# Kofax eFlow
## Design Smart User's Guide

Version: 6.1.0

Date: 2023-08-29

KOFAX

# Table of Contents

# Preface

The eFlow Design Smart module enables you to build a model for automatic classification of documents. This guide explains how to use Design Smart in an application, create a classification model, analyze classification result, tune a model, and configure parameters.

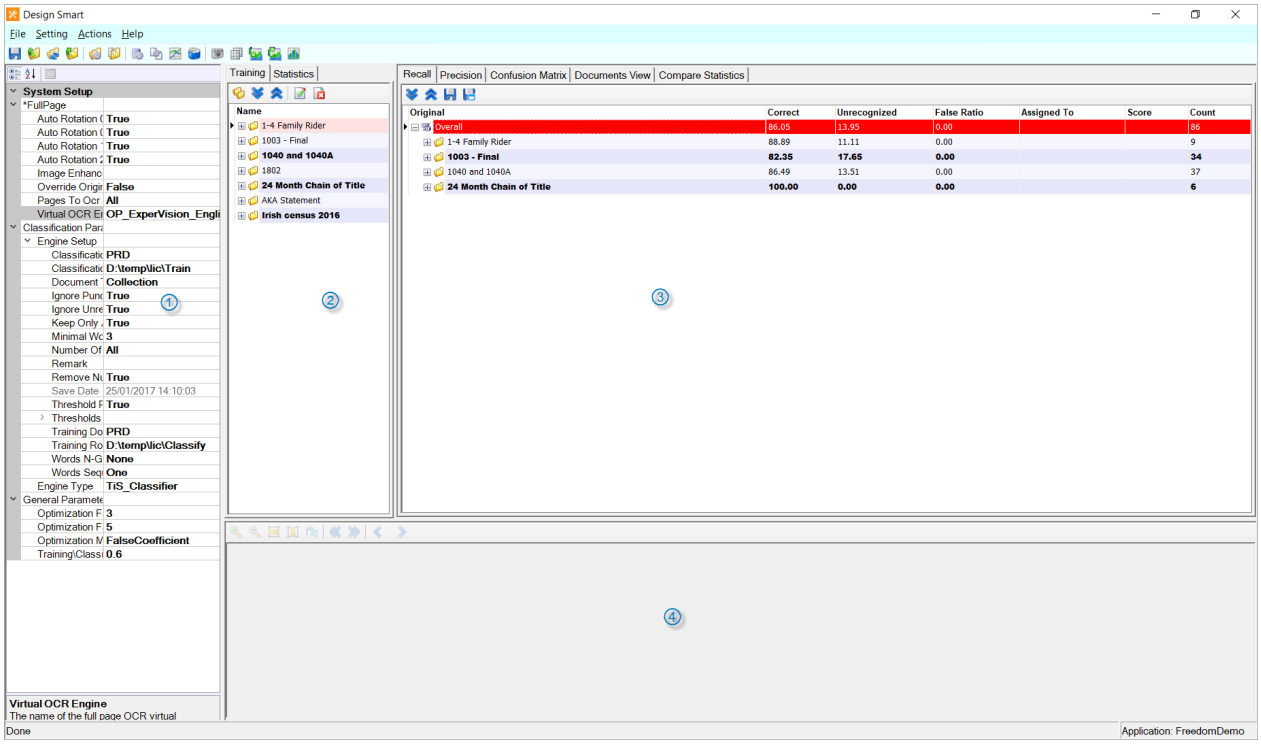
## Product documentation

To access the full Kofax eFlow documentation set online, see the Kofax eFlow Product Documentation page. For a complete set of Kofax eFlow documents, refer to the Kofax eFlow Release Notes.

Chapter 1

# Design Smart window

In the **eFlow LaunchPro**, select **Design Smart** from the **Tools** list. The **Design Smart** window appears.




| | |
|---|---|
| 1 | System setup |
| 2 | Image folders |
| 3 | Statistics |
| 4 | Image viewer |

## System setup

The System Setup pane area contains parameters for configuring and tuning the classification model. See Configuration parameters for more information.

## Images folders

The folders containing the training and test images are displayed in the following two tabs:

- **Training**: Displays all folders and files that have been manually classified to train the system.
- **Statistics**: Displays all folders and files available for testing the automatic classification.

See Create the classification model for more information.

## Statistics

Displays statistics generated after classification.

See Analyze the classification results for more information.

## Image viewer

You can view the image of a document, by double-clicking on the document in the Recall, Precision, or Documents View tab.

The following buttons are available to work with the image viewer.

| Button | Description |
|---|---|
|  | Increases the image size. |
|  | Decreases the image size. |
|  | Displays the full page width. |
|  | Displays the entire page. |
|  | Rotates the image 90 degrees clockwise. |
|   | For multi-page images, goes to the first, last, previous, or next page. |

Chapter 2

# Use Design Smart in eFlow application

To use Design Smart in an eFlow application, first create, test, and tune the Design Smart model for the application.

Add a PageOCR activity and a Smart activity to the Recognize workflow.

The PageOCR activity reads the OCR data of incoming documents. The Smart activity applies the model that was created with Design Smart to the PRD files.

If you want to do OCR on the recognized Smart documents, you will need more activities, such as Integra or Freedom.

The result will be a classification for each batch.

The classification can be kept as meta tags and can be seen in the eFlow Control module.

In the Smart Categorizer, you can see details for the collection, form, class, and confidence.

Chapter 3

# Create a classification model

To create a classification model with Design Smart, you must provide sample TIF image files for each document type or class to be learned and split 60% for training and 40% for testing.

## Guidelines for choosing suitable files

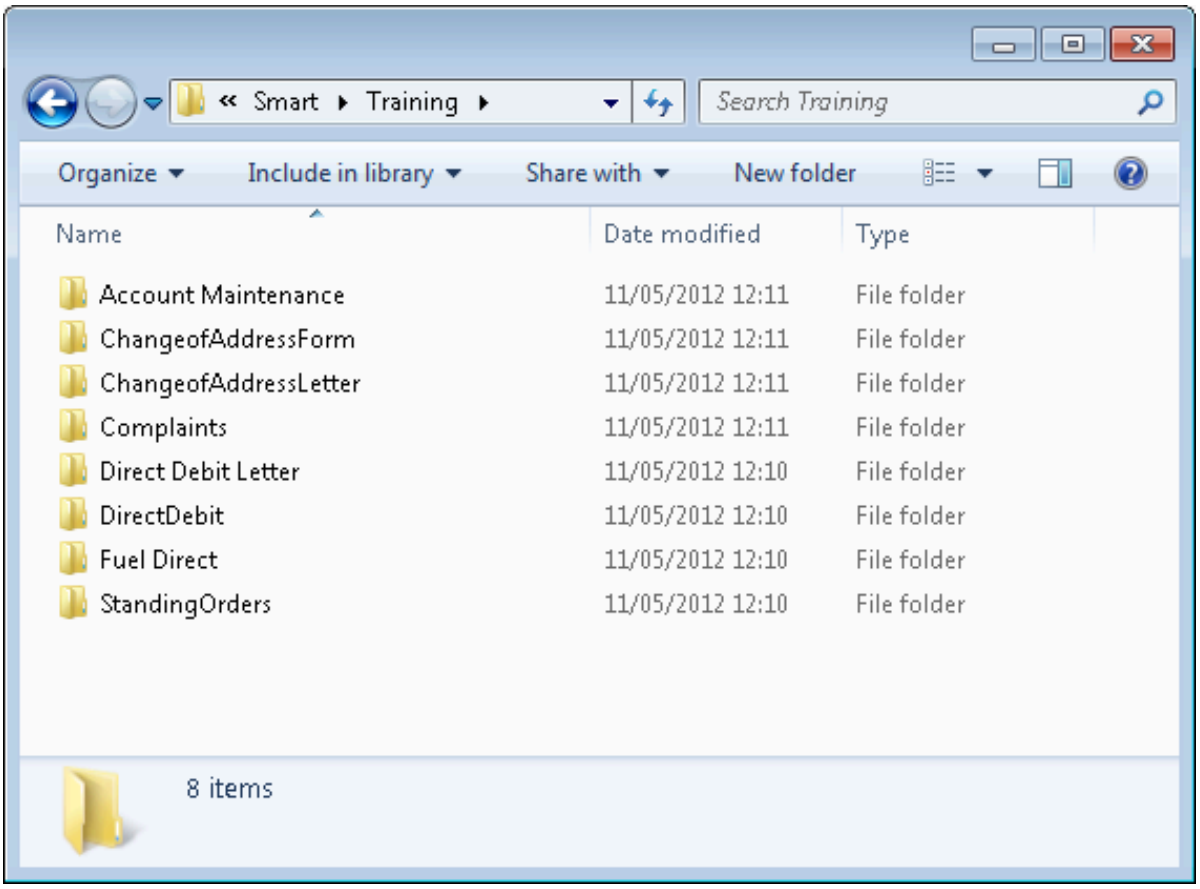
When collecting the samples, follow these guidelines to identify appropriate candidates to be learned:

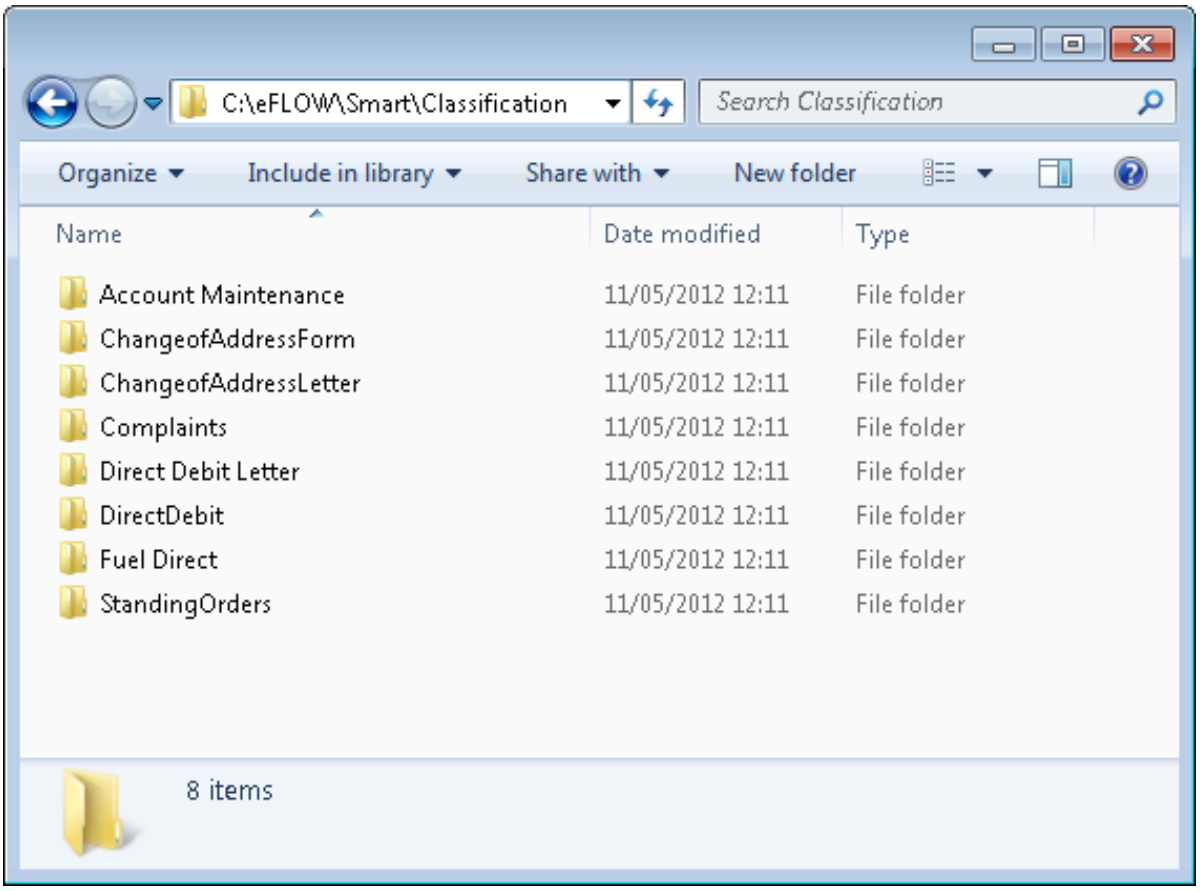
- Do not use handwritten pages unless the handwriting is used within a machine-printed page.
- Rotate images to the correct orientation.
- Do not include blank images.
- Do not include poor-quality images.
- Make sure that images for a single class are not from just one source (from one client or supplier).
- Provide a total of 300 images (documents) per class.
- Do not keep the same image in two different classes.
- Use different images for training and classification.
- Make sure that the relevant folders have full permissions to read, write and create.

## Prepare the files

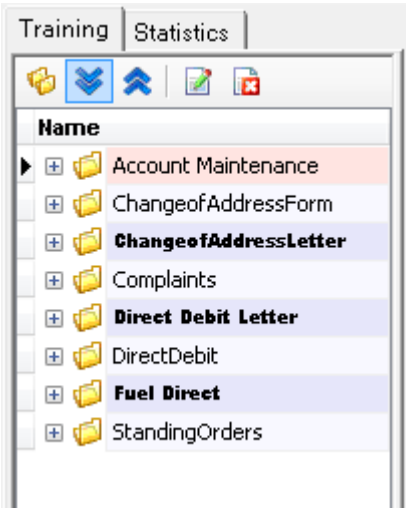
1. Create a folder for storing the images to be learned.
   For example, `C:\eFLOW\Smart`
2. Create two subfolders as shown below:
   - **Training**: This folder contains 60% of the overall files available for each type. These files are used to build the model.
   - **Classification**: This folder contains 40% of the overall files available. These files are used to test the model and display the results in Design Smart.

   **Training** and **Classification** are example folders. You can name the folders as needed.
3. In each folder, create a subfolder for each document type and copy 60% of the TIF image files in the Training subfolders and 40% in the Classification subfolders.


« Smart ▸ Training ▸
Search Training
Organize
Include in library
Share with
New folder
Name
Date modified
Type
Account Maintenance 11/05/2012 12:11 File folder
ChangeofAddressForm 11/05/2012 12:11 File folder
ChangeofAddressLetter 11/05/2012 12:11 File folder
Complaints 11/05/2012 12:11 File folder
Direct Debit Letter 11/05/2012 12:10 File folder
DirectDebit 11/05/2012 12:10 File folder
Fuel Direct 11/05/2012 12:10 File folder
StandingOrders 11/05/2012 12:10 File folder
8 items




4. In the **Design Smart** window, on the **Setting** menu, select **Training Dir**.
5. Select the Training folder, then click **OK**.
   The subfolders are displayed in the **Training** tab.




6. On the **Setting** menu, select **Classification Dir**.

7. Select your Classification folder, then click **OK**.
   The subfolders are displayed on the **Statistics** tab, and the selected folders appear in the **Classification Parameters** under **Classification Root Directory** and **Training Root Directory**.

## Automatic file preparation

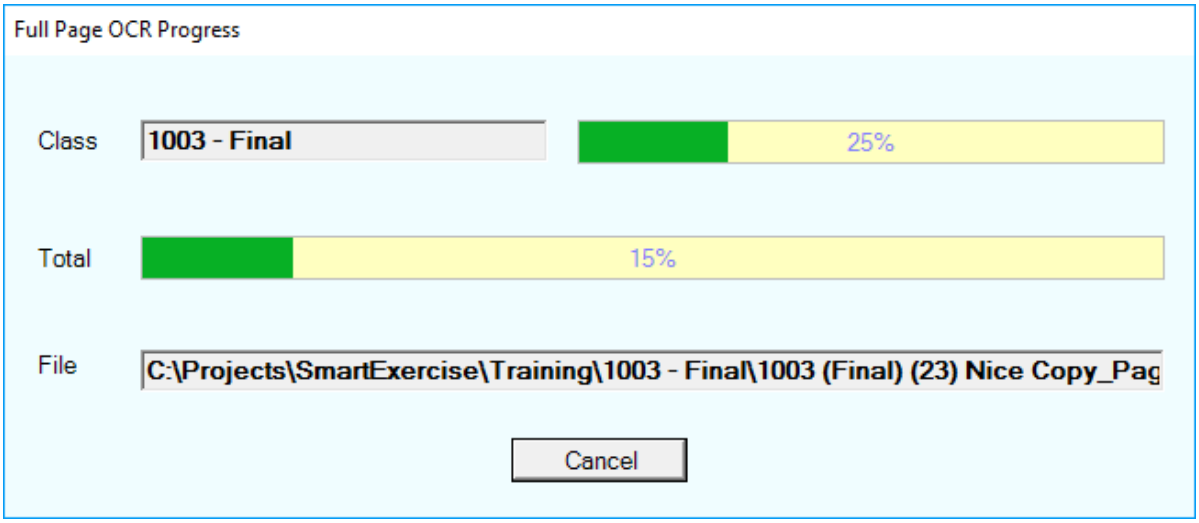
Rather than adding files manually to your Training and Classification folders, you can have Design Smart automatically split the files into these folders.

1. In the **System Setup** pane, under **General Parameters**, type a value in the **Training/ Classification Ratio field**.
   The default value is 0.6, which means that 60% of the available files will be added to the Training folder and the remaining 40% to the Classification folder.
2. On the **Actions** menu, click **Split** .

## Create full-page OCR

You must perform full-page OCR on all images. Performing OCR creates .PRD files in the Training and Classification folders.

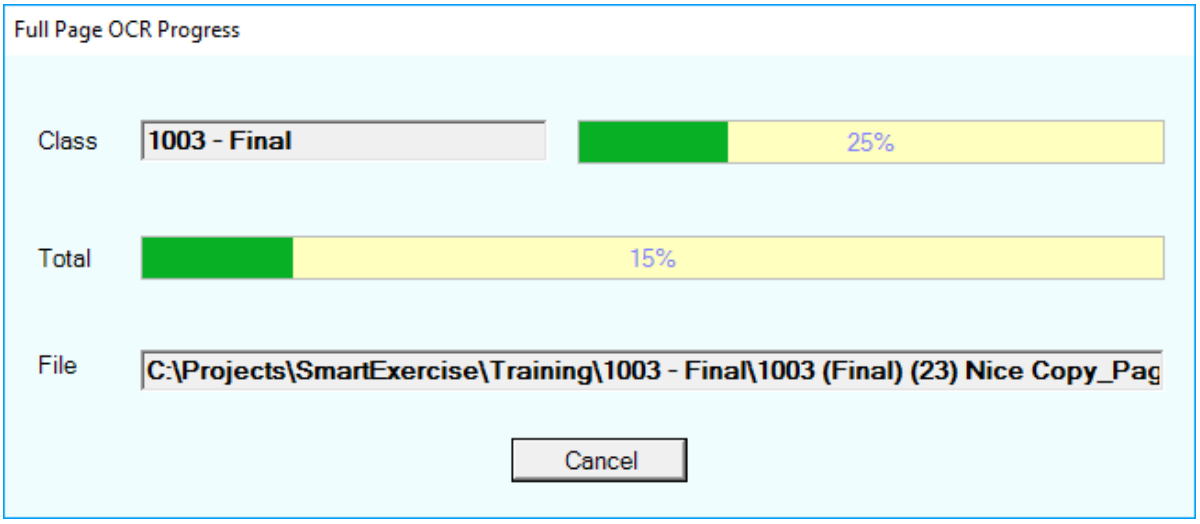
1. In the **System Setup** pane, under the **Full Page** section, select a **Virtual OCR Engine**.
2. On the **Training** tab, click **Create full page OCR** .
   The **Full Page OCR Progress** dialog box displays the progress of the full-page OCR creation process.




   When the process is finished, the dialog box closes with a message, "Full page OCR data was created successfully" appears in the status bar below the **System Setup** pane.
3. On the **Statistics** tab, click **Create full page OCR** .

## Train and classify

1. On the **Actions** menu, select **Train**, or on the toolbar, click **Train** .
   The **Training Progress** dialog box displays the progress of the training process. When the process is finished, the dialog box closes and the message "Done Training" appears on the status bar, below the **System Setup** pane.
2. On the **Statistics** tab, click **Actions** > **Classify**, or on the toolbar, click **Classify** .
   The **Classification Progress** dialog box displays the progress of the classification process. When the process is finished, the dialog box closes and the message "Done" appears on the status bar, below the **System Setup** pane.
   Statistics are now displayed in the **Recall** tab.




## Add new classes and files

You can add new classes and files to the model at any time.

1. In the configuration parameters, set the **Threshold per Category** to **False**.
2. Create new subfolders for the document classes in the Training and Classification folders, and add the appropriate files to these subfolders.
3. Re-select the top-level Training and Classification folders in Design Smart and follow the processes described for building, analyzing, and tuning the model.

## Remove classes and files

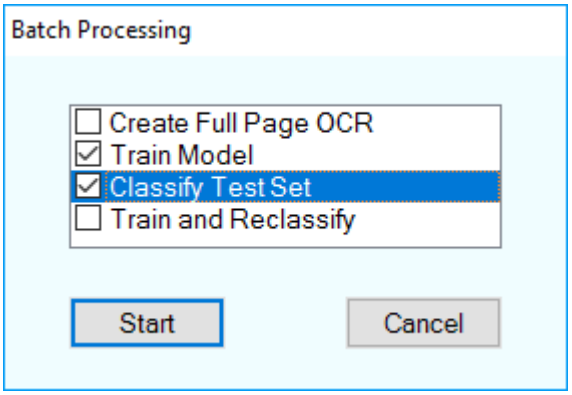
When removing a document type or specific examples from a document class folder, make sure both TIF and PRD files are removed.

There may be more than one PRD file per TIF image.

## Batch processing

In batch processing, you can perform several actions in one go, rather than performing each action manually.

1. On the **Actions** menu, select **Batch Processing**, or on the toolbar, click **Batch Processing** .
2. In the **Batch Processing** dialog box, select the actions you want to perform.




3. Click **Start**.

Chapter 4

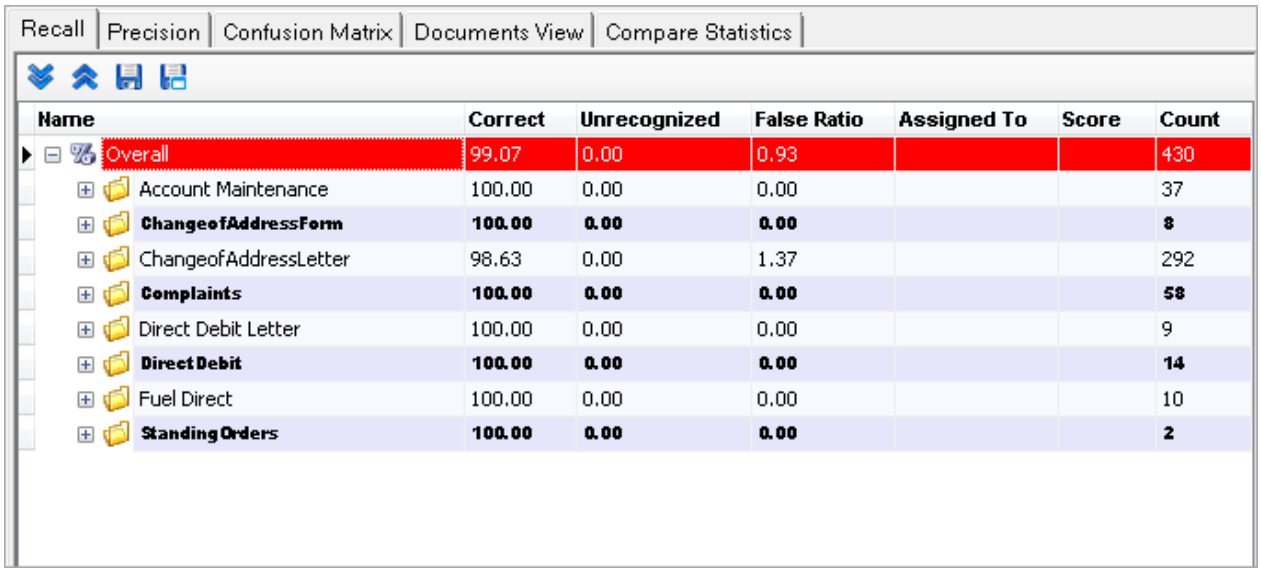
# Analyze the classification results

The results of testing appear in the tabs to the right of the workspace. The results are statistics that measure how accurate the automatic classification is compared to manual classification in the test folder. Your classification of test documents into the class subfolders is assumed to be correct.

## Recall and Precision tabs

The Recall tab lists all tested classes and their documents (that is, the contents of the Classification folder), indicating which of those documents were correctly or incorrectly classified, or not recognized.

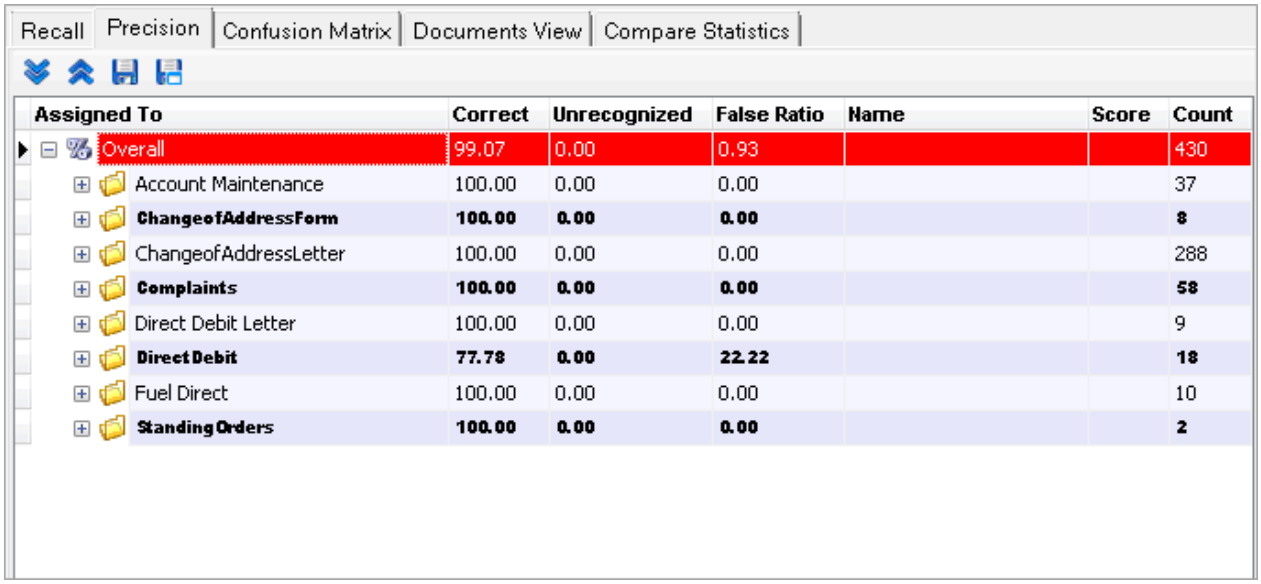
The Recall tab shows the ratio of documents that were correctly classified to a class to the total number of documents that truly belong to this class. The number of documents incorrectly assigned to this class has no bearing on this figure.




| Name | Correct | Unrecognized | False Ratio | Assigned To | Score | Count |
|---|---|---|---|---|---|---|
| Overall | 99.07 | 0.00 | 0.93 | | | 430 |
| Account Maintenance | 100.00 | 0.00 | 0.00 | | | 37 |
| **ChangeofAddressForm** | **100.00** | **0.00** | **0.00** | | | **8** |
| ChangeofAddressLetter | 98.63 | 0.00 | 1.37 | | | 292 |
| **Complaints** | **100.00** | **0.00** | **0.00** | | | **58** |
| Direct Debit Letter | 100.00 | 0.00 | 0.00 | | | 9 |
| **Direct Debit** | **100.00** | **0.00** | **0.00** | | | **14** |
| Fuel Direct | 100.00 | 0.00 | 0.00 | | | 10 |
| **Standing Orders** | **100.00** | **0.00** | **0.00** | | | **2** |

The Precision tab lists all classes to which documents were assigned, indicating which of those documents were correctly or incorrectly classified, or not recognized.

The Precision tab shows the ratio of documents that were correctly classified to a class to the total number of documents that were classified (correctly or in error) to this class. Precision for a class is lower if there are documents that were incorrectly assigned to this class.




| Assigned To | Correct | Unrecognized | False Ratio | Name | Score | Count |
|---|---|---|---|---|---|---|
| Overall | 99.07 | 0.00 | 0.93 | | | 430 |
| Account Maintenance | 100.00 | 0.00 | 0.00 | | | 37 |
| **ChangeofAddressForm** | **100.00** | **0.00** | **0.00** | | | **8** |
| ChangeofAddressLetter | 100.00 | 0.00 | 0.00 | | | 288 |
| **Complaints** | **100.00** | **0.00** | **0.00** | | | **58** |
| Direct Debit Letter | 100.00 | 0.00 | 0.00 | | | 9 |
| **Direct Debit** | **77.78** | **0.00** | **22.22** | | | **18** |
| Fuel Direct | 100.00 | 0.00 | 0.00 | | | 10 |
| **Standing Orders** | **100.00** | **0.00** | **0.00** | | | **2** |

| Column | Description |
|---|---|
| Correct | Percentage of documents correctly assigned to each class.<br>Correctly assigned documents are indicated by . |
| Unrecognized | Percentage of documents that could not be assigned to any class.<br>Unrecognized documents are indicated by . |
| False Ratio | Percentage of documents assigned to the wrong class.<br>Incorrectly assigned documents are indicated by . |
| Assigned To | Class to which the document was incorrectly assigned. |
| Score | The score awarded by Smart Classifier during classification for that particular class. |
| Count | Number of documents overall and in each category.<br>If you set the Document Type parameter to Page, each page of a multipage TIF is treated as a separate document. If you set it to Collection, a multipage TIF is treated as a single document. |

Consider the following example, where a set of 150 documents, 100 of class A, and 50 of class B need to be classified. One set of thresholds returns the following results:

- Of the 100 documents of class A, 80 are classified correctly and 20 are classified as B.
- Of the 50 documents of class B, all 50 are classified as B.

For the A documents:

- The recall for A is 80: It is calculated by the ratio of 80/100 (80%).
- The precision for A is the highest possible: 80/80 (100%), as all 80 documents assumed to be A are actually A.

For the B documents:

1. The recall for B is 50/50 (100%), that is, all 50 B documents were correctly identified.

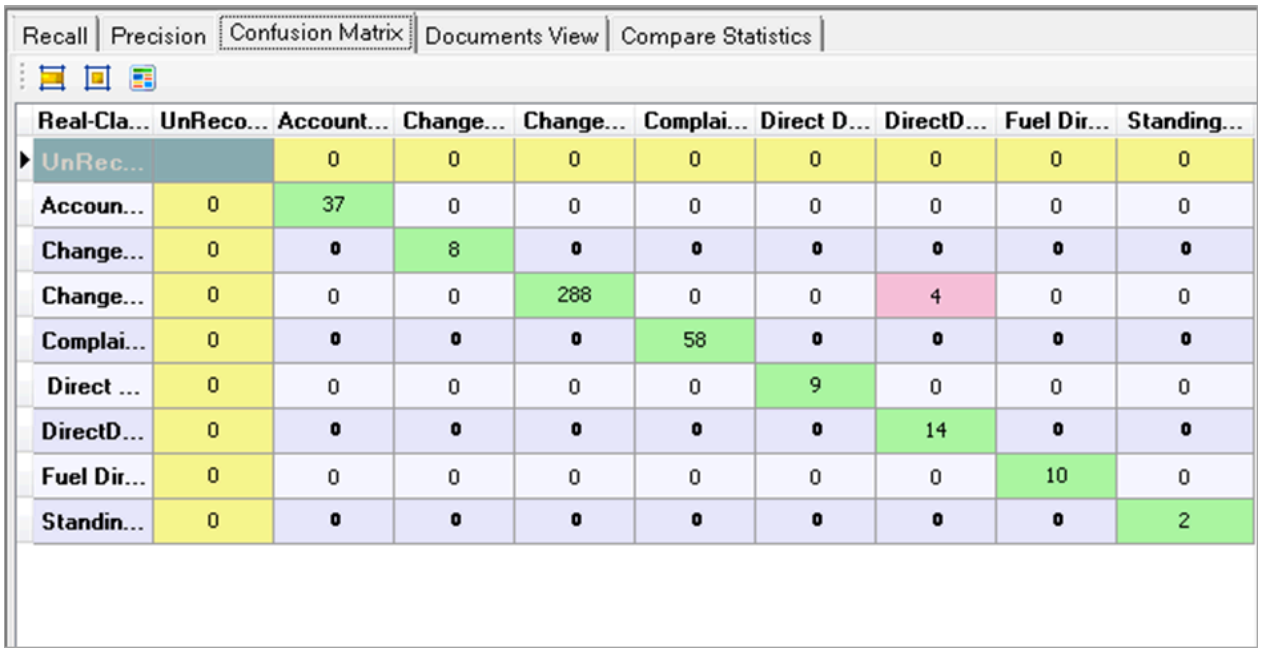
2. The precision for B is not as high, only 50/70 (71.42%), as there were also 20 A documents that were assigned to class B.

Without both recall and precision, it would be necessary to view all the categories to see where the wrong classification was made. For example, if Financial Statements have a low recall rate, you may need to know the other missing categories that are incorrectly assigned. Checking the classes with low precision helps you to focus on those categories.

## Confusion Matrix tab

The essential conclusions of the classification testing are displayed in this tab. Here it is easy to identify patterns in the behavior of the algorithm for a test with a particular set of thresholds and other parameters.

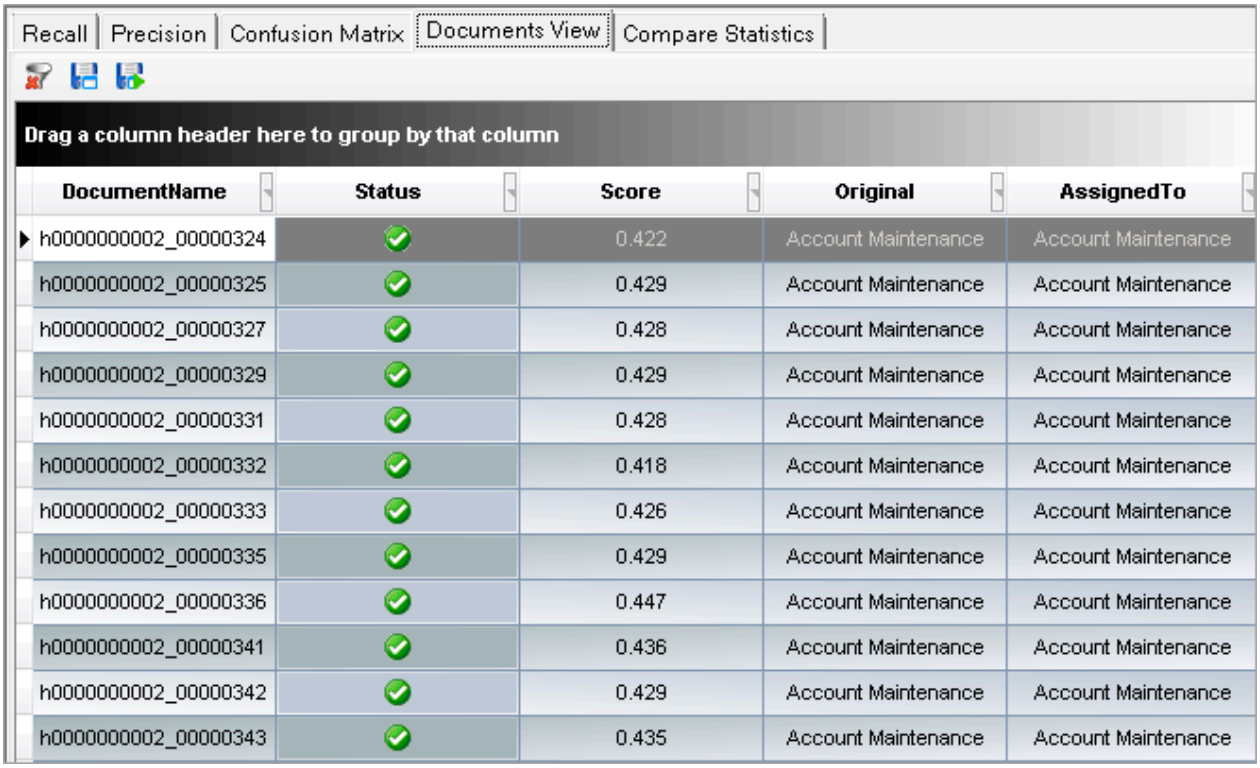
Each cell in the matrix represents a set of documents. Double-click in a cell to view the document set in the Documents View.




| Real-Cla... | UnReco... | Account... | Change... | Change... | Complai... | Direct D... | DirectD... | Fuel Dir... | Standing... |
|---|---|---|---|---|---|---|---|---|---|
| UnRec... | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Accoun... | 0 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Change... | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Change... | 0 | 0 | 0 | 288 | 0 | 0 | 4 | 0 | 0 |
| Complai... | 0 | 0 | 0 | 0 | 58 | 0 | 0 | 0 | 0 |
| Direct ... | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 |
| DirectD... | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 |
| Fuel Dir... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 |
| Standin... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

## Documents View tab

In the Documents View tab, you can view a defined set of documents selected via the Recall, Precision, or Confusion Matrix tabs.



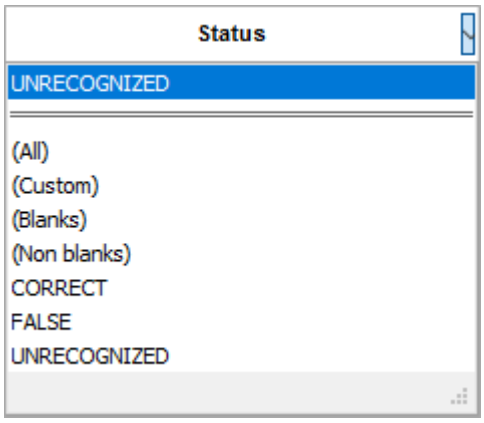

| DocumentName | Status | Score | Original | AssignedTo |
|---|---|---|---|---|
| h0000000002_00000324 | ✓ | 0.422 | Account Maintenance | Account Maintenance |
| h0000000002_00000325 | ✓ | 0.429 | Account Maintenance | Account Maintenance |
| h0000000002_00000327 | ✓ | 0.428 | Account Maintenance | Account Maintenance |
| h0000000002_00000329 | ✓ | 0.429 | Account Maintenance | Account Maintenance |
| h0000000002_00000331 | ✓ | 0.428 | Account Maintenance | Account Maintenance |
| h0000000002_00000332 | ✓ | 0.418 | Account Maintenance | Account Maintenance |
| h0000000002_00000333 | ✓ | 0.426 | Account Maintenance | Account Maintenance |
| h0000000002_00000335 | ✓ | 0.429 | Account Maintenance | Account Maintenance |
| h0000000002_00000336 | ✓ | 0.447 | Account Maintenance | Account Maintenance |
| h0000000002_00000341 | ✓ | 0.436 | Account Maintenance | Account Maintenance |
| h0000000002_00000342 | ✓ | 0.429 | Account Maintenance | Account Maintenance |
| h0000000002_00000343 | ✓ | 0.435 | Account Maintenance | Account Maintenance |

## Sort the list

Click once to sort the list in ascending order, and again to sort in descending order.

## Filter the list

- Click the arrow next to the column heading and select an item from the filter list.
- Click **Reset filter** to view all documents again.



### Group the list

- Click on a column header and drag it to the area labeled as **Drag a column header here to group by that column**.
- To remove a grouping, click on it and drag it back into the list.

### Save the list to a file

You can save the complete document list or a filtered view to an Excel, text, HTML, or XML file.

- To save the file for viewing later, click **Save to file** .
- To save the file and view it immediately in the application associated with that file type, click **Save to file and open with default application** .

### View the file image

Double-click anywhere in the row associated with the document.

## Save and export statistics

When the classification process is complete, you can export the results to an XML file and compare them with up to four different sets of saved classification model statistics.

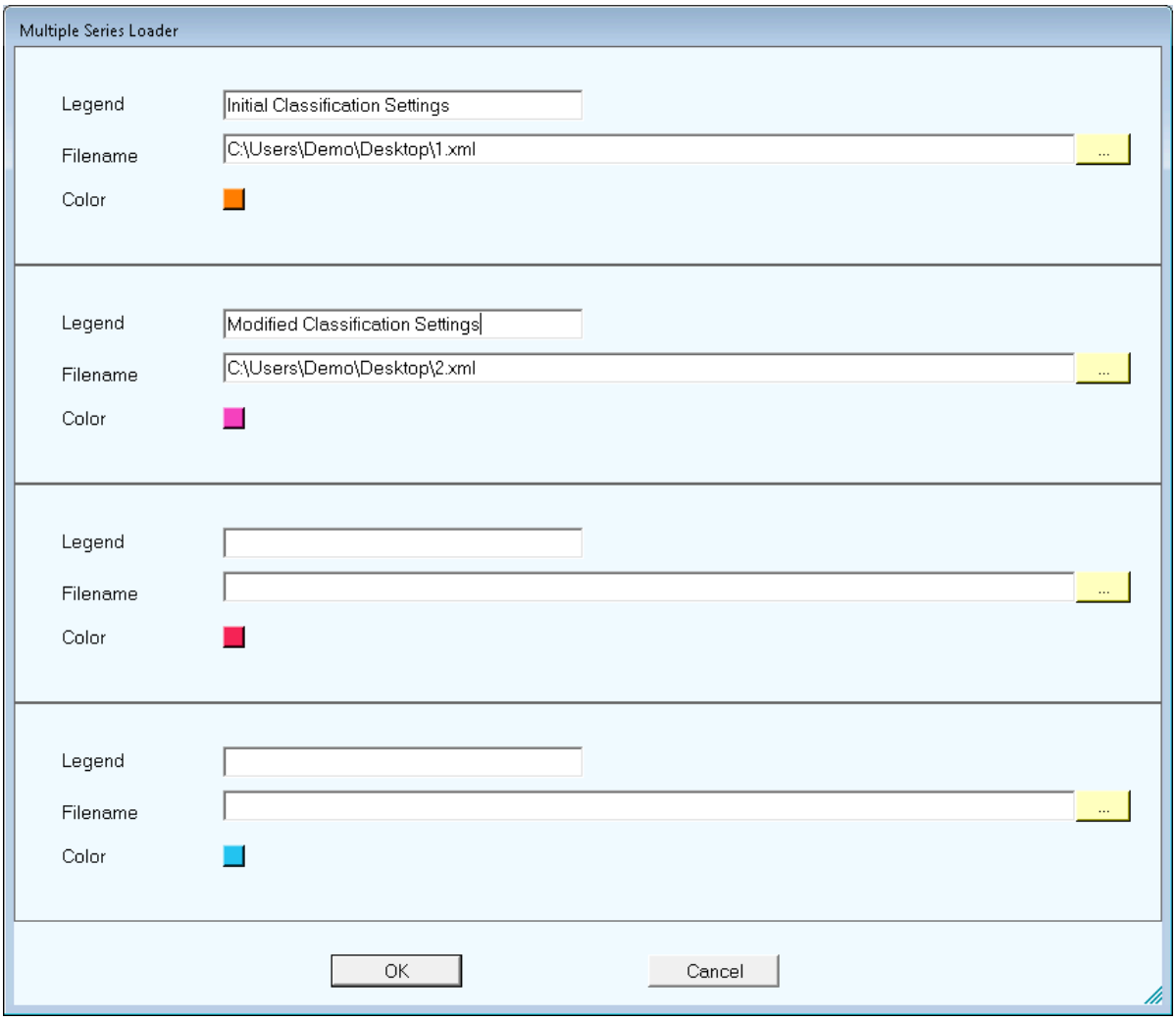
After you update the configuration parameters and the thresholds, you should export both the classification configuration and the classification statistics to a file. Name the files in the same way so that it is easy to locate the correct configuration to re-import following statistics comparison.

- To export the configuration, on the **Actions** menu, select **Export Configuration to file**.
- To export statistics, on the **Actions** menu, select **Export Statistics**, or click **Export Statistics** .

## Compare statistics

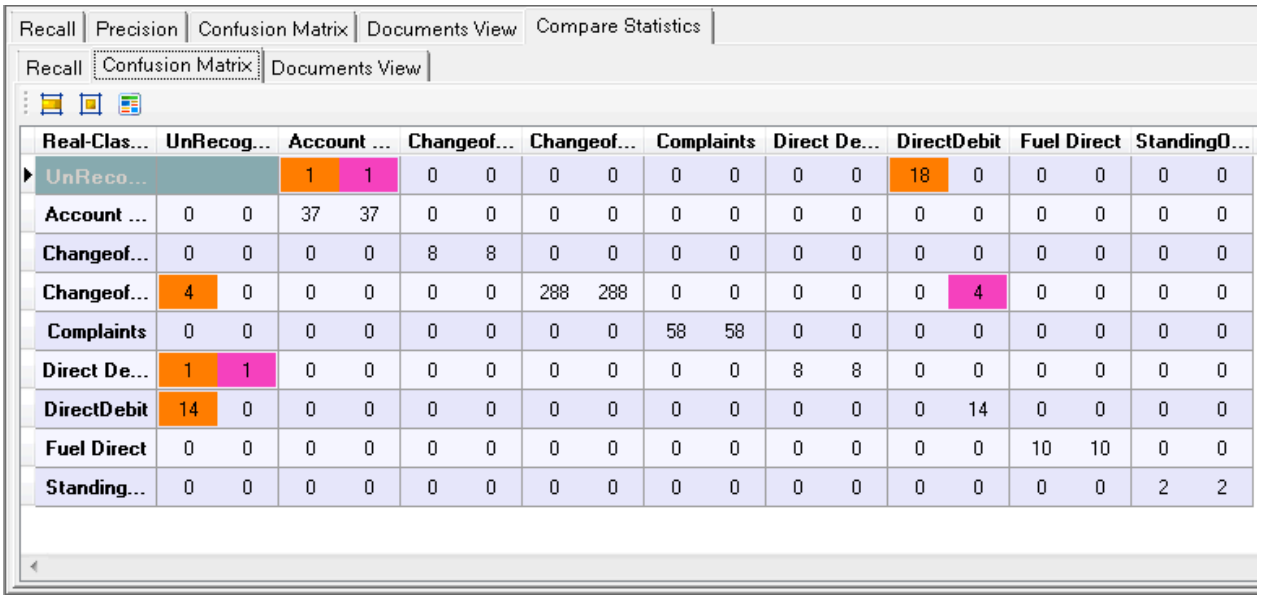
You can compare up to four sets of statistics. You must have previously exported and saved the statistics that you want to compare. See Save and export statistics.

1. On the **Actions** menu, click **Compare Statistics** .
2. In the **Multiple Series Loader** dialog box, click ... right to the **Filename** field and select each statistics file that you want to compare.
3. Optional. In the **Legend** field, type a description of the statistics.
4. Optional. Click **Color** to select a color for each file.




5. Click **OK**.

The following image shows the results of two configurations being compared using the Confusion Matrix view.




| Real-Clas... | UnRecog... | | Account ... | | Changeof... | | Changeof... | | Complaints | | Direct De... | | DirectDebit | | Fuel Direct | | StandingO... | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| UnReco... | | | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 |
| Account ... | 0 | 0 | 37 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Changeof... | 0 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Changeof... | 4 | 0 | 0 | 0 | 0 | 0 | 288 | 288 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 |
| Complaints | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 58 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Direct De... | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| DirectDebit | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 |
| Fuel Direct | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 10 | 0 | 0 |
| Standing... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |

Chapter 5

# Tune the Smart model

Each classified document is given a classification score, which is shown in the Recall, Precision, and Documents View tabs. The desired outcome of classification can be affected by altering the threshold required for positively assigning a document class.

You can control the ratio of unrecognized to false positive using of the Threshold parameter(s), found in the properties grid on the left, under Classification > Engine Setup. For example, if you wish to lower the number of unrecognized documents at the cost of more false positives, lower the threshold. You can specify one general threshold for all categories or a different threshold for each category.

Save the classification model before setting the Threshold per Category value to True.

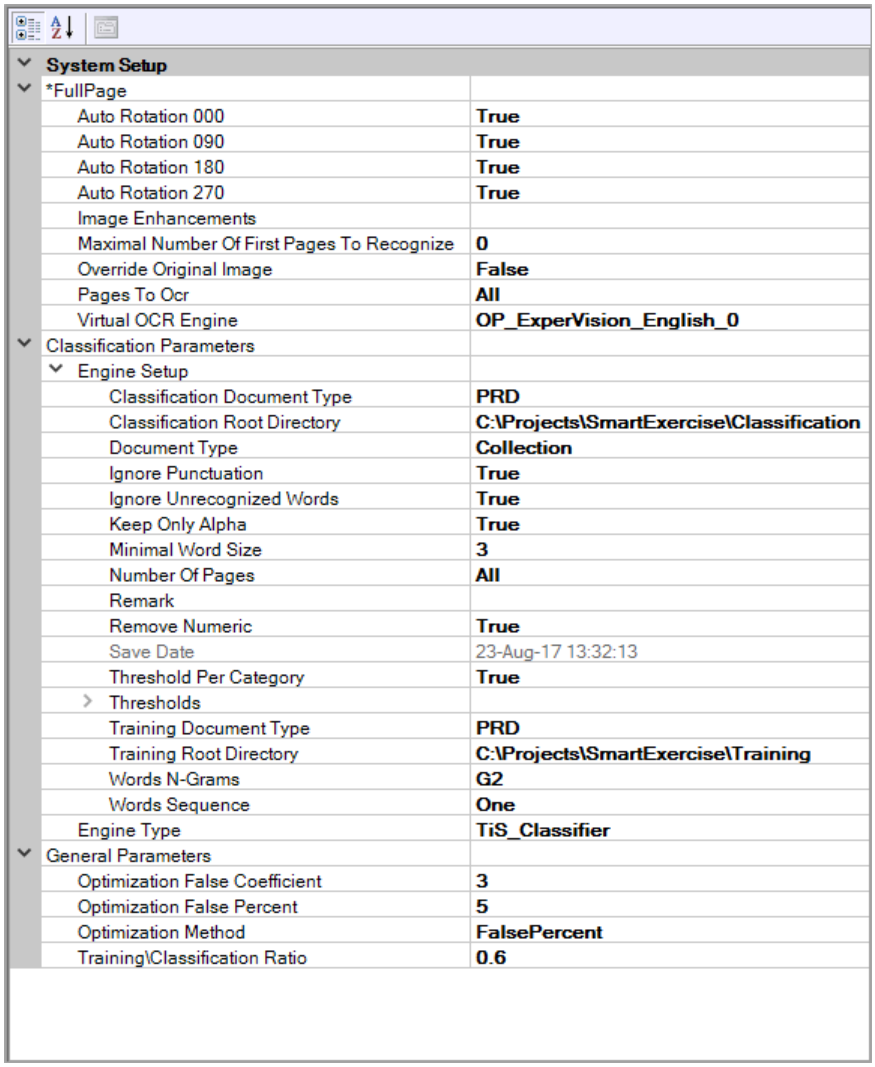
| | |
|---|---|
| ⊟ Thresholds | |
| Accounting Invoices | 0.28 |
| Add Co-Apps | 0.28 |
| Auto Pay | 0.28 |
| Bankruptcies | 0.28 |
| CCCs | 0.28 |
| Death Certificates | 0.28 |
| Disputes | 0.28 |
| Divorces | 0.28 |
| Financial Statements | 0.28 |
| Fixed Pay | 0.28 |
| General Correspondence | 0.28 |
| Remove Co-signer | 0.28 |
| Training Document Type | FLAT_T |
| Training Root Directory | E:\2006 |
| Words N-Grams | None |
| Words Sequence | One |
| Engine Type | TiS_Cla |

| | |
|---|---|
| Threshold Per Category | True |
| ⊟ Thresholds | |
| Accounting Invoices | 0.005 |
| Add Co-Apps | 0 |
| Auto Pay | 0.225 |
| Bankruptcies | 0 |
| CCCs | 0 |
| Death Certificates | 0.275 |
| Disputes | 0.255 |
| Divorces | 0 |
| Financial Statements | 0 |
| Fixed Pay | 0 |
| General Correspondence | 0.29 |
| Remove Co-signer | 0 |
| Training Document Type | FLAT_T |
| Training Root Directory | E:\2006 |
| Words N-Grams | None |
| Words Sequence | One |
| Engine Type | TiS_Cla |

Chapter 6

# Configuration parameters

This chapter describes the Design Smart configuration parameters.



| System Setup | |
|---|---|
| *FullPage | |
| Auto Rotation 000 | **True** |
| Auto Rotation 090 | **True** |
| Auto Rotation 180 | **True** |
| Auto Rotation 270 | **True** |
| Image Enhancements | |
| Maximal Number Of First Pages To Recognize | **0** |
| Override Original Image | **False** |
| Pages To Ocr | **All** |
| Virtual OCR Engine | **OP_ExperVision_English_0** |
| Classification Parameters | |
| Engine Setup | |
| Classification Document Type | **PRD** |
| Classification Root Directory | **C:\Projects\SmartExercise\Classification** |
| Document Type | **Collection** |
| Ignore Punctuation | **True** |
| Ignore Unrecognized Words | **True** |
| Keep Only Alpha | **True** |
| Minimal Word Size | **3** |
| Number Of Pages | **All** |
| Remark | |
| Remove Numeric | **True** |
| Save Date | 23-Aug-17 13:32:13 |
| Threshold Per Category | **True** |
| Thresholds | |
| Training Document Type | **PRD** |
| Training Root Directory | **C:\Projects\SmartExercise\Training** |
| Words N-Grams | **G2** |
| Words Sequence | **One** |
| Engine Type | **TiS_Classifier** |
| General Parameters | |
| Optimization False Coefficient | **3** |
| Optimization False Percent | **5** |
| Optimization Method | **FalsePercent** |
| Training\Classification Ratio | **0.6** |

# FullPage parameters

| Parameter | Description |
|---|---|
| Autorotation | Rotates the image at the set angles. |
| Image Enhancements | Sets image enhancements that are applied on each image before the full page OCR. |
| Maximal Number Of First Pages | The number of pages per document that the full-page OCR will process. |
| Override Original Image | Whether to replace the original image with the auto-rotated TIF. |
| Pages to OCR | Creates full-page OCR for several pages in the document (relevant only to Smart Designer). |
| Virtual OCR engine | The virtual OCR engine is used to perform OCR on the images and create PRD files. You can define the virtual engines to be used with the eFlow application in the eFlow Design module.<br>Refer to the *Kofax eFlow Design User Guide* for more information. |

# Classification parameters

| Parameter | Description |
|---|---|
| Classification Document Type | Type of documents that are be used for classification and by the Categorizer:<br>**PRD**: PRD file<br>**STD**: Text file |
| Classification Root Directory | The folder that contains all classes used for classification. |
| Document Type | **Collection**: Treats each multipage TIF or collection as a separate document.<br>**Page**: Treats each page of a multipage TIF or collection as a separate document. |
| Ignore Punctuation | Removes punctuation signs from the text. |
| Ignore Unrecognized Words | Removes words that contain any unrecognized character from the text. |
| Keep Only Alpha | Keeps only the alphabetical words. |
| Minimal Word Size | Keeps only words whose length is equal to or larger than the specified number of characters. |
| Number Of Pages | Number of pages to consider during the training and classification phases. |
| Remark | General remark that is saved with the model file. |
| Remove Numeric | Removes all instances of numbers. |
| Save Date | Date and time at which the model was last saved. This is entered automatically and cannot be edited. |
| Threshold Per Category | **False**: Applies a global threshold for all document classes.<br>**True**: Allows a different threshold for each class. |

| Parameter | Description |
|---|---|
| Thresholds | Defines individual thresholds for each document class.<br>The Threshold Per Category parameter must be set to True.<br>To define thresholds, in the Thresholds field, click ... See Tune the Smart model for more information on defining thresholds. |
| Training Document Type | Type of documents that will be used for training:<br>**PRD**: PRD file<br>**STD**: Text file |
| Training Root Directory | The folder that contains all classes used for training. |
| Words N-Grams | Breaks words into low-level sub-words.<br>Captures words in long strings and searches for meaningful words of N characters. |
| Words Sequence | Possibly considering a group as words with specific meaning.<br>Important phrases contain a given number of words (if unknown/irregular, leave them as default). |
| Engine Type | Classification engine type. |

## General parameters

| Parameter | Description |
|---|---|
| Optimization False Coefficient | The coefficient of the % false in the linear objective function: % Correct - (coefficient * % False).<br>Used when the Optimization Method is set to FalseCoefficient. |
| Optimization False Percent | The false ratio that the system should reach. Used when the Optimization Method is set to FalsePercent. |
| Optimization Method | Whether to optimize according to the false ratio or false coefficient. |
| Training/Classification Ratio | The ratio of documents that will be copied to the training folder when splitting a folder into training and classification. |