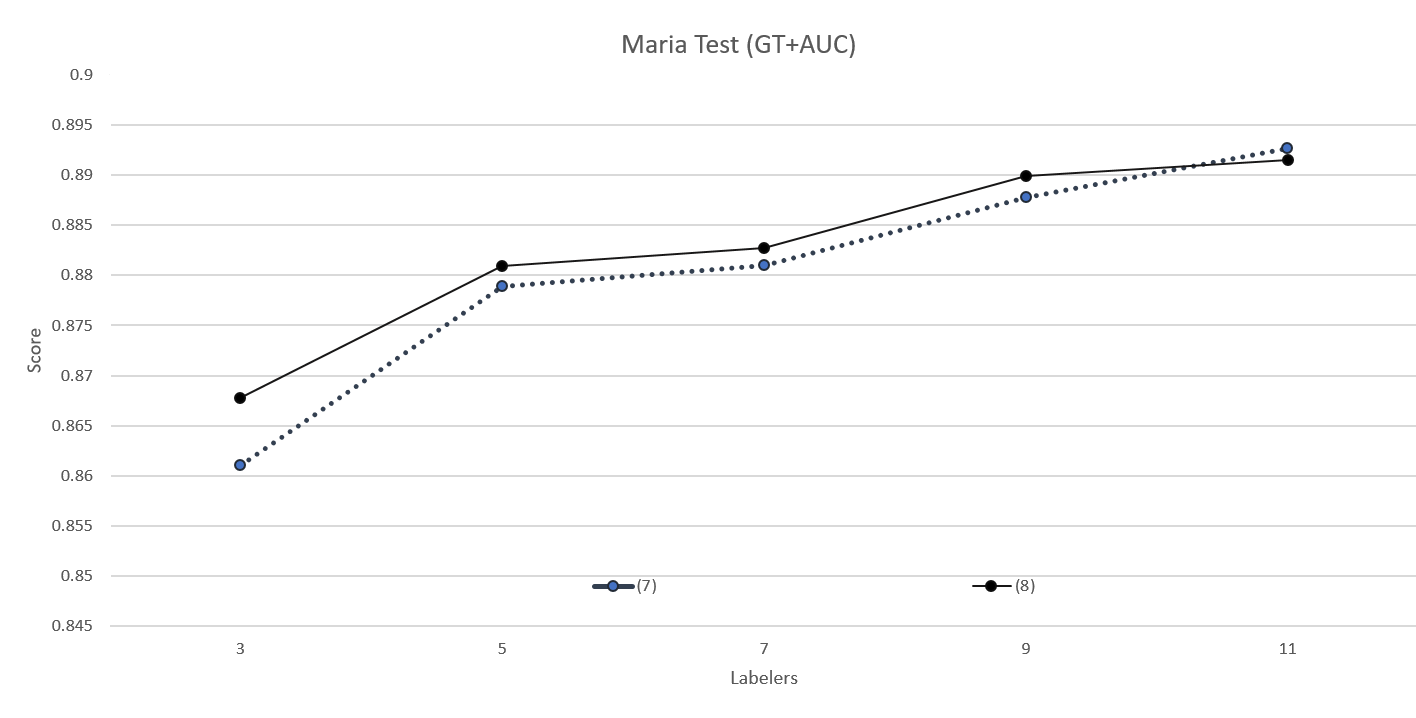


Fig. 9. b. (7) and (8) AUC comparisons for Set 3 Maria testing trials compared to ground truth.

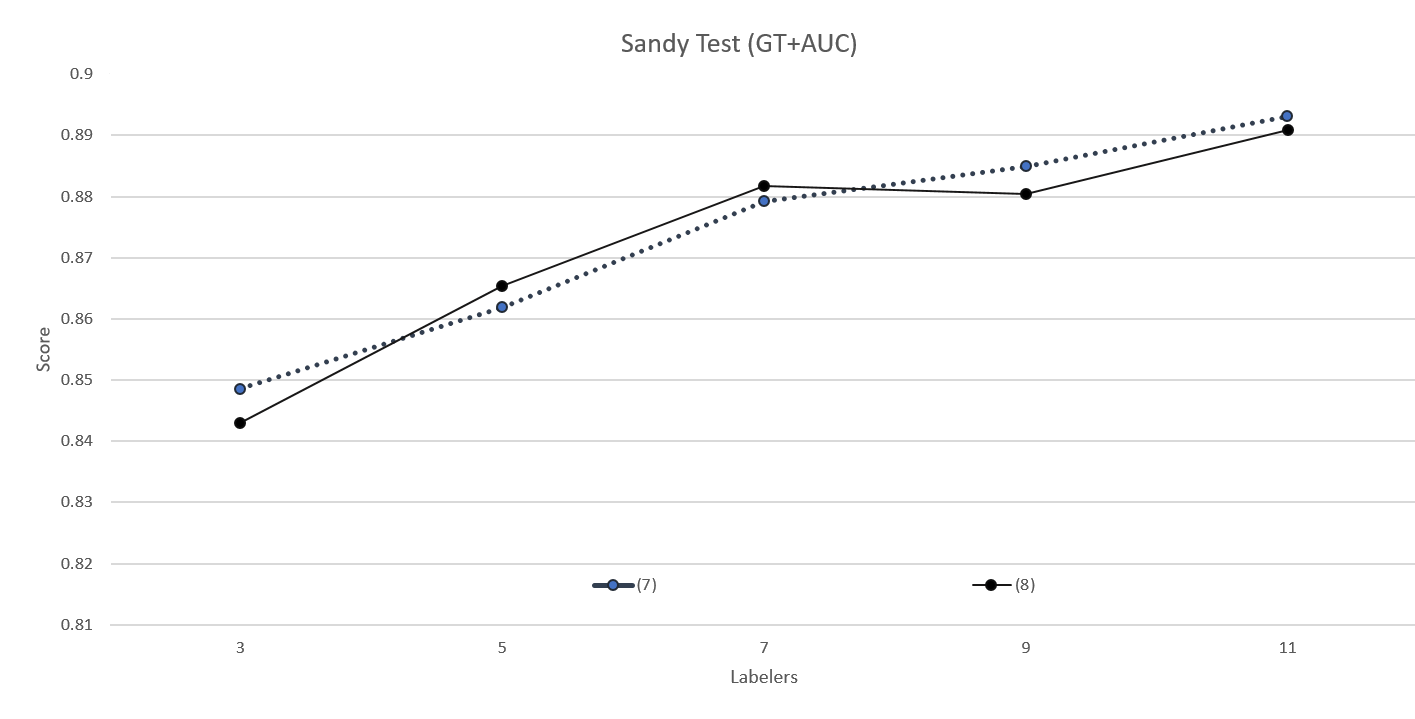


Fig. 9. c. (7) and (8) AUC comparisons for Set 3 Sandy testing trials compared to ground truth.

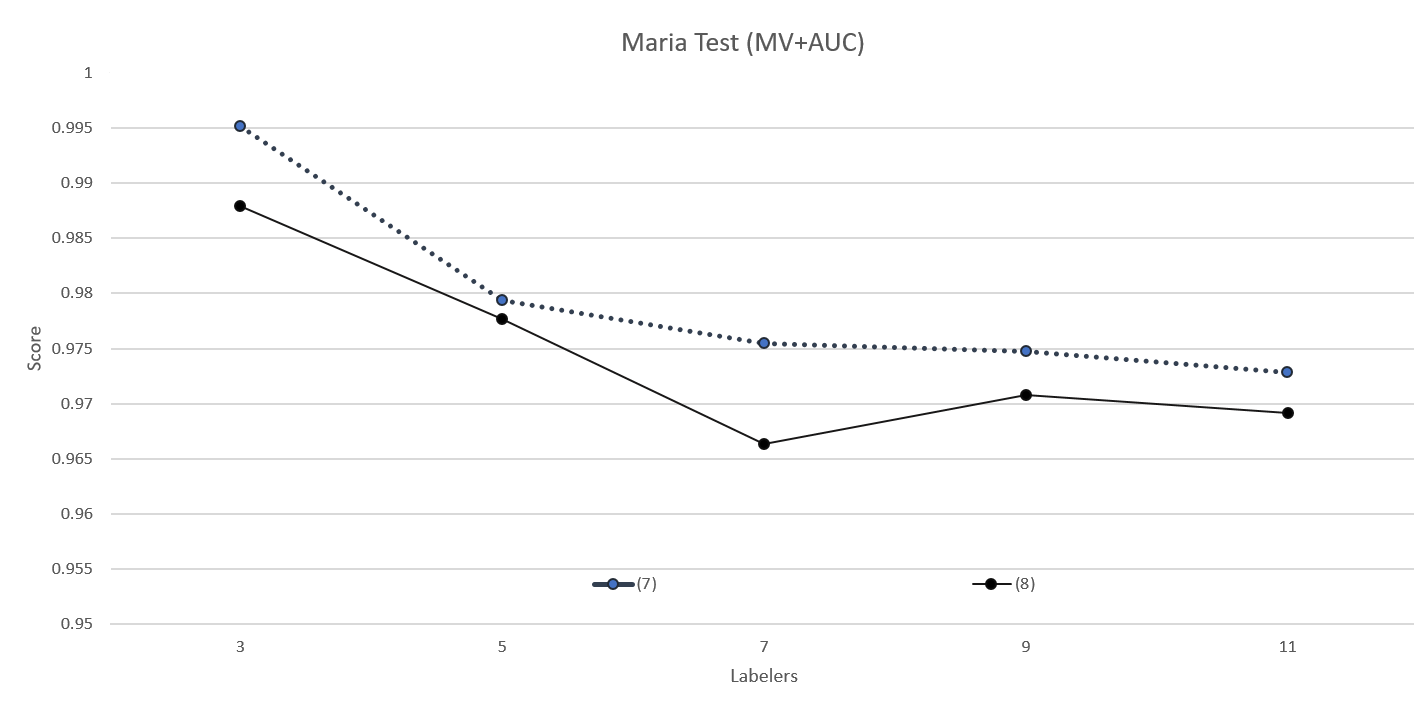


Fig. 9. d. (7) and (8) AUC comparisons for Set 3 Maria testing trials compared to majority voting.

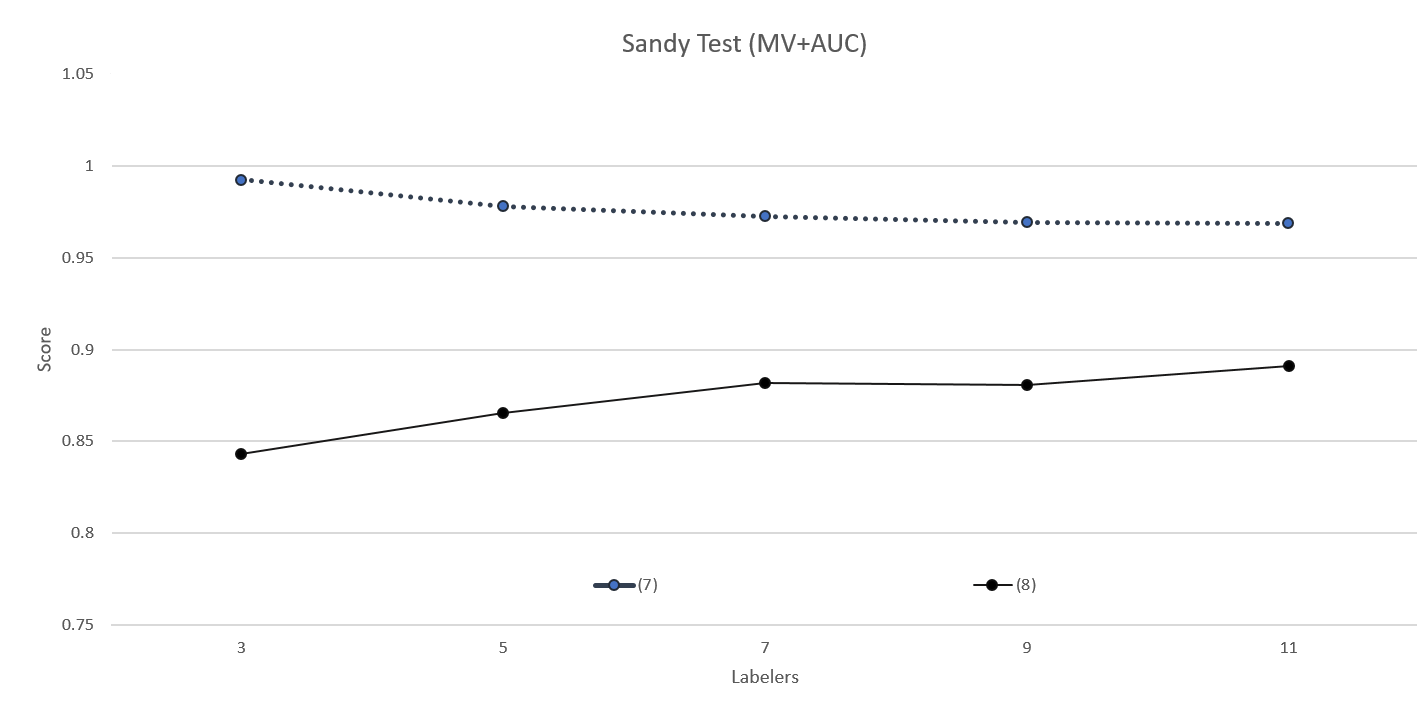


Fig. 9. e. (7) and (8) AUC comparisons for Set 3 Sandy testing trials compared to majority voting.

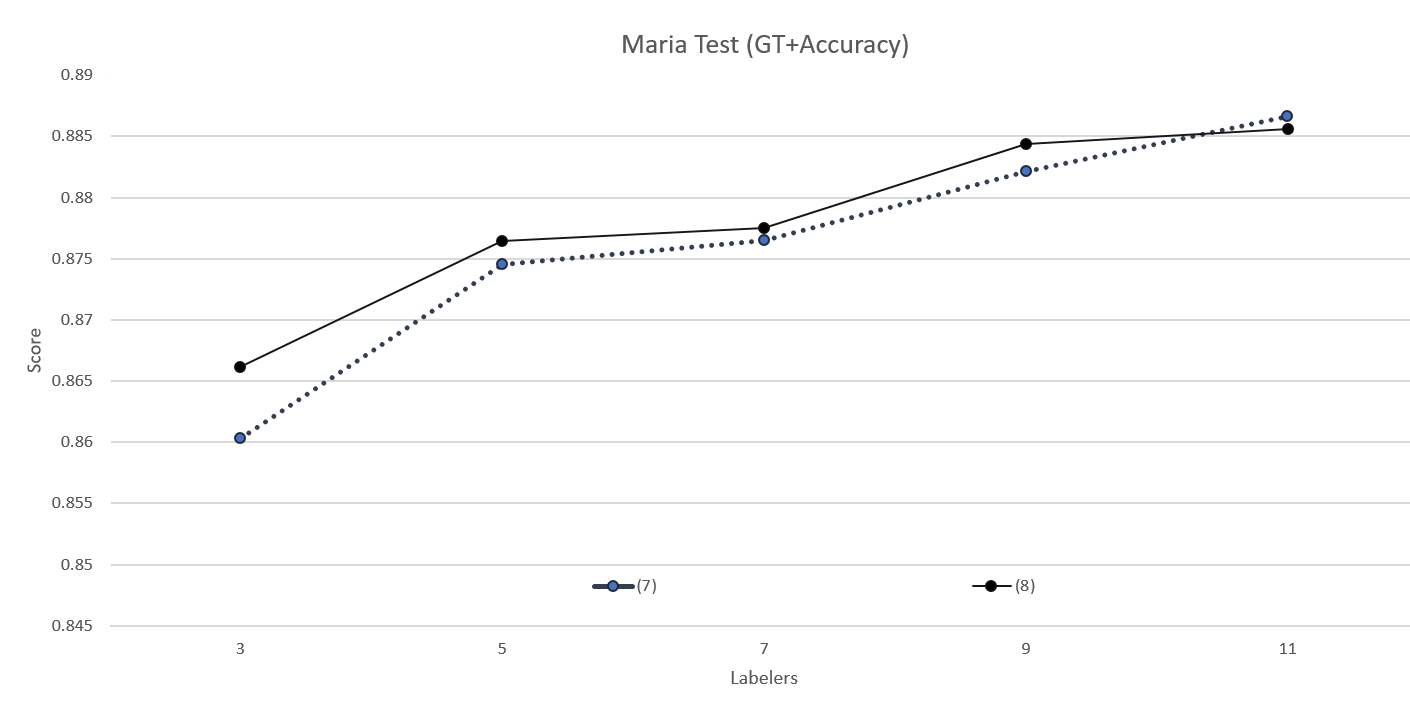


Fig. 9. f. (7) and (8) accuracy comparisons for Set 3 Maria testing trials compared to ground truth.

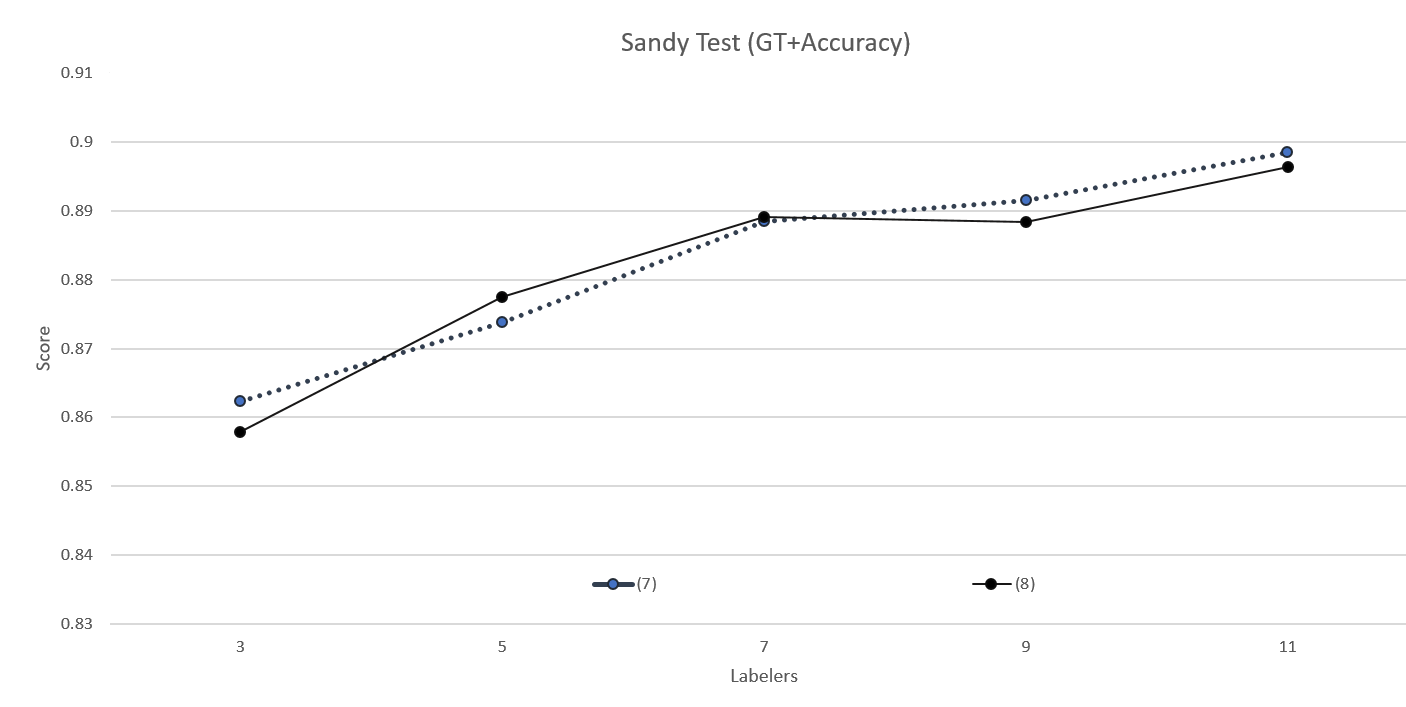


Fig. 9. g. (7) and (8) accuracy comparisons for Set 3 Sandy testing trials compared to ground truth.



Fig. 9. h. (7) and (8) accuracy comparisons for Set 3 Maria testing trials compared to majority voting.

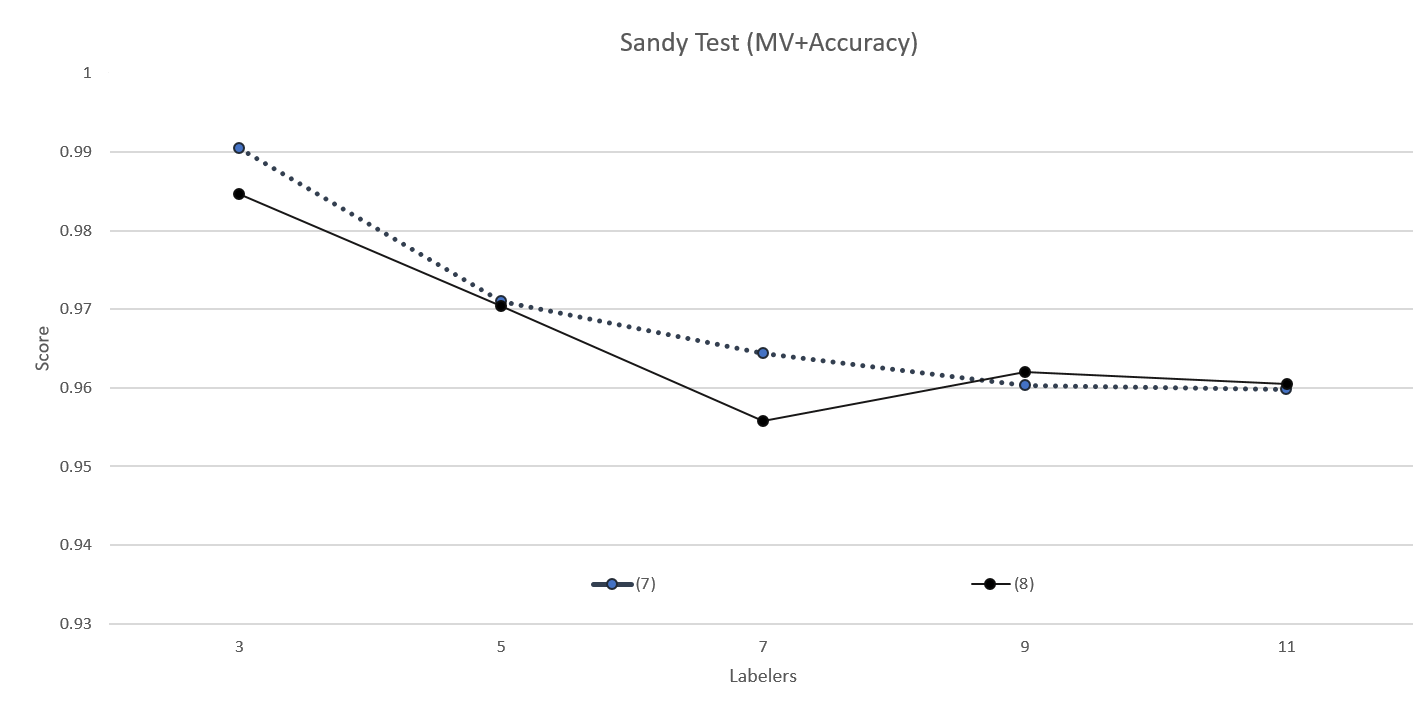


Fig. 9. i. (7) and (8) accuracy comparisons for Set 3 Sandy testing trials compared to majority voting.