“”””  
My thought:  
In the first 2 datasets, all went well since the background was the same in each image and the subject was easily recognizable. The third dataset was hard for the AI because the subjects were hard to recognize since the clothes were different in each picture as well as the background. The face, the important "feature" was small and sometimes covered. The last one was a medium-difficult because the subjects were recognizable but the background made go mad the AI. A way to solve these problems is to cut the background and analyze only the subjects  
”””

Actually, an AI is always right. It’s just the result we don’t expect. When we get a result that we think it’s wrong, it’s just because the algorithm is learning the training data set, but not the features we are looking for.  
The secret of a good training data set is to get a lot of images under all the possible conditions (light conditions, make-up conditions, clothes conditions, etcetera). We need to get pictures that “tell” to the algorithm what we want to make it learn.  
If we need to train a model to recognize two people, we ask the people to make a lot of photos with a lot of differences (clothes, makeup, light conditions,…) so the algorithm learns the faces, not the clothes or some other particular. Another example is a model that recognizes cats and dogs. If we have no gray dog in the training data set, but a lot of grey cats, even if we give to the model the picture of a gray dog, it would be classified as a gray cat. To fix this problem, just look for a gray dog picture and add it to the testing data set.