Author: Chen, X. (G.)
Title: Prediction sets via parametric and nonparametric Bayes: With applications in pharmaceutical industry
Issue date: 2021-01-05

Chapter 5

Application: Image-Based Filter for Silicon Oil Particles[1]

5.1 Introduction

The objective of this chapter is to classify images as showing either a silicon oil (SO) or a non-silicon oil (NSO) particle. Silicon oil particles have a clear profile, but non-silicon oil particles may have many, also unknown forms. For this reason our chosen method of classification is based on agreement of a future image with a data-base of images of SO particles, rather than on training a classifier on samples of SO and NSO particles. In machine learning terms our approach is unsupervised, rather than supervised. We measure the agreement by data depth, whence this chapter also presents an application of data depth as described in the preceding chapter.

The images of interest are acquired from *flow imaging microscopy* (FIM), more precisely Micro-Flow Imaging (MFI) 5200 in this study, which is prevalent in the pharmaceutical industry. Their resolution varies by the total number of pixels, and by the average and range of brightness across the pixels. Since our proposed SO filter/classifier is image based, rather than on the metadata of the images, we expect that the methods developed here will lend themselves also to other FIM systems with an output of 2D images in gray scale, such as FlowCAM.

The industrial background of the application is described in Section 5.2. In Section 5.3 we review the state-of-the-art classification methods (or 'filters') currently used in the industry. These are all

---

[1] This chapter is based on joint work with Björn Boll and Migle Grauzinyte in Novartis Global Drug Development. A significant part of the chapter is reworked into a paper, "Applying pattern recognition as a robust approach for silicone oil droplet identification in flow-microscopy images of protein formulations", that has been accepted and published by Journal of Pharmaceutical Sciences, DOI:https://doi.org/10.1016/j.xphs.2020.10.044.

based on metadata of the images, in contrast to our new approach, which takes the full images as inputs. Section 5.4 presents the data used, and the methods used to prepare the raw images for analysis. The construction of our SO classifier is described in Section 5.5.

Section 5.6 compares the new filter against three benchmark approaches (summarized in Section 5.3) on three groups of test data, consisting of images of particles that are clearly SO (I), likely SO (II), or clearly NSO (III). The selection of these test datasets is motivated at the beginning of the section. Evaluation criteria are the false negative rate for datasets (I) and (II), and the false positive rate for dataset (III). The newly proposed filter is overall an improvement on the benchmark approaches. Detailed conclusions and thoughts for future improvements are provided in Sections 5.7 and 5.8.

5.2 Industrial Background

Injections and parenteral infusions are typically contaminated by extraneous, mobile and undissolved particulates, which are unintentionally present in the solutions [39]. Sources of contamination in sterile and/or biologic products include the solution itself, its ingredients, the manufacturing process, environmental factors, packaging, and the preparation of the product for administration. See [55] for discussion and further sources.

Particles can be categorized by size, as visible (>100 μm), micron (1-100 μm), sub-micron (100-1000 nm) and nanometer (<100 nm) particles [65]. In a vaguer sense, all particles of size <100 μm can be called subvisible. Currently the US and European pharmacopoeia only have clear requirements on the number of subvisible particles of sizes $\geq$10 and of sizes $\geq$25 μm, which are usually quantified by light obscuration in the parenteral solution. However, regulatory changes are expected ([15] and USP<1787>), spurring industrial interest in the quantification of subvisible particles with techniques that are orthogonal to light obscuration. A recent option is *flow imaging microscopy* (FIM), which uses a camera with high magnification power to capture images of a liquid solution passing through a thin flow cell. A detailed account of FIM, with a comparative evaluation of several instruments, is offered in [70].

It is presently not uncommon to see FIM applied in quality control of protein therapeutics, which takes an increasing share in the product portfolio of many pharmaceutical companies. Distinguishing proteinaceous from other particles is important for assessing stability and for controlling other potential risk factors (see [52]). However, the irregular shape of proteinaceous particles and their aggregation make this task genuinely difficult. As a practical work-around, one tends to identify non-protein particles of known shape, and counts the remaining observed particles as aggregated protein.

A routine task is to identify silicon oil (SO) particles and distinguish them from all other, non-silicon oil (NSO) particles. SO particles are ball-shaped and are commonly present in prefilled syringes, cartridges or vials, as lubrication for the glass barrel, the plunger or the stopper. Because they look spherical from any viewing angle, classifying SO particles is considered feasible even when common FIM only captures 2D projections. Air bubbles are among the few kinds of entities, if not the only one, that also have spherical projections and can occur in the pharmaceutical products under consideration. However, their presence is avoided or minimized by sample preparation.

There are many types of possible NSO particles, which would make it difficult to compile a representative data-base of such particles for training purposes. On the other hand, images of SO particles are very similar, in particular after some pre-processsing. This makes it attractive to classify a future image by similarity to SO pictures, rather than train a supervised classifier, which would necessarily have to be trained on a non-representative set of example images.

From an industrial background preventing false positives (unjustly labeling particles as SO) is more serious than false negatives (labelling SO as NSO). This is because SO particles are not immunogenic, in contrast to some NSO particles.

In the industrial context classification algorithms are also referred to as *filters*.

5.3 Three Benchmark Approaches

The standard output from the MFI system (and similar systems) includes images and their corresponding metadata. The images consist of the intact picture frames taken in a defined frequency and sub-frames of “individual” particles. The sub-frames are captured in the intact frames by some generic algorithm in the MFI system. Below we take this capturing as given and refer to the sub-frames when using the word "images", unless specified otherwise. The metadata contains features of each sub-frame, including morphological (e.g. aspect ratio, perimeter and circularity) and optical aspects (e.g. intensity mean, SD, min and max across pixels), together with some descriptive information (e.g. from which intact frame, ECD, area, coordinates of the sub-frame in an intact frame).

Existing filters of SO particles ([53],[46]) are all based on the metadata. We discuss two industry standards, and a filter based on random forests.

Filter 1 (Benchmark-SRD)**.** *Strehl et al. [53] propose a filter based on five attributes from the metadata. They first calculate an ‘S-factor’ as the product*

$$\textbf{\textit{S-factor}} = Circularity \times AspectRatio \times IntensityMax \times IntensitySD.$$

They next determine a cut-off for the S-factor that depends on the particle size (ECD). But since their original work were for the MFI 4100 system, we use the following cut-offs that are suited for the MFI

5200 system, updated by as Torres et al. [47].

$$\textbf{\textit{Cut-Off}}(ECD) = \begin{cases} 29732 \times ln(ECD) + 3913, & 1.125 \leq ECD < 2.875 \\ 46609 \times ln(ECD) - 10464, & 2.875 \leq ECD < 6.125 \\ 3368 \times ECD + 49124, & ECD \geq 6.125 \end{cases}$$

If the S-factor of a sub-frame is larger than the Cut-Off value, then the image is classified as "SO", and otherwise as "NSO".

Filter 2 (Benchmark-AR)**.** *A sub-frame with Aspect ratio (AR)* ≥ 0.85 *is classified as "SO", and otherwise as "NSO". This simple, empirical rule is sometimes used in practice.*

Filter 3 (Benchmark-RF)**.** *In [46] the authors propose to use a random forest classifier based on the metadata. As their training data are not accessible and it may be worth while to tune the random forest to typical data in our domain, we fitted a random forest classsifier to our own dataset. We compiled the metadata of samples of 1990 NSO and 2316 SO particles, including the 10 predictors: ECD, Area, Perimeter, Circularity, Aspect Ratio, Maximum Feret Diameter, Intensity mean, STD, min and max from MFI 5200. The sample of SO particles included carefully selected SO particles and data from experiments measuring only SO particles. The sample of NSO particles was from a study where the presence of SO was minor and experts had carefully excluded the SO particles. We fitted a random forest model by the function* randomForest() *from the R package {randomForest, 4.6-14}, using the default setup, with the maximum number of terminal nodes (*maxnodes*) as a tuning parameter. The value 4 for* maxnotes *was selected via a cross-validation procedure, to minimize the chance of overfitting (see Figure 5.1). The prediction accuracy was 99.6% to the whole training set. Two important features in the fitted model, as measured by the mean decrease in Gini index when dropping the predictor, were the maximum intensity across all pixels and the aspect ratio.*

5.4 Data and Image Pre-processing

Even though SO images acquired by the MFI system share some key feature, they also display visible differences. Inspection of a set of images (see Appendix 1) reveals that, within their spherical shape, SO particles display an inner white spot. This is due to the reflection of camera light. The position of this spot depends on the position of the SO particle in the intact picture frame. Its size varies with the size of the particle, and its shape is not necessarily spherical. When the white spot is close to the edge of the spherical shape and the resolution of the image is low (which is typical for small particles), the spherical shape may be gently breached. The spot can be very small or blur into the background, in particular in the images of small particles. Thus small SO particles tend to show different patterns than bigger SO particles.

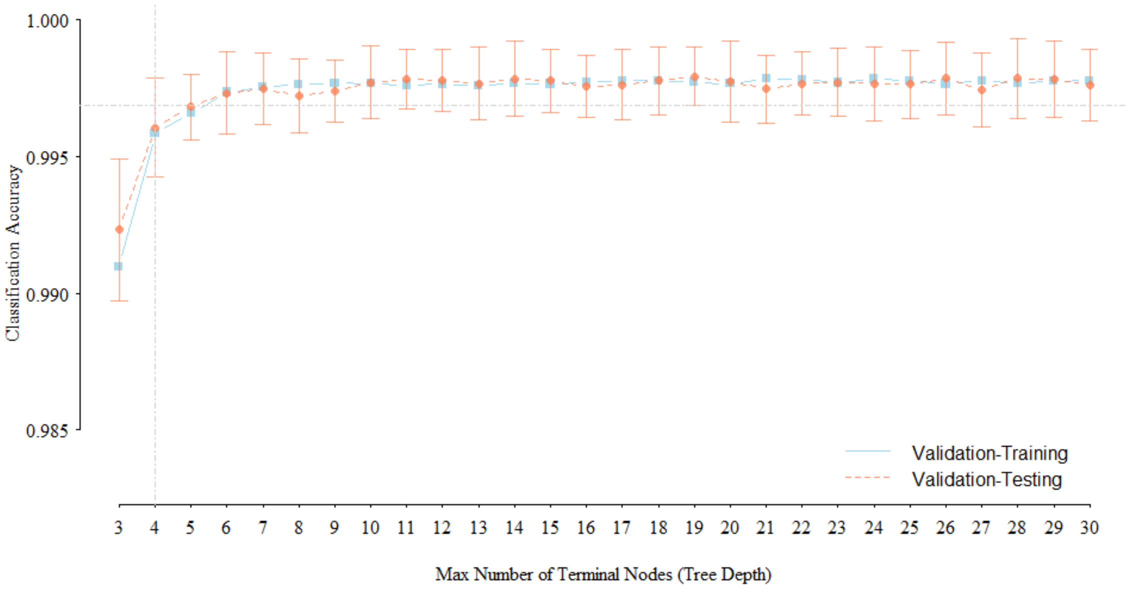


Fig. 5.1 Cross-validation curve to determine the maximum size of terminal nodes in benchmark filter 3. The data were randomly split 70:30 % in training and testing samples, with 50 repeats at each candidate setting of the variable *maxnodes* (shown on the horizontal axis). The vertical axis gives the proportion of correctly classified samples, with a blue square indicating the average in the training set, a red dot the average in the test set with red bar indicating the standard deviation across 50 repeats. The horizontal gray dashed line marks the lower end of the red bar at the candidate setting with the highest average accuracy. The selected value 4 is the minimum integer for which the red bar contains the horizontal gray dashed line.

Algorithm 6: Pre-processing of raw images to a square image

```
Data: A $a \times b$ image/matrix $X$ and an integer $h$.
1 Rescale the values in image $X$ to improve the contrast:
    $\forall j,k, \quad X_{jk} := (X_{jk} - \min(\{X_{jk}\}_{j,k})/(\max(\{X_{jk}\}_{j,k}) - \min(\{X_{jk}\}_{j,k})))$ ;
2 Chop off empty edges: for every row and column, determine its average and range length
    (maximum-minimum), chop off the row(s) and column(s) if the average is >0.78 and the
    range length is <0.28;
3 Repeat step 1-2 one more time, resulting in a $a' \times b'$ matrix;
4 Pad the image close to a square before resizing to avoid pattern distortion: If $|a' - b'| \leqslant 1$, no
    padding is needed. If $a' > b' + 1$, pad $\lfloor |a' - b'|/2 \rfloor$ blank columns to each of the two sides of
    the matrix. If $b' > a' + 1$, pad $\lfloor |a' - b'|/2 \rfloor$ blank rows to each of the two sides of the matrix.
    A blank column/row has elements equal to the average of the maximum of the first
    column/row and the maximum of the the last column/row;
5 Resize the matrix to an $h \times h$ matrix by linear interpolation, to ensure comparability across
    images of different sizes: resize() from R package {EBImage, 4.22.1} was used in this
    chapter;
Result: An $h \times h$ matrix
```

Features such as these are not captured in the metadata of the images, which motivates to build a filter based on the complete images. To train our image-based (IB) filter, we selected 798 SO images from the MFI system, aiming to include a representative set to cover the variation in the aforementioned features (e.g. images of big and small sizes, the inner white spot having various positions, shapes and sizes, the spherical hull gently breached by the white spot).

An image of $a \times b$ pixels can be viewed as an $a \times b$ matrix of numbers with values between 0 and 1, where 0 indicates pure black, 1 indicates pure white, and intermediate values correspond to gray with different degrees of hue. We pre-processed the images to cut off empty edges, thus centering the SO particle in the images, and to correct for the varying size and contrast and of the images. The procedure is illustrated in Figure 5.2 and described precisely in Algorithm 6. This pre-processing is meant to be consistently applied to all images, also future ones, and hence is part of the filter constructed here.

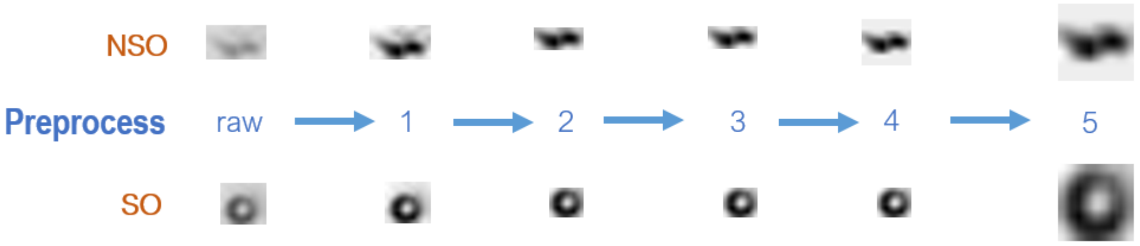


Fig. 5.2 Illustration of the pre-processing procedure for images

As a convention, images of particles of sizes smaller than 5 μm (as determined by the variable ECD in the metadata of the image) are not processed by any SO filters. Their low resolution prevents making a confident judgement, also by human experts after close scrutiny. Images of particle of sizes exceeding $5\mu m$ usually have at least 7 rows or columns of pixels.

5.5 Construction of Image-Based (IB) Filter

Our objective is to build an accurate binary filter (classifier) for SO/NSO assignment using full images as inputs. We approach this from the perspective of unsupervised learning, because images of NSO particles, in contrast to SO particles, do not have a clear description and can take infinitely many forms. Therefore, it would be difficult to form a good training set, and a filter based on supervised learning would not be able to generalize its success against NSO stereotypes that are not seen in the training data.

On the other hand, SO particles have fairly predictable 2D projections from any angle, making it possible to use similarity of a new image to images in the training set the basis for classification.

Our filter consists of two sub-filters (or ‘tests’), each of which outputs a binary label SO/NSO. A future image will be labeled as SO if either or both tests label it as SO. The first filter uses data depth to assess similarity of a future image to a set of training SO images. The data depth is computed on features of the images extracted by a principal component analysis. The second filter is more crude and checks a pixel-wise distance of a future image to set of SO stereotypes in a reference book of nonstandard SO images.

We found that the first filter rarely misclassified an image as NSO while it was SO. On the other hand, the second filter was good at recognizing some nonstandard SO images, which were missed by the first filter. Details on how each test was constructed are presented in the following subsections.

5.5.1 IB-Test 1: General SO Pattern Recognition

After pre-processing, an image is represented as a 20×20 matrix, and can be identified with a vector in 400-dimensional space. The first filter is based on measuring the data depth of a future vector in the distribution of a typical vector obtained from an image of a SO particle. Data depth enables ranking of multivariate data in a center-outwards manner, low depth indicating an outlier, and can be used as a measure of similarity. In the present case lower depth means that the future image does not represent a SO particle. See Section 3.3 for a detailed description on depth and its usage in building nonparametric prediction sets.

To reduce noise, the 400-dimensional image vector is first projected on the linear span of the first 20 principal components of the distribution of the SO particles. If the projection of a future image is very different from the original image, then this is a first indication that it does not represent a SO particle (see Figure 5.3). We used the 99.5% percentile of the distribution of the distance between a typical vector obtained from an image of a SO image and its projection as a cut off for future images.

The loadings of the 400-dimensional image vector on the first 20 principal components is a vector in 20-dimensional space. The distribution of a typical vector obtained from an image of a SO image was approximated by a mixture distribution, and data-depth of a future vector is computed relative to all mixture components separately. The maximum data depth over the 30 components is taken as the final similarity measure. The maximum is taken, because the component weights resulting from the training set are arbitrary, while a future image should be labeled as "SO" if it is close to any SO stereotype. As a cut off for classification of a future image as SO, we chose again the 99.5% quantile of the similarity measure in the training data.

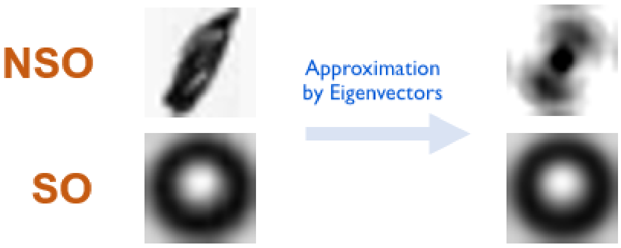


Fig. 5.3 Illustration of Bad (top row) and Good (bottom row) approximation by extracted image features.

We next elaborate on these steps.

The basic training set included 798 SO images, of various sizes. After pre-processing, each image was represented by a 20×20 matrix, with entries between 0 and 1. The training set was augmented by rotating every matrix by 90, 180 and 270 degrees, gving a total of 3192 matrices. Each matrix was vectorized by sequentially concatenating its column vectors. The resulting 3192 vectors were joined as rows in a data matrix X of dimensions 3192×400.

Principal component analysis (PCA) was performed using the covariance matrix $X^T X$ of the vectorized images, yielding (orthogonal) eigenvectors $v_1, \cdots, v_{400}$ (or ‘image features’), scaled to unit length ($\|v_i\| = 1$), and corresponding eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_{400}$. The first 20 eigenvectors, which preserved 98% of the total variation, were retained, and are visualized in Figure 5.4. (In the picture the values in each eigenvector were rescaled to the interval $[0,1]$, and the vectorization was reverted back to a 20×20 matrix.) The visualization enhances the interpretation of the eigenvectors. For instance, the first eigenvector (top-left of Figure 5.4) captures the spherical shape of the droplet and the inner white spot, the second and the third eigenvectors (the second and the third in the first row of Figure 5.4) encapsulate the different camera lighting angles on the droplet, and the forth and fifth features represent a diagonal shift of the inner white spot. The decision to retain the first 20 eigenvectors depended, besides on explained variance, on these graphical evaluations. The 21^{th} and onward eigenvectors showed features that are not desired for a typical SO image (e.g. vertical or horizontal elongation of a sphere).

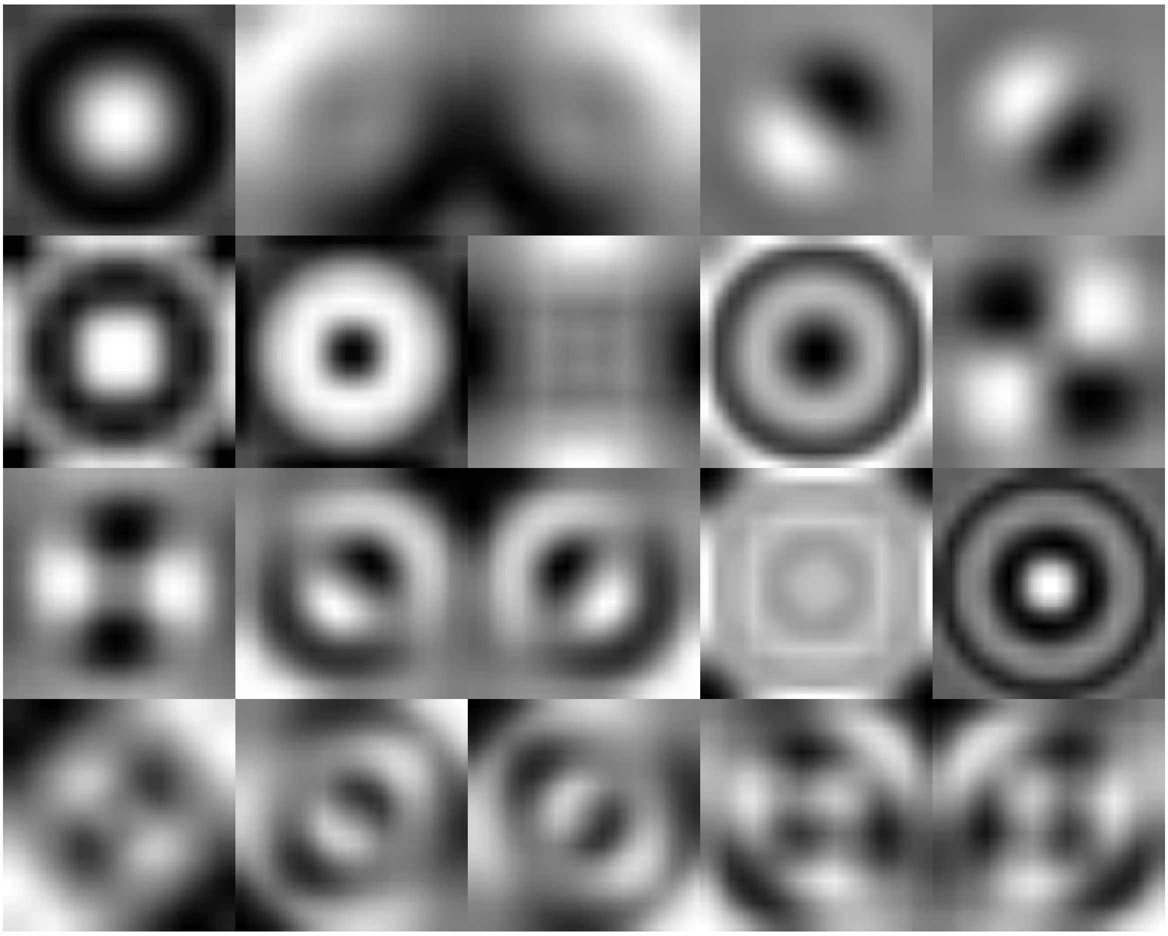


Fig. 5.4 First 20 (rescaled) eigenvectors of the covariance matrix of the 3192 training data (798 pre-processed 20×20 images with 3 augmentations for each via rotation). The first 5 eigenvectors are in the first row from left to right, the second 5 in the second row, and so on.

For $V_k = [v_1, \cdots, v_k]$ the $400 \times k$ matrix with as columns the first k eigenvectors, the loading matrix of the original data can be computed as $\Lambda_k = XV_k$. (We used $k = 20$.) The projection of a training image $X_{j\cdot}$ (1×400 vector, j^{th} row of X) on the first k eigenvectors is given by $\sum_{i=1}^{k}(X_{j\cdot} v_i)v_i$, which leads to the matrix $X^* = \Lambda_k V_k^T$ for the projections of all training images relative to the standard basis. The squared Euclidean distanced between the original training images and their projections (the squared distances between the rows of X and X^*) ranged from 0.58 to 2.04, with 99.5% sample percentile approximately equal to 1.4. The latter sample percentile was chosen as the upper threshold for assessing whether a future image is well approximated by its projection on the eigenvectors (and hence similar to SO), or not.

We clustered the loadings (the rows of Λ_{20}) of the training data using the K-Means algorithm, with $K = 30$ clusters. The latter value was based on evaluation of the average patterns across the clusters, with a focus on assessing their variability and avoiding duplication. Intuitively, increasing the number of clusters (i.e. choosing larger K) would enhance the ability to recognize a future SO image, but lower the cardinalities of the clusters and increase the chances of duplicated stereotypes. In the end we determined that 30 clusters are appropriate in this case, leading to cluster sizes between 44 and 163. Trimmed means of the images of the 30 clusters are shown in Figure 5.5.

For each of the 30 clusters, the mean vector and covariance matrix of the loadings of the training data were calculated, and next the 0.5% percentile of Mahalanobis depth was derived via simulation. The details are as follows. Let $c_j \in \{1, \cdots, 30\}$ be the member indicators of the training data ($j = 1, \cdots, 3192$), and $B_l = \{j\colon c_j = l, \forall j\}$ be the indices of the clusters. For each cluster, we simulated 5000 data points from the multivariate Gaussian distribution with parameters the sample mean and sample covariance matrix of the vectors $\{\Lambda_{j\cdot}\}_{j \in B_l}$, and then calculated 0.5% percentile of the Mahalanobis data depth of these simulated data relative to $\{\Lambda_{j\cdot}\}_{j \in B_l}$. Taking 0.5% percentile as the lower-sided threshold gives a 99.5% prediction set per cluster, since larger depth means the data point is closer to the center. Across the 30 clusters these 0.5% percentiles range from 0.023 to 0.025, which motivated a lower threshold of 0.02 for the maximum depth over the 30 clusters of a future image.

The projection and loadings of a future image, and its data depth to the clusters, can be computed similarly. The filter will consider the image to be SO if the square Euclidean distance between the image and its projection is small enough and its depth relative to at least one of the clusters is large enough. The two thresholds were set at 1.4 and 0.02, respectively, both reflecting a 99.5 % prediction set based on the training data.

In summary, the first part of our filter is as follows.

IB-Filter: Test 1. *A future image is pre-processed in the manner described in Section 5.4 and vectorized to a* 1×400 *vector* y. *Its loadings to the extracted 20 features* V_{20} *are calculated as* $\eta_y = yV_{20}$. *The projected image is calculated as* $\eta_y V_{20}^T$ *and its square distance to the original image is* $d_{PC}(y) = (y - \eta_y V_{20}^T)(y - \eta_y V_{20}^T)^T$. *The Mahalanobis depths relative to the clusters* $l = 1, \cdots, 30$

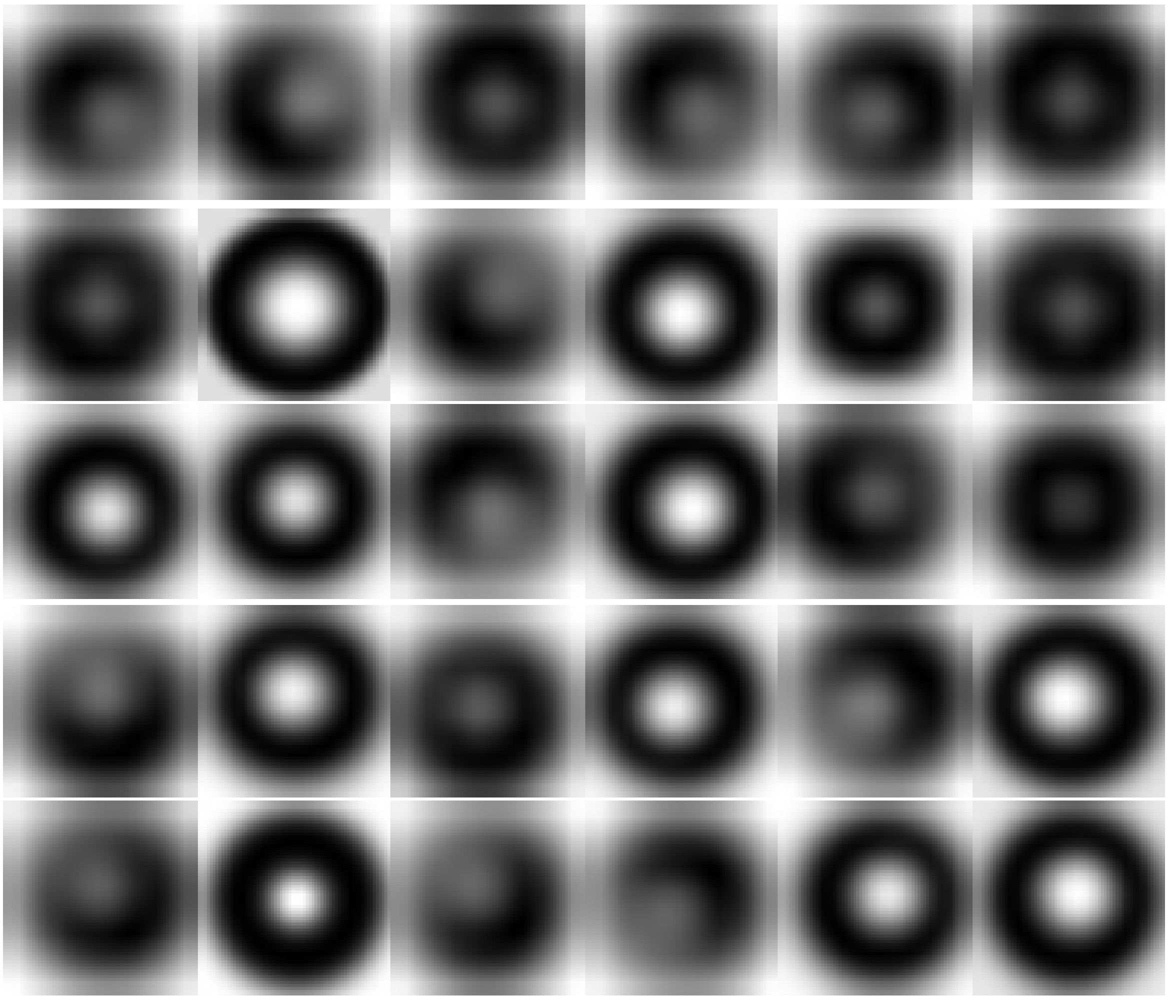


Fig. 5.5 Trimmed means of the 30 clusters obtained via 30-Means applied to the loading matrix. Trimmed means were computed as pixel-wise averages, after discarding the 10% most extreme values.

are $MD_{y,l} = MD(y, \{\Lambda_{j\cdot}\}_{j \in B_l})$*, giving the maximum depth* $MD_y = \max(MD_{y,1}, \cdots, MD_{y,30})$ *as the final similarity measure. The future image is considered to be a SO particle if* $d_{PC}(y) < 1.4$ *and* $MD_y > 0.02$.

5.5.2 IB-Test 2: Reference books of SO stereotypes

The second part of the filter is to check similarity of a future image against a reference book of SO stereotypes. The rationale of this test is our expectation that the first part of the filter might neglect some important nonstandard (unpopular) SO patterns (e.g. non-spherical white spot for a modestly large SO image, or a small SO image with gently breached sphere hull).

This test is performed separately on 'large' and 'small' images, small images being defined as images that originally had fewer than 9 rows or columns. The reason is that small-size SO images tend to show more idiosyncratic patterns than modest or large ones (discussed at the start of Section 5.4), and sizing them up to 20×20 would introduce profound distortion. Instead we resized smaller images to 9×9, by the same pre-processing procedure discussed in algorithm 6.

Thus two reference books (or 'galleries') of SO images were formed, one of 68 images of 20×20 pixels and a second one of 33 images of 9×9 pixels (see Appendix of the chapter). A future image is pre-processed to either 20×20 or 9×9, based on whether its number of rows and columns is greater than 10 or smaller than 9, and its pixel-wise distance is computed to all images in the corresponding reference book. If the minimum distance is smaller than a threshold, the future image is classified as SO. The distance employed was the L_1-distance between the pixel values, and the threshold was set to the 99% and 90% percentiles of the distances of the images in the training datasets, for larger and smaller images, respectively.

The reference books were prepared from training datasets, as follows. The training dataset included the same matrix of dimensions 3192×400 containing as rows the vectorized 20×20 images, used to construct the first part of the filter, and in addition a 2060×81 matrix, in which each row is a vectorized 9×9 image. The second matrix was prepared from the 419 'small' images in the original training sample of 798 images, resized to 9×9 pixels by algorithm 6, and augmented to a set of 1676 images by rotating each image over 90, 180 and 270 degrees. The data matrices were reduced in dimension by PCA and next clustered to 30 clusters by the K-means algorithm applied to the rows of the loading matrix, following the procedure described in Section 5.5.1. For small pictures we retained 22 principal components, explaining 99% of the variation in the loadings (see Figure 5.6 for a graphical representation of the eigenvectors) and also 30 clusters. The 30 clusters means were placed in initial reference books, separately for large and small images, which were next enlarged as follows. We computed the L_1-distance

$$d(x,y) = \sum_{i=1}^{h} |x_i - y_i|,$$

between every training image (row of the data matrix) and every element of the initial reference book (cluster average), where $h = 400$ or $h = 81$ for large and small images, respectively, and selected the 10 % images that were most deviant from the reference book, i.e. had the largest minimum distance. These images were added to the iniital reference books. Following visual inspection the latter sets of images were next reduced to sets of 68 and 33 images. Visual inspection focused on the picture clearly being a SO particle, and, for instance, excluded pictures with a severe breach of the spherical shape of the SO particle.

Having thus formed references books $R^{(20}$ and $R^{(9)}$ of large and small images (matrices of dimensions 68×400 and 33×81), we computed the minimum L_1-distance of all training images to the corresponding reference book, and choose the 99% and 90% percentiles of these minima (38 and 6, respectively) as cut offs for the L_1-distance of a future distance. Choosing 90% percentile as the cut off for small images is due to our finding of large portion of unsatisfactory SO pattern during the visual inspection for its final reference book.

In summary, the second part of the filter is as follows.

IB-Filter: Test 2. *A future image of dimensions* $a \times b$ *is pre-processed to* 20×20 *pixels if* $max(a,b) > 9$, *and to* 9×9 *otherwise, in the manner described in Section 5.4, and vectorized to a* 1×400 *or* 1×81 *vector* y. *The* L_1*-distance of* y *to every row of the reference book* $R^{(h)}$ *is computed, where* $h = 20$ *or* $h = 9$ *for large and small images, respectively. The distance* $d_{REF}(y, R^{(h)})$ *of* y *to the reference book* $R^{(h)}$ *is defined as the minimum of these distances. If* $d_{REF}(y, R^{(h)}) < c$, *the image is labeled as SO, otherwise NSO. Here* $c = 38$ *when* $h = 20$ *and* $c = 6$ *when* $h = 9$.

5.6 Performance Evaluation

Three groups of data were used to evaluate the performance of the benchmark filters and our proposal, categorized as: (I) clearly SO, (II) unclear SO, (III) clearly NSO. Table 5.1 summarizes these data. These three groups correspond to common situations, which a specialist faces in practice when manually labeling an image as SO or NSO. For each group, the metadata were used for the benchmark filters, and the raw images were the input to our image-based filter. The majority of test images are small to modest-size (almost all particles are $< 20\mu m$), with around 30% of the images falling into the small category (number of rows and columns ≤ 9). This is useful for testing purposes, since the main challenge will be in recognizing SO images of lower resolution. Simple cases of NSO particles, e.g. very elongated images, were also excluded from the datasets. The performance of the filters is expected to be significantly better when applied to real sample measurements.

The filters were compared on their false negative rate (i.e. error in labeling SO as NSO) for test sets (I) and (II), and on their false positive rate (i.e. error in labeling NSO as SO) for test set (III). As SO particles are not immunogenic, while some types of NSO particles could cause immunogenic

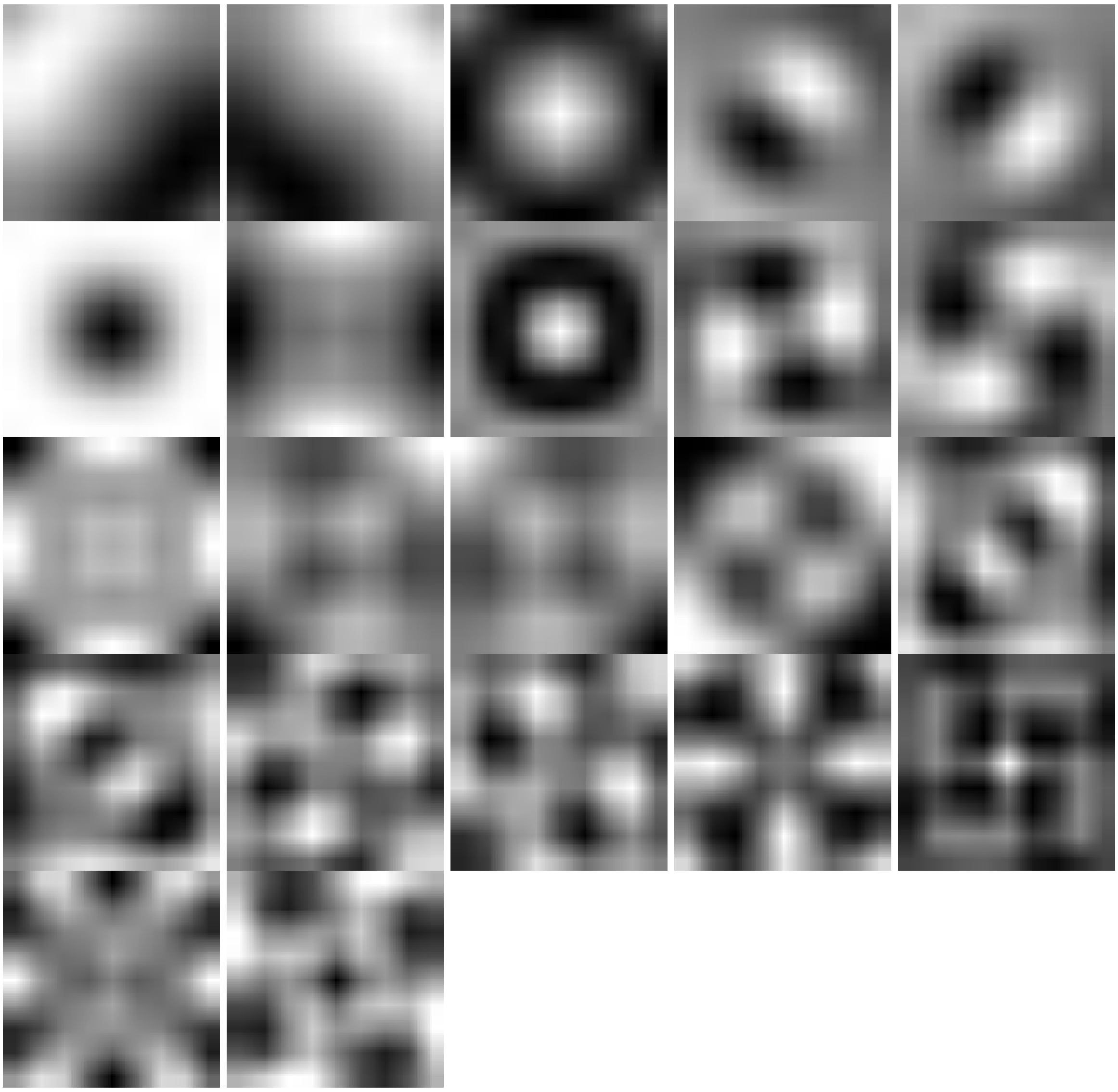


Fig. 5.6 First 22 (rescaled) eigenvectors of the covariance matrix of 1676 × 81 training data (419 pre-processed 9×9 images with 3 augmentations for each via rotation). The first 5 eigenvectors are in the first row from left to right and the sequence follows.

Table 5.1 Summary of Testing datasets

			Small Image		Not-Small Image	
	Category	Selection Criteria	#	ECD (99% perc.)	#	ECD (99% perc.)
(I)	Clearly SO	Recognize easily	233	7.375	700	16.375
(II)	Unclear SO	Moderately certain	537	7.125	373	8.625
(III)	Clearly NSO	Recognize easily	263	6.47	1361	19.375

concerns for bio-pharmaceuticals, the false positive rates could be considered as the more important criterion. The results are summarized in Tables 5.2–5.4, and visualized in Figure 5.7.

Table 5.2 Misclassification rates (False Negative) in group (I): Clearly SO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	233	0.013	0.021	0.652	0.021	0.266	0.009
Not-Small	700	0.163	0.051	0.491	0.099	0.023	0.017
Total	933	0.125	0.044	0.532	0.079	0.084	0.015

Table 5.3 Misclassification rates (False Negative) in group (II): Unclear SO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	537	0.084	0.324	0.873	0.235	0.378	0.212
Not-Small	373	0.496	0.603	1.000	0.464	0.249	0.206
Total	910	0.253	0.253	0.925	0.329	0.325	0.210

Table 5.4 Misclassification rates (False Positive) in group (III): Clearly NSO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	263	0.517	0.319	0.015	0.137	0.076	0.163
Not-Small	1361	0.248	0.282	0.324	0.012	0.071	0.071
Total	1624	0.291	0.288	0.274	0.032	0.071	0.086

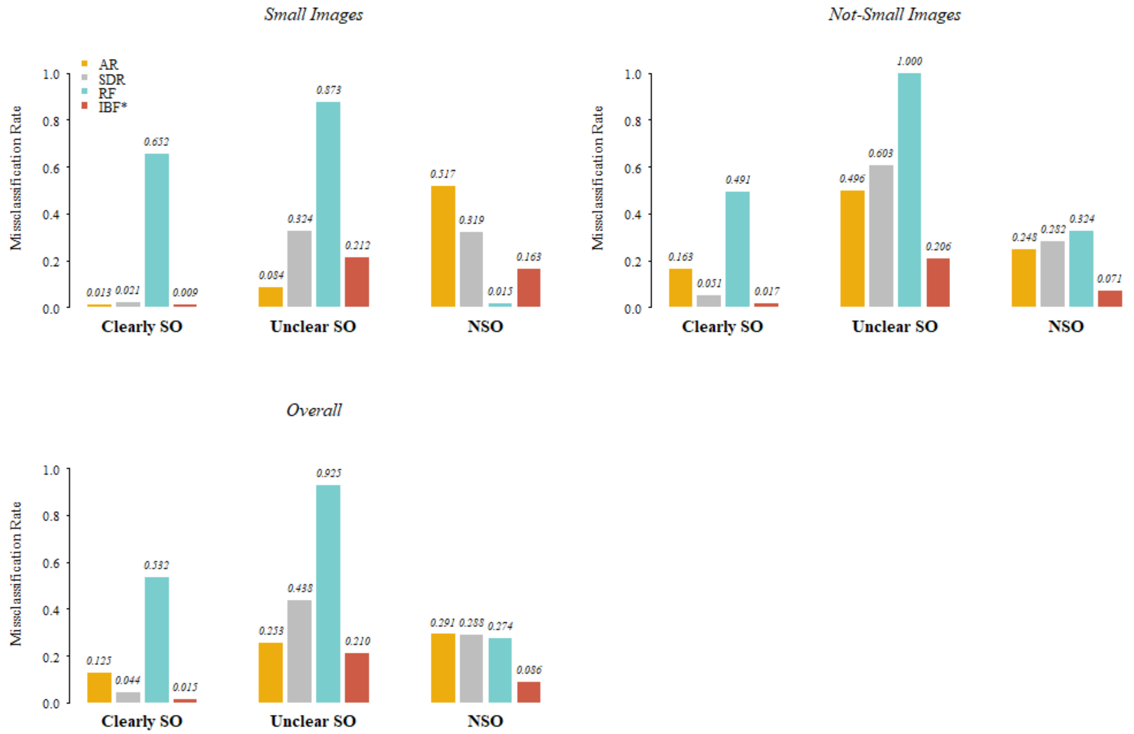


Fig. 5.7 Misclassification rates of the three benchmark filters (AR, SDR, RF) and our IB-filter for the three testing set (Clearly SO, Unclear SO and Clearly NSO)

5.7 Conclusion

As shown in the previous section, our IB-filter keeps a good balance in trading off the false positive and false negative rates. It outperforms the benchmark filters in almost all cases, the exceptions being that it gives the second lowest false negative rate for small unclear SO images (trailing AR) and the second lowest false positive rate for small NSO images (trailing RF). AR achieves the lowest error rate in labeling small unclear SO at the cost of having the highest error rate in labeling small NSO images. RF achieves its excellent false positive rate for small images by labeling the majority of testing set as NSO, which leads to very high error rates for classifying small clearly SO images (65.2%) and unclear SO images (87.3%).

Although it is not the focus of this chapter, we further studied the highly unsatisfactory performance of RF, which may be surprising, as the error on the training set was only 0.4%, while the tree sizes had been properly tuned by cross validation to prevent overfitting. We suspected that the key issue was that the NSO images in the RF training set were obtained from different materials than the NSO material in the testing set, giving little overlap between the NSO images in training and test sets. We prepared another training set for RF, consisting of the original 2316 SO samples and a new set of 1317 NSO samples, obtained from measurements from similar material as used for the NSO samples in the test set. The maximum number of terminal nodes of the trees in the random forrest was determined at 11, by a similar cross-validation procedure as stated in the caption of Figure 5.1. The performance of the two RF models (training the original and the new dataset) is visualized in Figure 5.8. It can be observed that the error rate in labeling small NSO images increases, but the overall false positive and false negative rates decrease. This is a warning to practitioners that an RF-based SO filter will need to be trained per drug product. Even with this improvement, the performance with respect to the false positive rate comes close to our IB-filter, but still falls behind with respect to the false negative rates.

5.8 Discussion

This study confirms the potential of image-based filters based on the data depth of a future image in the distribution of SO images. In particular, our proposed filter makes a good trade-off between false positive and false negative rates. The advantages over the benchmark approaches come from two parts: (1) metadata has limited representation of image features that are important to distinguish SO and NSO particles, particularly for small particles, and (2) the benchmark filters were mostly built in a supervised manner.

We elaborate on the second point. The patterns of NSO images are arbitrary, as they are random 2D projections of free-shape 3D objects. A formulation of the binary classification problem SO/NSO as a supervised learning task is therefore ill-posed. A training dataset will necessarily differ from

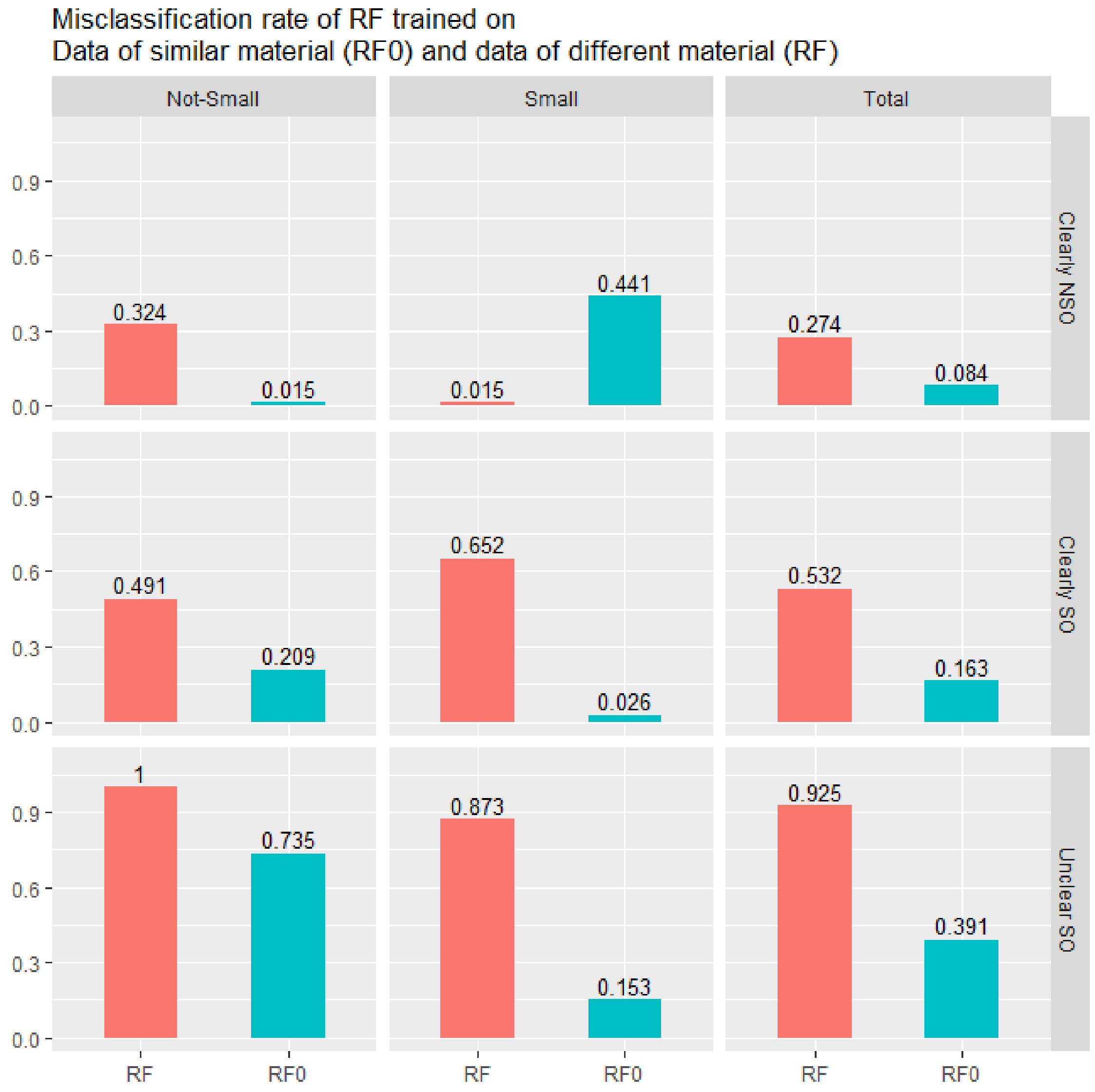


Fig. 5.8 Misclassification rates of the two RF models across the three testing sets. The original RF model (RF, red bars) was trained on NSO images of a material different from the NSO testing set. The second model (RF0, blue bars) was trained on NSO images from a similar material as the NSO testing set.

future data in essential features, and the usefulness of a supervised filter will be limited to a restricted domain.

By choosing to train a filter based only on SO images, we accept that there will be some patterns of NSO images that overlap with the distribution of SO images, but this is acceptable if the primary focus is on recognizing particles that give strong evidence that they are are SO. The overlapping patterns introduce an inevitable trade-off between false positives and false negatives, which is shared by any image-based or metadata-based filters. We learnt that this challenge is more serious for small particles, since small SO images have less regular patterns, e.g. the spherical hull may be breached by the light reflection, or the white spot may all but disappear.

Another minor merit of pursuing an image-based filter is its applicability across different versions of MFI systems. For instance, our image-based filter is theoretically applicable to images in gray scale, regardless of the type of flow image system it originates from. In contrast, metadata from different versions of MFI system may summarize the images under different definitions, even if the summarized features are labeled with the same names.

The following points can be improved in future work.

- While pre-processing is necessary to standardize input images of arbitrary size for analysis, the current algorithm for padding and resizing can be improved with respect to distortion of the original patterns.
- Although this study is adequate to demonstrate the value of an image-based filter, more careful selection of a larger set of SO training images could easily improve performance. Furthermore, instead of our distinction between small and particles (number of rows and columns smaller than 10), a systematic study may be carried out to determine appropriate bin sizes.
- The settings of the first part of the filter (Test 1) in our image-based filter may be further optimized. Instead of using the maximum depth across 30 clusters, one may seek the average of the highest two depth values, or measure another degree of agreement with several key clusters. The intuition is that, if the testing image has a standard SO pattern, it is likely to get high depth values in more than one cluster.
- One may develop a better algorithm to extract individual images from the intact images, in order to minimize the chance of presence of multiple particles in one image, or truncation of a particle.

5.9 Appendix 1: Gallery on variations of SO/NSO images

This Appendix shows the gallery of selected SO and NSO particles: examples of Clearly SO images in Figure 5.9, Unclear SO images in Figure 5.10, and Clearly NSO images in Figure 5.11.

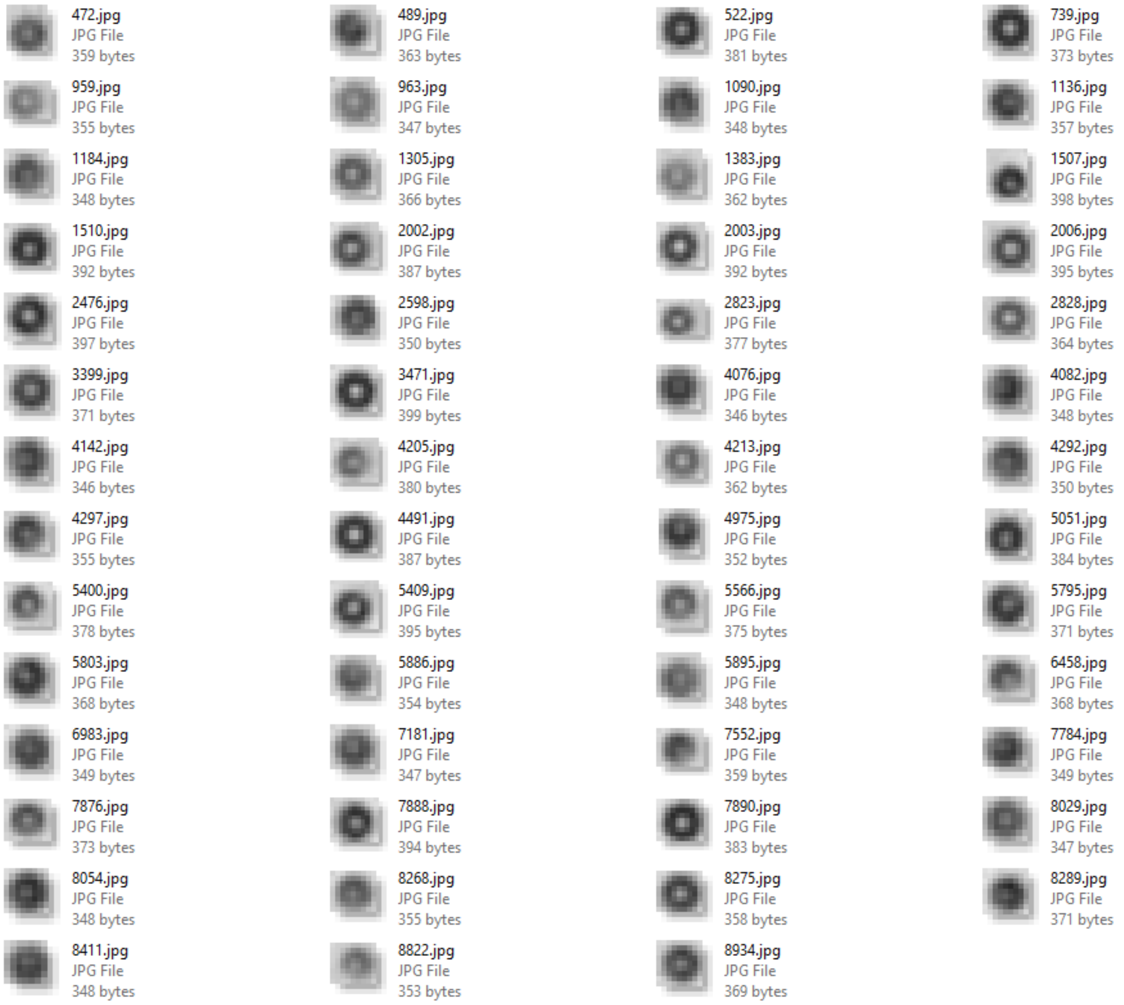


Fig. 5.9 Examples of Clearly silicon oil (SO) particles that can be easily recognized.

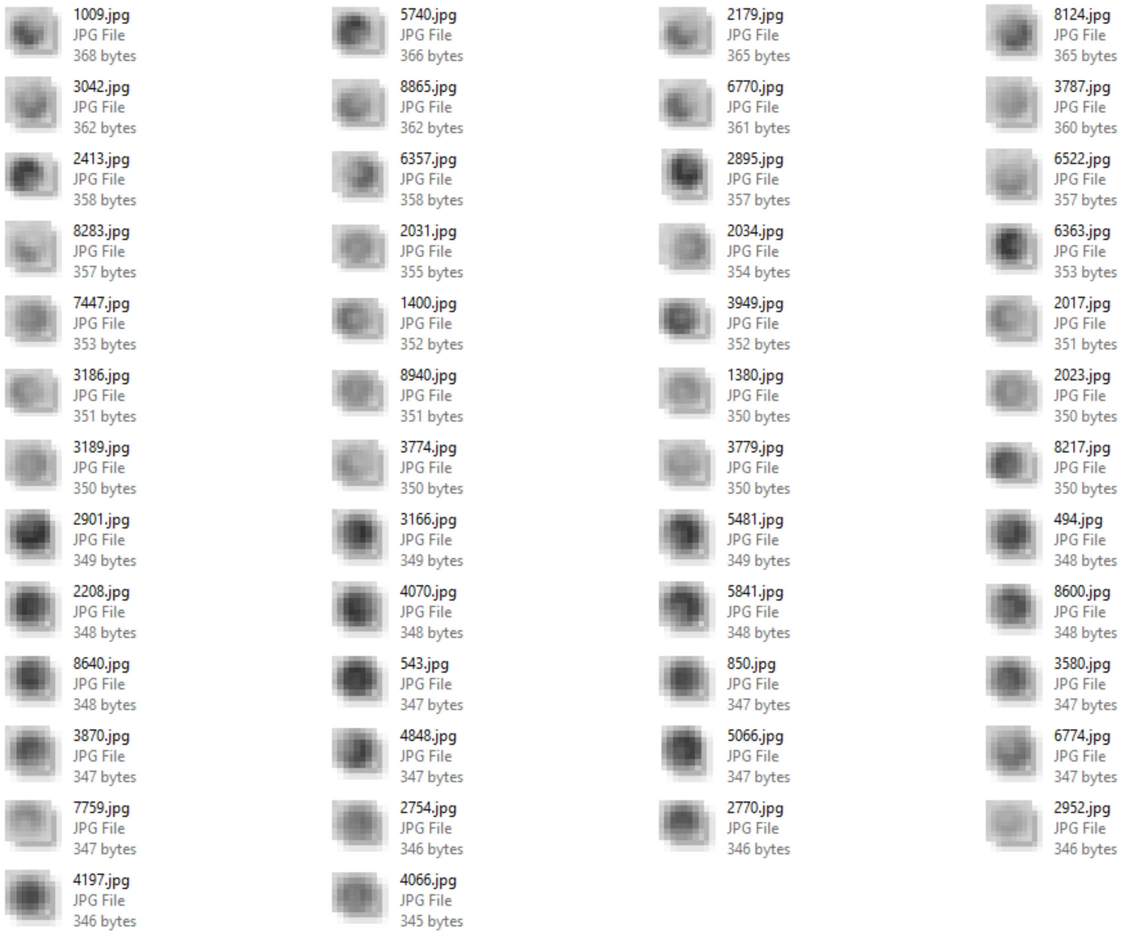
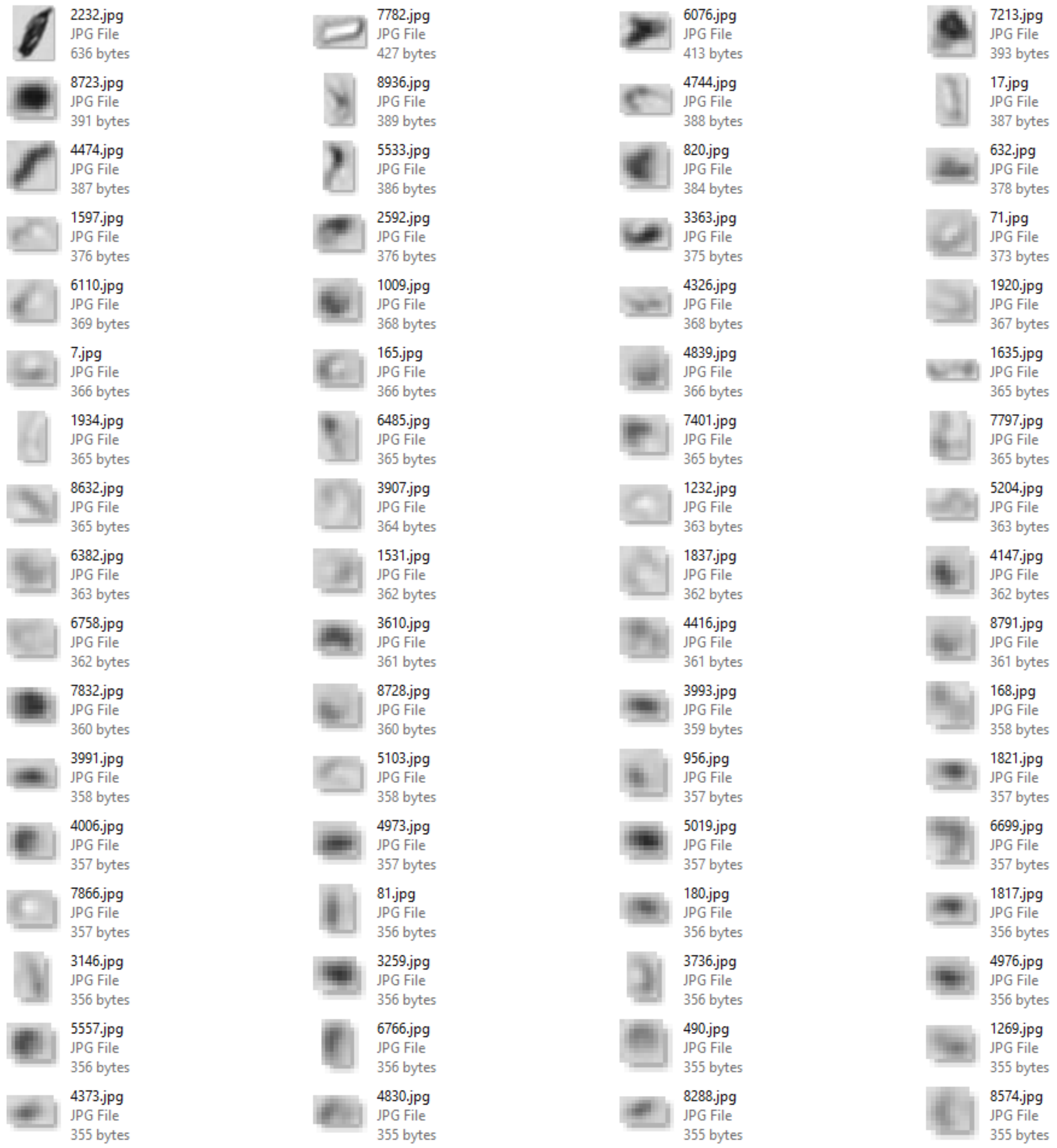


Fig. 5.10 Examples of Unclear silicon oil (SO) particles that experts are moderately certain that they are SO based on images.



```
2232.jpg     7782.jpg     6076.jpg     7213.jpg
JPG File     JPG File     JPG File     JPG File
636 bytes    427 bytes    413 bytes    393 bytes
8723.jpg     8936.jpg     4744.jpg     17.jpg
JPG File     JPG File     JPG File     JPG File
391 bytes    389 bytes    388 bytes    387 bytes
4474.jpg     5533.jpg     820.jpg      632.jpg
JPG File     JPG File     JPG File     JPG File
387 bytes    386 bytes    384 bytes    378 bytes
1597.jpg     2592.jpg     3363.jpg     71.jpg
JPG File     JPG File     JPG File     JPG File
376 bytes    376 bytes    375 bytes    373 bytes
6110.jpg     1009.jpg     4326.jpg     1920.jpg
JPG File     JPG File     JPG File     JPG File
369 bytes    368 bytes    368 bytes    367 bytes
7.jpg        165.jpg      4839.jpg     1635.jpg
JPG File     JPG File     JPG File     JPG File
366 bytes    366 bytes    366 bytes    365 bytes
1934.jpg     6485.jpg     7401.jpg     7797.jpg
JPG File     JPG File     JPG File     JPG File
365 bytes    365 bytes    365 bytes    365 bytes
8632.jpg     3907.jpg     1232.jpg     5204.jpg
JPG File     JPG File     JPG File     JPG File
365 bytes    364 bytes    363 bytes    363 bytes
6382.jpg     1531.jpg     1837.jpg     4147.jpg
JPG File     JPG File     JPG File     JPG File
363 bytes    362 bytes    362 bytes    362 bytes
6758.jpg     3610.jpg     4416.jpg     8791.jpg
JPG File     JPG File     JPG File     JPG File
362 bytes    361 bytes    361 bytes    361 bytes
7832.jpg     8728.jpg     3993.jpg     168.jpg
JPG File     JPG File     JPG File     JPG File
360 bytes    360 bytes    359 bytes    358 bytes
3991.jpg     5103.jpg     956.jpg      1821.jpg
JPG File     JPG File     JPG File     JPG File
358 bytes    358 bytes    357 bytes    357 bytes
4006.jpg     4973.jpg     5019.jpg     6699.jpg
JPG File     JPG File     JPG File     JPG File
357 bytes    357 bytes    357 bytes    357 bytes
7866.jpg     81.jpg       180.jpg      1817.jpg
JPG File     JPG File     JPG File     JPG File
357 bytes    356 bytes    356 bytes    356 bytes
3146.jpg     3259.jpg     3736.jpg     4976.jpg
JPG File     JPG File     JPG File     JPG File
356 bytes    356 bytes    356 bytes    356 bytes
5557.jpg     6766.jpg     490.jpg      1269.jpg
JPG File     JPG File     JPG File     JPG File
356 bytes    356 bytes    355 bytes    355 bytes
4373.jpg     4830.jpg     8288.jpg     8574.jpg
JPG File     JPG File     JPG File     JPG File
355 bytes    355 bytes    355 bytes    355 bytes
```

Fig. 5.11 Examples of Clearly non-silicon oil (NSO) particles that can be easily recognized.

5.10 Appendix 2: Gallery of the Reference book in Test 2 of IB-Filter

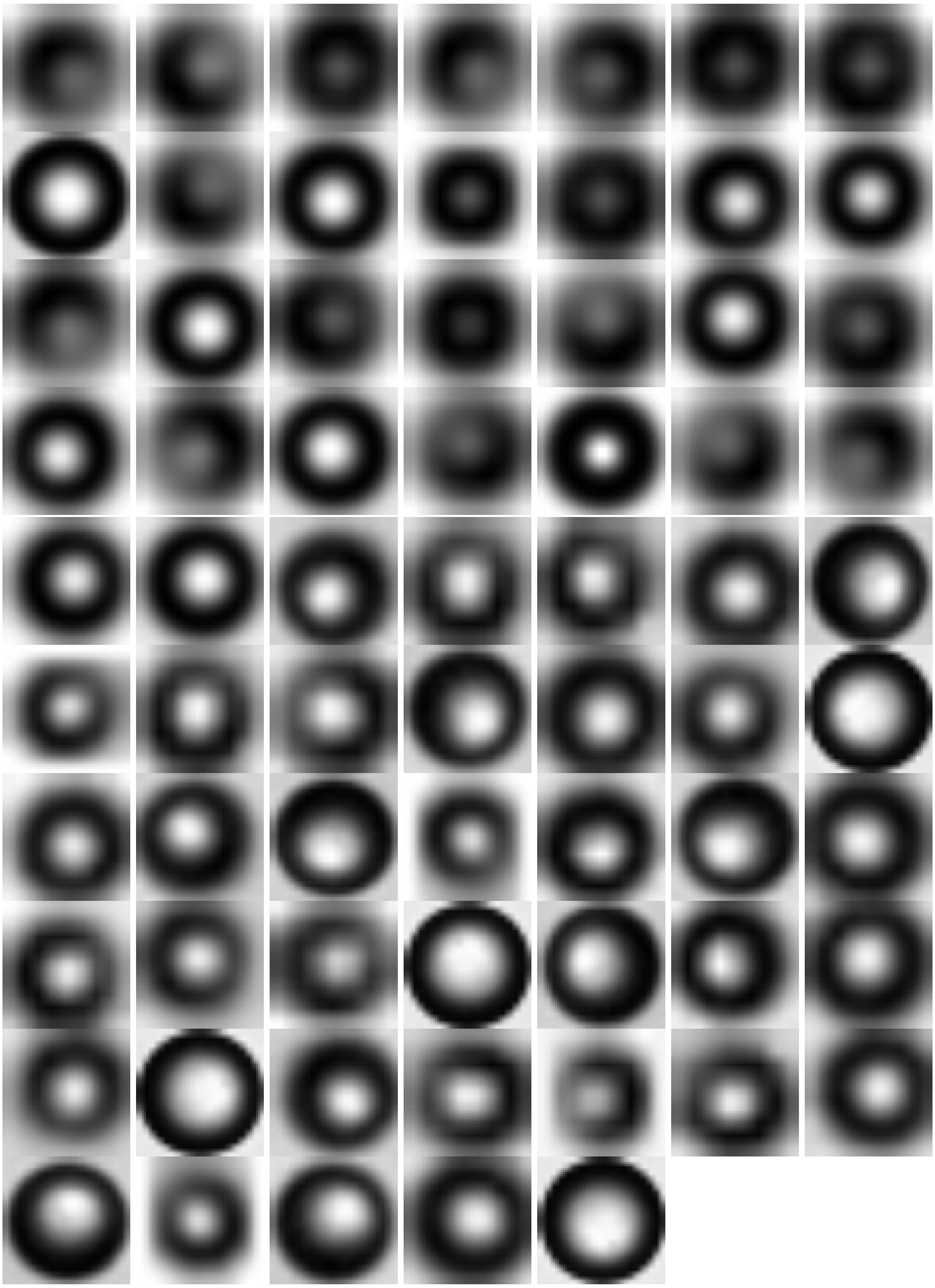


Fig. 5.12 Reference book of SO images of size 20×20.

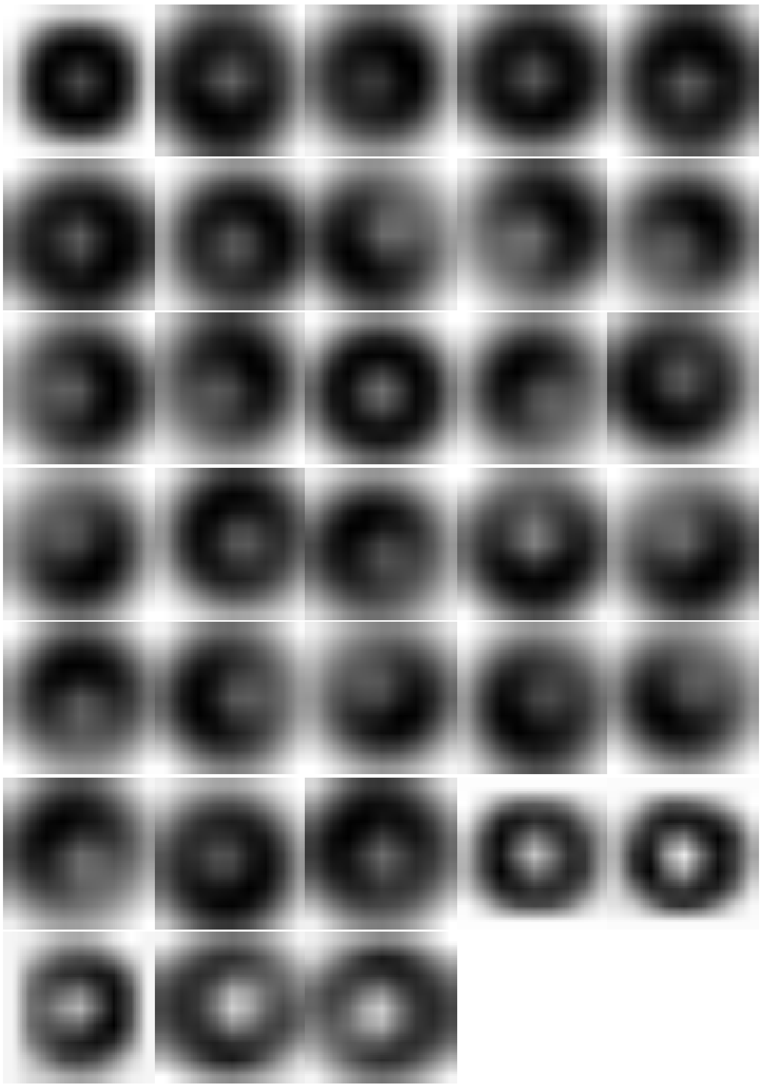


Fig. 5.13 Reference book of SO images of size 9×9.

5.11 Appendix 3: Gallery of example images from the testing dataset.

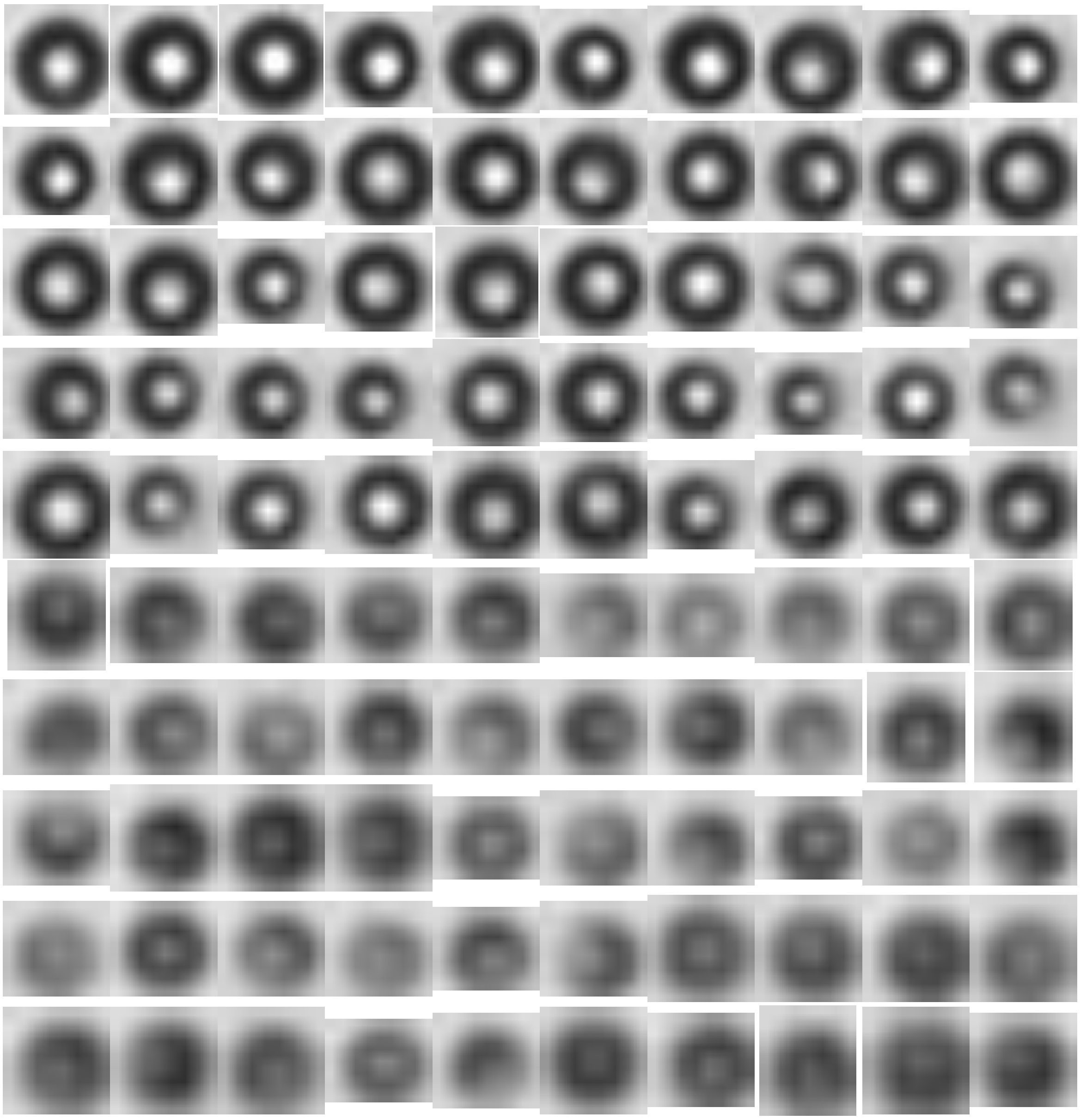


Fig. 5.14 Examples of Clearly Oil images in the testing set.

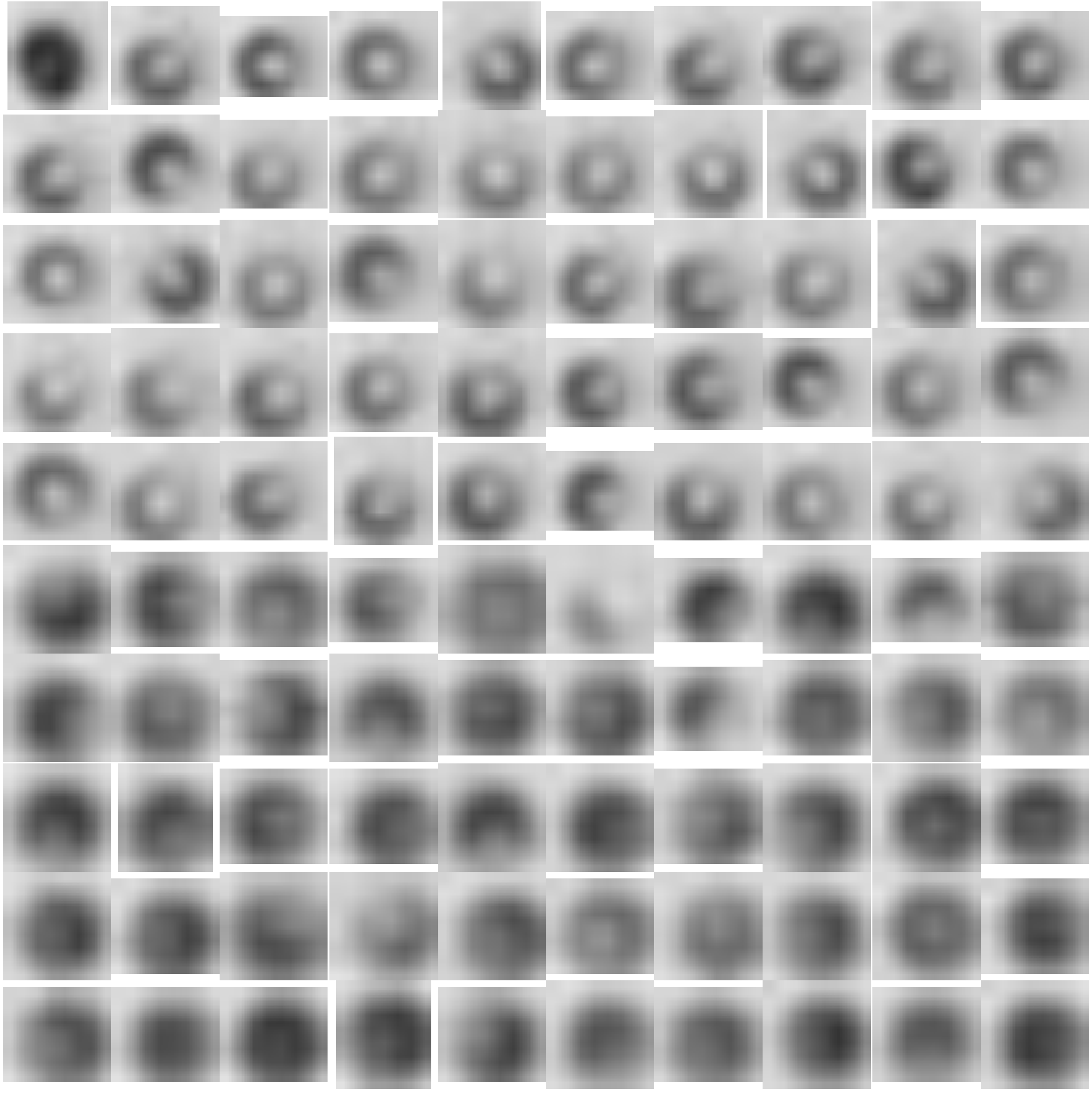


Fig. 5.15 Examples of Unclear Oil images in the testing set.

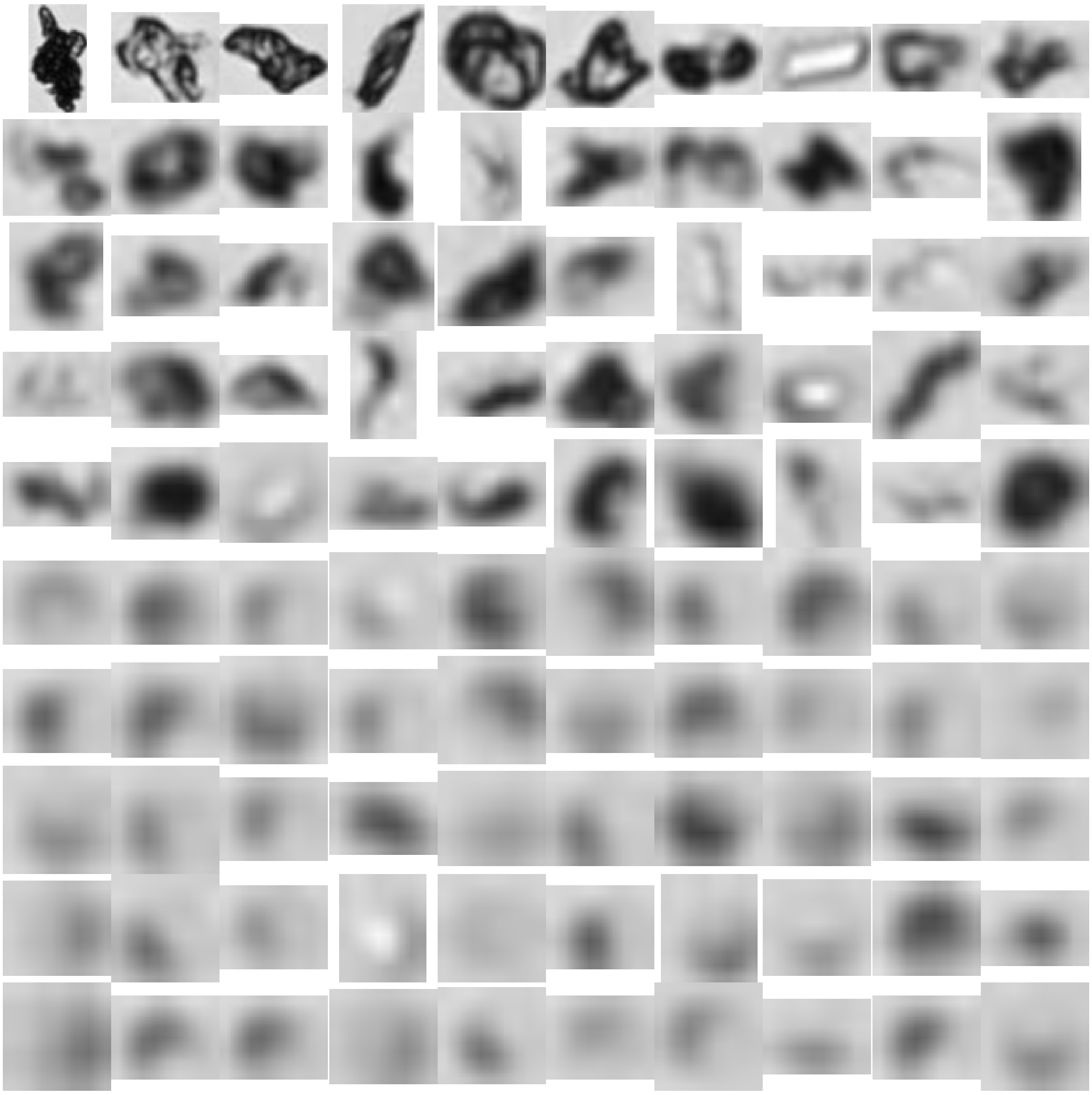


Fig. 5.16 Examples of Clearly NSO images in the testing set.