PortiK: A computer vision based solution for real-time automatic solid waste characterization – Application to an aluminium stream

Rémi Cuingnet, Yannik Ladegaillerie, Jérôme Jossent, Aude Maitrot, Julien Chedal-Anglay, Williams Richard, Marine Bernard, Jake Woolfenden, Emmanuel Birot, Damien Chenu

PortiK: a computer vision based solution for real-time automatic solid waste characterization – Application to an aluminium stream

Remi Cuingnet[1], Yannik Ladegaillerie[1], Jérôme Jossent[1], Aude Maitrot[1], Julien Chedal-Anglay[1], Williams Richard[1], Marine Bernard[1], Jake Woolfenden[2], Emmanuel Birot[3], Damien Chenu[1]

[1] Veolia Scientific & Technical Expertise Department, Maisons-Laffitte, France

[2] Veolia UK, London, United Kingdom

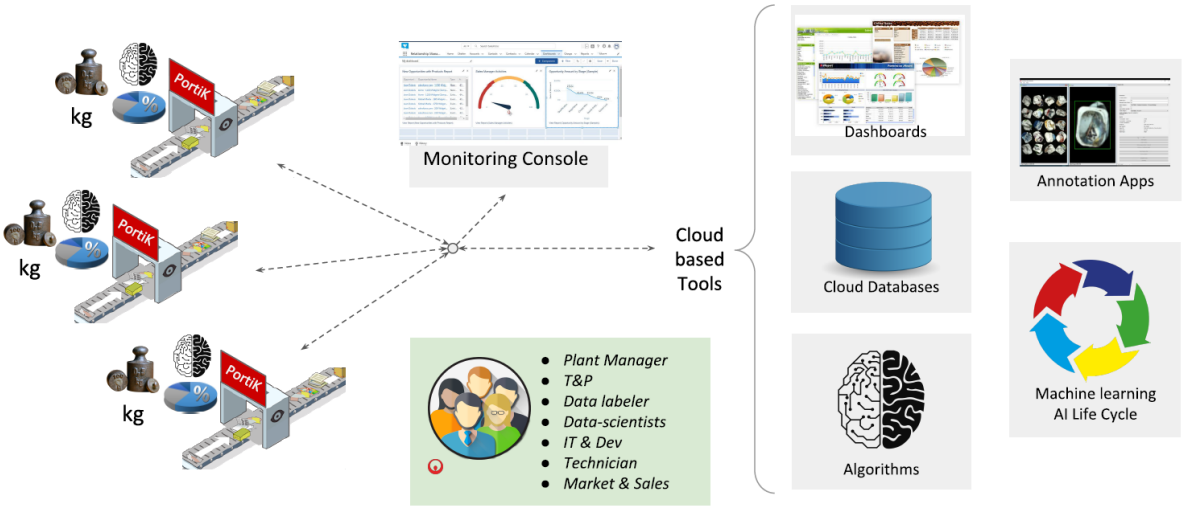
[3] Veolia Recyclage & Valorisation des Déchets, Aubervilliers, France

Abstract

In Material Recovery Facilities (MRFs), recyclable municipal solid waste is turned into a precious commodity. However, effective recycling relies on effective waste sorting, which is still a challenge to sustainable development of our society. To help the operations improve and optimise their process, this paper describes *PortiK*, a solution for automatic waste analysis. Based on image analysis and object recognition, it allows for continuous, real-time, non-intrusive measurements of mass composition of waste streams. The end-to-end solution is detailed with all the steps necessary for the system to operate, from hardware specifications and data collection to supervisory information obtained by deep learning and statistical analysis. The overall system was tested and validated in an operational environment in a material recovery facility.

PortiK monitored an aluminium can stream to estimate its purity. Aluminium cans were detected with 91.2% precision and 90.3% recall, respectively, resulting in an underestimation of the number of cans by less than 1%. Regarding contaminants (i.e. other types of waste), precision and recall were 80.2% and 78.4%, respectively, giving an 2.2% underestimation. Based on five sample analyses where pieces of waste were counted and weighed per batch, these counts were used to estimate purity and its confidence level. The estimation error was calculated to be within ±7% after 5 minutes of monitoring and ±5% after 8 hours. These results have demonstrated the feasibility and the relevance of the proposed solution for online quality control of aluminium can stream.

Keywords: *material recovery facilities; MRF; solid waste characterization; deep-learning; deep neural network; computer vision*



Graphical abstract – PortiK's ecosystem with many different stakeholders. It is composed of two subsystems: an on-site system and a suite of cloud-based services to handle all the back-end operations: data management, image annotations, machine learning and operation services.

Nomenclature

AI	artificial intelligence
cdf	cumulative distribution function
CNN	convolutional neural network
FN	number of false negatives
FP	number of false positives
GPU	graphics processing unit
IoU	intersection over union
MRF	Material recovery facility
TP	number of true positives

1. Introduction

1.1. Need for automatic characterization

In Materials Recovery Facilities (MRFs), solid waste materials collected from the kerbside are separated into different categories by a combination of automatic and manual sorting. These separation steps are generally (1) mechanical screenings, (2) magnetic separation, (3) optical sorting, (4) then manual sorting at the end of the process. The different categories of recycling materials and their expected level of purity are often defined in contractual *Minimum Technical Requirements* with the end buyers and local authorities. Usual categories are plastics, aluminium cans, ferrous metal cans, newspapers, pamphlets and magazines, mixed paper and cardboard.

To be efficient in a competitive, changing, tense environment, the sorting process has to ensure a good ratio between the purity of the sorted material and its processing cost. It should also be flexible to handle new waste and new types of recyclates. Finally, the sorting process and the quality of its outcome should be controlled to meet technical requirements.

Currently, controlling sorted recyclates' purity is achieved (i) quantitatively by manually sampling and sorting waste materials at the end of sorting lines, and (ii) qualitatively with visual inspections of the baled output materials. Hence, regular and frequent waste stream characterizations are impracticable. Besides being intrusive to the sorting process, this task is time consuming and results are obtained after the characterised waste has been baled. In a nutshell, there is no solution to know in real time the waste characterization at any point of the process.

1.2. Artificial Intelligence for waste recognition

In 2012, new machine learning algorithms based on deep learning, together with the arrival of powerful graphics processing units (GPU) led to breakthroughs in artificial intelligence (AI). In particular, outstanding results were obtained by Krizhevsky et al. (2012) in computer vision for the automatic recognition of objects in images with deep convolutional neural networks (LeCun et al. 1998, Goodfellow et al. 2016). They are now even performing as well as humans on specialised visual recognition tasks (Liu et al. 2018, Yang et al 2018).

These great advances in the field of artificial intelligence led different teams to propose new methods based on computer vision to automatically recognize waste. More specifically, standard near-infrared spectrometers measurements were replaced with hyperspectral images (Lachaize et al. 2016; Zheng et al. 2018; Xiao et al. 2019) or even with RGB images to analyse waste (e.g. Adedeji et al. 2019; Srinilta et al. 2019). While solutions based on near-infrared sensors efficiently discriminate between different materials (Zheng et al. 2018), RGB cameras have been increasingly used. Their low costs and easy setups make them valuable alternatives, leading to some automation for waste collection (Donati et al. 2020; Melinte et al. 2020) and waste sorting (Gundupalli et al. 2017; Strollo et al. 2020) or even to new solutions such as the smart bins that control the type of waste brought in (White et al. 2020; Chen et al. 2020) or automatically sort the pieces of waste into different compartments (Sheng et al. 2020).

Most solutions based on RGB images considered waste analysis as a supervised image classification problem. It consists in automatically assigning a category (among a predefined set) to a given image (Krizhevsky et al. 2012). The classification problem is named supervised when the assignment function is selected and calibrated on a set containing images together with their expected categories. To classify waste images, Chu et al. (2018), Adedeji et al. (2019) and Toğaçar et al. (2020) used pre-trained neural networks such as the AlexNet (Krizhevsky et al. 2012), GoogLeNet (Szegedy et al. 2015) or ResNet-50 (He et al. 2016) to extract features followed by classifiers such a multi-layer perceptron or support vector machine. Unlike previous approaches, White et al. (2020) and Mao et al. (2021) used end-to-end convolutional neural networks. Since many neural network architectures for the image classification problem exist in the litterature (Dhillon and Verma 2020), some authors compared the performances of some of them to classify waste images. More specifically, Aral et al. (2018) compared the Dense-Net (Huang et al. 2017), Res-Net, Mobile-Net (Howard et al. 2017; Sandler et al. 2018), and the Xception (Chollet 2017) architectures for the classification of images from Trashnet dataset (Yang and Thung 2016). They obtained the best accuracies (95% and 94%) with a Dense-Net and a ResNet respectively. Srinilta et al. (2019) compared the performances of different neural network architectures for the waste classification task of 9,200 municipal solid waste images. Among the VGG-16 (Simonyan and

Zisserman 2014), ResNet-50, MobileNet V2 and DenseNet-121, the best performances were obtained with the ResNet-50 architecture with 95% accuracy. Finally, on a different dataset, Thanawala et al. (2020) obtained the best performances (93% accuracy) with a Mobile-Net architecture compared to with a VGG-16, Dense-Net or Google-Net architecture to classify waste images.

While their results are very promising, classifying the whole image often leads to strong constraints on image acquisition. This limitation has been overcome using object recognition algorithms (Ren et al. 2015). They automatically identify both the categories and the positions of the different pieces of wastes in images. To monitor the waste brought into a bin in real time Chen et al. (2020) detected the motion of falling waste using an *Adaptive Gaussian mixture model*. Once detected, the items were classified using a hybrid model based on MobileNetV2 (Sandler et al. 2018) and on the action recognition network R3D (Tran et al. 2018). The detection and classification can be performed jointly with a dedicated neural network. Sheng et al. (2020) jointly detected and classified items in the bin with an SSD neural network (Liu et al. 2016). The SSD neural network was also used to detect and identify waste on the ground (Melinte et al. 2020). De Carolis et al. (2020) used YOLO (Redmon et al. 2016), another object recognition neural network to identify waste in public city areas or in places outside of town. Unlike the previous approaches, Donati et al. (2020) considered the waste recognition problem as a semantic classification problem: it consists in assigning a (waste) category to each pixel in an image.

Although these new techniques will yield major innovations in the waste management sector, some limitations still need to be overcome (Abdallah et al. 2020). First, insufficient field data is often a major obstacle affecting the implementation and evaluation of AI systems. To the best of our knowledge, none of the mentioned approaches has used actual data from MRF. According to Abdallah et al. (2020), another limitation is that most studies were direct applications of AI models and more efforts should be made to adapt these technologies to SWM.

1.3. Paper outline

This paper describes *PortiK*, a solution for automatic waste analysis on conveyors. It automatically computes stream quality evolution based on image analysis. The contribution of this paper is threefold. First, an end-to-end solution is described with all the steps necessary for the system to operate, from hardware specifications and data collection to supervisory information. Then, the overall system was validated in an operational environment. *Portik* monitored an aluminium stream in a material recovery facility. To the best of our knowledge, no published solution has been evaluated on real data from material recovery facilities. Finally, *PortiK* computes some monitoring indicators dedicated to waste management such as mass purity.

The overall system is described in section 2.1. Conveyor images acquired in a material recovery facility are analysed in real time with an object detection algorithm (section 2.2). The detection results are then transformed into different performance indicators specific to waste sorting facilities (section 2.3). To be operative, the detection algorithm has to be calibrated on a specific dataset (section 2.4). Finally, the proposed system has been evaluated in an operational environment in a MRF (section 3). Results are discussed in section 4.

2. Methods

2.1. Overall system description

PortiK has been designed as a solution for continuous, real-time, non-intrusive measurement of the composition of waste streams. It relies on image based waste recognition by artificial intelligence. It is composed of two subsystems: an on-site system and a suite of cloud-based services.

PortiK on-site system (figure 1) is composed of a vision system on a mechanical structure, a monitoring console and a communicating computing unit. The vision system consists of a 2464x2056 resolution colour matrix camera (JAI GO 5100C-USB) and LED lighting ramps above the conveyor. For a good trade-off between the amount of data to analyse and the image quality, the proposed solution is based on image analysis rather than video analysis. The system has been set up with a 20 mm depth of field and a 1Hz image acquisition frequency to analyse a conveyor running at 0.4 m/s with a 50% overlap between the successive images. Once acquired, the raw images are analysed in real time with an industrial computer (Intel Core I7-7700 CPU, NVIDIA Geforce GTX 1080 GPU and 32 Go RAM) (sections 2.2 and 2.3). The results are displayed on the local interface (figure 2) and sent to the remote cloud system.

For the on-site system to operate, the overall solution relies on cloud-based services to handle all the back-end operations: data management, image annotations (section 2.4), machine learning pipelines (section 2.2) and operation services including model deployment and system performance monitoring (section 2.2.4).

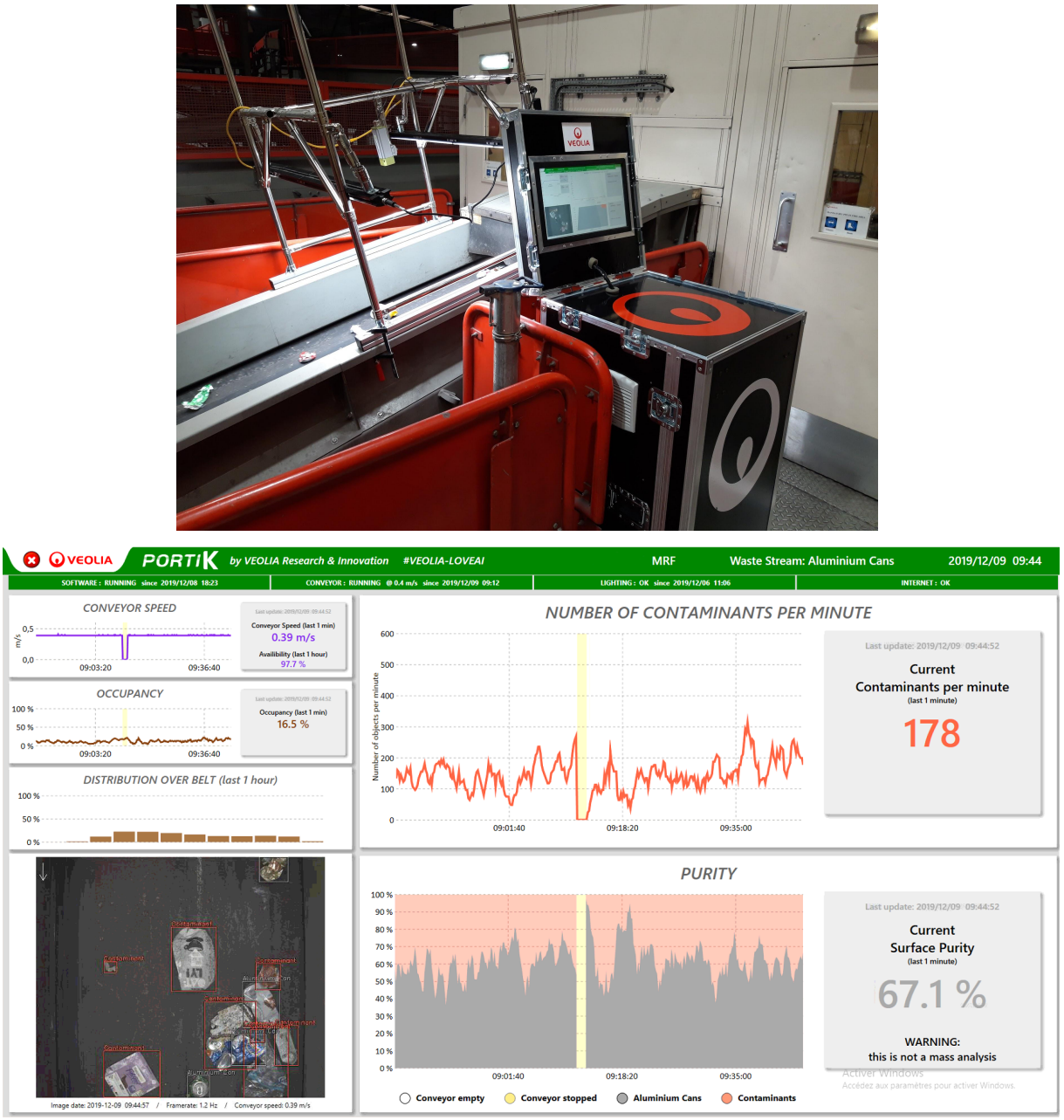


Figure 1 *– PortiK prototype's local system and its local on-site interface as installed in a material recovery facility (London area) for purity analysis of the aluminium can stream at the entrance to the picking room. The interface showed the results of the object recognition algorithm together with dedicated indicators: conveyor speed, the occupancy rate, the distribution over belt, the object rate and the stream purity.*

2.2. Object recognition

2.2.1. Waste characterization as an object recognition problem

The solution proposed in this paper aims at measuring different indicators based on image analysis (section 2.3): the mass flow rate, the stream purity and the occupancy rate. Directly regressing these indicators on images raises several limitations. In particular, building the required training and test datasets would be impossible in practice. Hence, the problem was divided into two steps. First, the different pieces of waste and their categories are detected. Then, based on these detections, the desired indicators are computed. Not only does this approach make the construction of training and test datasets much easier but it also lessens the black box nature of such an AI model (Abdallah et al. 2020). Having quantitative and qualitative (visual) intermediate results smoothens both the development of such a solution and its adoption by end-users.

In the computer vision literature, the problem of detecting all occurrences and locations of some objects of interest and automatically assigning them a category is either called *object recognition* or *instance segmentation* depending on location accuracy. When the position in the image is coarsely defined, usually as axis-aligned bounding boxes surrounding the objects' instances, this is called *object recognition* (*eg.* Ren et al. 2015). On the other hand, for the fine delineation of objects, the term *instance segmentation* (e.g. Silberman et al. 2014; He et al. 2017) is used instead. However, precise segmentation comes at the expense of greater computational resources, both for training and for inference. It also requires fine object delineation in the training and test sets, which is time consuming. Object recognition therefore seems a good trade-off. For object recognition to be used in waste characterization, the waste stream has to be assumed countable. It should be composed of countable and separately identifiable pieces of waste as in figure 2. Note that such hypotheses do not hold for very dense fibre streams.

2.2.2. Object recognition model

Modern object recognition detectors address this problem in two steps. First, a features extractor convolutional neural network (CNN) sometimes called *backbone architecture* takes image as input and makes pertinent information available through a features map. Some usual features extractors are the Resnet (He et al. 2016), the VGGNet (Simonyan and Zisserman 2014), the GoogLeNet (Szegedy et al. 2015), the AlexNet (Krizhevsky et al.2012), the DenseNet (Huang et al. 2017) and the MobileNet (Howard et al. 2017).

Figure 2 *– Results of the object recognition on two images from the aluminium stream in an MRF. Grey boxes indicate the detections of aluminium cans and red boxes the detections of contaminants.*

Second, surrogate regression and classification problems are solved to detect, locate and assign a category for every object of interest. More specifically, based on the extracted feature map and a large set of position proposals, a classification block detects, for each proposal, whether there is an object or not at this location and assigns a category to it. A regression block adjusts this proposed position to predict the location. Depending on the methods, the position proposals are encoded as boxes (Ren et al. 2015; Cai et al. 2029), anchors position (Liu et al. 2016; Lin et al. 2017), pixel position (Redmon et al. 2016), or window centres (Tian et al. 2019). Recently, Carion et al. 2020 proposed a different approach based on the transformer architecture (Vaswani et al. 2017) to avoid the surrogate classification and regression of position proposals.

Choosing an object detection model comes down to choosing a feature extractor CNN and a detection method. Srinilta et al. (2019) compared feature extractors for the waste classification task of 9,200 municipal solid waste images. Among the VGG-16 (Simonyan and Zisserman 2014), ResNet-50 (He et al. 2016), MobileNet V2 (Sandler et al. 2018) and DenseNet-121 (Huang et al. 2017), the best performances were obtained with the ResNet-50 architecture. Nevertheless, in machine learning, conclusions drawn from how a model behaves on a use case can difficulty predict precisely its behaviour on other use cases. In this study, the chosen criteria to pick a model were *(i)* an extensive use in the literature and state of the art performances on databases such as PASCAL VOC challenges (Everingham et al. 2010) an*d (ii)* an out-of-the-shelf implementations in standard machine learning libraries (tensorflow and pytorch) with available weights from different training on standard

databases. Furthermore, to be used in real time, the inference should take less than one second per image with a standard GPU card (e.g. NVidia GTX1080 or Quadro P5000). As a matter of fact, the selected detection model was a Faster R-CNN model (Ren et al. 2017) with a ResNet-101 from the ResNet family (He et al. 2016) as feature extractor.

2.2.3. Model training

Similarly to existing deep learning object recognition approaches detailed in section 1.2, our image analysis algorithm is based on supervised learning (section 2.2). To be operative, it has to be calibrated with annotated data. This *training* step set the model's numerous parameters for the detection and classification to get as close as possible to the expected results. To that end, a few hundred images were acquired with the local PortiK and manually annotated as described in section 2.4. On each image bounding boxes were manually drawn around each piece of waste. The pieces of waste were then assigned categories.

Most neural networks for image recognition can be described as a stack of simple parameterized modules that roughly mimics the human vision system (Goodfellow et al. 2016). Their main drawbacks are the large number of required data to properly calibrate the numerous modules' parameters. Since operations computed in the first layers are low level image processing, one may assume that these operations could be shared between different recognition problems. This is the rationale behind most transfer learning strategies. Transfer learning focuses on how to use knowledge gained by solving a different but related problem (Yosinski et al., 2014,, Oquab et al. 2014). Oquab et al. (2014) showed that the first layers of a convolutional neural network trained on large-scale annotated datasets can be efficiently used for different visual recognition tasks with limited amount of training data. Yosinski et al. (2014) further studied the transferability of the different layers of a convolutional neural network. They reached the conclusion that rather than directly transferring convolutional layers, which might cause some issues, initialising a network with transferred features from almost any number of layers and *fine-tuning* them can produce a boost to performance generalisation.

In this study, transfer learning was then performed by *fine-tuning* a model pre-trained on MSCOCO Dataset (Lin et al. 2014)[1]. Our model was trained with the default hyper-parameters using data augmentation (Krizhevsky et al. 2012).

When training a model, there is a risk that the optimised parameters values are too specific to a particular training set of data and thus fail to predict future observations reliably. This problem is *called overfitting*. Hence, to ensure that decision rules have been inferred properly, it is necessary to evaluate trained models on an independent set called the *test set*.

[1] https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md

2.2.4. Model evaluation metrics

For an object recognition algorithm, both the accuracy of the position and of the classification is to be evaluated. In practice, this evaluation is transformed into the evaluations of binary detections by *(i)* considering the different categories independently and *(ii)* by binarizing the position estimation evaluations as either correct or incorrect. The latter is done by computing matching scores between detections and true positions. The scores are then thresholded, thus classifying the detections either as correct or incorrect. For bounding box detection, the intersection over union (IoU) thresholded at 50% is often used as it corresponds to a good match (Everingham et al. 2010).

The metrics used for binary detections are the *precision* and *recall*. With *TP* and *FP* the number of true positive and false positive samples, the *precision* is defined by:

$$precision := \frac{TP}{TP + FP} . \qquad (1)$$

It quantifies how many selected items are relevant. With *FN* the number of false negative samples, the *recall* is then defined by:

$$recall := \frac{TP}{TP + FN} . \qquad (2)$$

It answers the question: how many relevant items were detected? A standard way to combine *precision* and *recall* is the *average precision, AP*. It is defined as the area under the precision-recall curve. In this report, unlike Everingham et al. (2010), the raw precision-recall curve was used**.**

As aforementioned, the mass purity estimation is based on object recognition. As a matter of fact, only the numbers of detected objects per category are used to estimate the masses and the mass proportions (section 2.3). Thus, while precision and recall are efficient at assessing the model performance, they are too strict. Error in counting only comes from the difference between the number of false negatives and the number of false positives. In other words, they compensate for each-other. The counting error is given by:

$$Counting\ Error := \frac{FP - FN}{TP + FN} = \frac{recall}{precision} - 1 \qquad (3)$$

2.3. From object recognition to waste sorting indicators

2.3.1. Object Rate and conveyor occupancy

The first indicator estimated from the object recognition is the object rate per waste category. While this indicator may seem straightforward to compute, there are some technical difficulties to take into account. First, to deal with image overlaps, the object rate is

computed per image and then averaged. Moreover, counting all the objects in the images would result in an overestimation of the object rate. Indeed truncated objects would be counted twice on average. A simple but robust way to unbias the estimation is to weight truncated objects by 0.5.

The object recognition results make it also possible to estimate the conveyor occupancy and the waste distribution over the conveyor width. These indicators are keys to monitor that the sorting machines operate in nominal conditions. It may also help prevent jamming and detect major dysfunctioning.

2.3.2. From vision to mass

So far, waste flows are estimated in terms of number of pieces of waste. To take another step forward, an easy to implement method is proposed to convert detection information into mass information. The goal is to *(i)* estimate the mass flow rate and the mass purity rate of solid waste, and to *(ii)* associate estimations with confidence intervals for reliability.

The only approach found in the literature that proposes to estimate the mass of an object from an image is by Standley et al. (2017). In a nutshell, the mass is directly regressed on the image. Although such an approach might be very promising, it would require a large database of images of pieces of waste together with their corresponding weights. This is too important a limitation for it to be used in practice in material recovery facilities right now.

2.3.2.1. Rationale and hypotheses

The proposed process is both simple and efficient with only a slight additional cost in terms of data acquisition. For each category of interest, only the number of objects is assumed to be known. To be consistent with this hypothesis, the waste stream must be composed of countable and separately identifiable pieces of waste. Such information could result from an object recognition algorithm (section 2.2.2) or an image annotation task (section 2.4.4). The mass is then regressed on these numbers. Least square regression may seem a natural choice. However the underlying homoscedasticity assumption is not met. Therefore, an alternative based on mass statistical distribution analysis is used instead. It comes down computing a confidence interval on the mass knowing some information about its distribution. The number of pieces of waste per category being very large, their mass distributions are assumed to be normally distributed. Finally, categories are assumed to be independent.

2.3.2.2. Calibration

The first step consists in obtaining, for each waste category, information about its mass distribution: mean and standard deviation. As weighing each piece of waste individually is too cumbersome a process to be accepted by end-users in material recovery facilities,

objects were weighed by batches . The input data for calibration is S sample analyses (at least five) with, for each sample *s*, *(i)* the number n_s of pieces of waste and *(ii)* the total mass w_s per category. Categories are treated independently. To take into account the uncertainty about the estimations, a Bayesian approach detailed in (Cuingnet 2021) is followed. As a result, for a single category, given the sample analyses, with a Jeffrey's noninformative prior, the posterior predictive of the mass *W* of *n* (new) objects follows a non-standardized Student *t*-distribution with *S* degrees of freedom as defined in equation (4).

$$W \mid (w_s,\ n_s)_{s=1,\ldots,S} \sim t_S\left(n \cdot \mu\ ;\ n\,\frac{\beta_s}{S/2}(1\ +\ n/N)\right) \tag{4}$$

where

$$N := \sum_{s=1}^{S} n_S \tag{5}$$

$$\mu := \frac{1}{N}\sum_{s=1}^{S} w_s \tag{6}$$

$$\beta_S :=\ \frac{1}{2}\sum_{s=1}^{S}\left(\frac{w_s - n_s \cdot \mu}{\sqrt{n_s}}\right)^2 \tag{7}$$

Note that $(n \cdot \mu)$ is the mean and $\left(n\,\frac{\beta_s}{S/2}(1\ +\ n/N)\right)$ is the square of the scale parameter.

2.3.2.3. Inference

The Bayesian approach allows for the computation of the posterior predictive distribution, the distribution of possible unobserved values conditional on calibration (equation (4)). Thus, predicting the mass from the number of pieces of waste (per category) comes down to estimating the quantile of the posterior predictive distribution. The 2.5th and the 97.5th percentiles give the 95% confidence interval. As for the stream composition or purity, it requires computing the quantile of the ratio distribution W_c / W_A between the mass of contaminants W_C and the mass of aluminium cans W_A. Estimating the quantile of ratio distribution can be carried out by inverting the cumulative distribution function (*cdf*) $F_{W_C / W_A}(r)$ with numerical inversion such as the secant method. Similarly to (Curtiss 1941) and (Hinkley 1969), W_A being mostly positive, $F_{W_C / W_A}(r)$ can be approximated by $F_{W_C - rW_A}(0)$, the cumulative distribution function at 0 of the linear combination $W_C - r.W_A$. Therefore, the *cdf* was computed with the Gil-Pealez (1951) inversion formula (Witkovsky 2016) as formulated in equation (8):

$$F_{W_C / W_A}(r) \approx F_{W_C - rW_A}(0) = \frac{1}{2} - \frac{1}{\pi} \int_0^{\infty} \frac{Im\left[\varphi_C(t) \cdot \varphi_A(-rt)\right]}{t} dt \quad (8)$$

where φ_C and φ_A are the characteristic functions of W_C and W_A respectively. This integral was numerically evaluated with Gauss–Kronrod quadrature formula implemented in the QUADPACK library (Piessens et al. 2012).

The width of the confidence interval comes from both the spread of the mass distribution and the lack of information on this distribution (uncertainties). The spread of the mass ratio distribution depends on several factors including the conveyor monitoring duration and the stream purity. The longer the monitoring is, the more accurate the estimation is. As for the stream purity, the closer it is to 100% (or 0%), the more concentrated the distribution is. The amount of information on the mass distribution mainly depends on the number of sample analyses used for calibration and on the sampling duration. Increasing the number of sample analyses improves the estimation of the mass variance. As for the sampling duration, the longer the sampling duration is, the more objects are sampled and weighed. As a result, the more accurate the estimation of the mean is.

In a nutshell, for each inference, it is necessary to (i) compute the parameters of the posterior predictive distribution (equation (4)) and (ii) approximate the 2.5th and 97.5th percentiles of the ratio distribution by numerically inverting equation (8) with the secant method.

2.4. Data collection

2.4.1. Image Annotation

PortiK is based on supervised object recognition. Thus, annotated image data are required both for training and evaluation (sections 2.2.3 and 2.2.4). In this context, image annotation consists in first drawing axis-aligned bounding boxes surrounding the pieces of waste. Then a category called *label* (for example: "*aluminium can"* or "*contaminant"*) is assigned to each of them.

2.4.2. How many images?

A recurring question in applied supervised learning is about the amount of data needed. How fast a supervised system learns? In statistical learning, this is called the *rate of convergence*. Knowing the rate of convergence may be critical for sizing such a system. Unfortunately, as stated in (Devroye et al. 1996) (Theorem 7.2) universal rate of convergence guarantees do not exist: it depends both on the model and on the data.

Although there are no statistical means of sizing the training set, a commonly accepted heuristic is to consider that, given a model and a detection task (e.g. detecting aluminium

cans and contaminants), the required amount of data is quite stable. For image classification, a rule of thumb is to start with about 1,000 training images per label (Google AutoML Vision Documentation). The overall idea is to get a representative sample of all the possible appearances. Note that, in computer vision, the appearance also depends on the items' backgrounds. For object detection, one can adapt this heuristic and start with about 1,000 object occurrences for each label.

Once trained, the performances of object recognition models are estimated empirically on an independent *test* set. While the required size of *training sets* cannot be mathematically estimated, sizing *test sets* is possible. As described in section 2.2.4, most metrics used to assess the models' performances are calculated from the outcome of a series of independent success–failure experiments (Bernoulli trials). Therefore, methods for binomial proportion confidence interval estimation (Newcombe 1998) can be used to estimate the required size of the test set. For instance, based on the *Wilson score interval with continuity correction*, if the estimated probability is 90%, at least 3,556 sampled objects (resp. 913) will be required for the estimation error to be lower than 1% (resp 2%) with 95% of confidence (cf figure S1 in supplementary materials).

2.4.3. Mass information

To predict the mass related indicators from image analyses, detections of pieces of waste are converted into a mass estimation following the approach described in section 2.3.2. This approach requires some mass information for each waste category. To gather such information, the standard sampling and characterization process was adapted. In material recovery facilities, the current process consists in taking a sample of a waste stream from a conveyor then manually sorting the waste into different categories. Once sorted, the waste is then weighed category per category. The number of categories depends on the local legislation and/or contract; it ranges from a dozen to a few dozen. To get the required information, a counting step was added. Thus, before weighing the waste, the number of pieces were counted.

3. Experiments and results

To evaluate the solution in an operational environment, PortiK device has been installed in an MRF in the London area. More specifically, it monitored an aluminium can stream just before the picking room. For this waste stream, any items other than aluminium cans are considered as contaminants.

3.1. Materials

3.1.2. Annotated images

A first set of 678 images with 7,404 pieces of waste was acquired by the *PortiK* device in Decembre 2018. To ensure that the images were representative of the monitored flow, they were randomly selected and the position of the device was fixed throughout the experiments. Two categories were considered for detection: "*aluminium can*" (main stream) and "*contaminant*". Hence, the annotation tasks consist in drawing bounding boxes around each piece of waste and assigning it a binary category. These tasks were performed by waste characterization specialists. To assess the feasibility of having the annotation performed by non-specialists of waste characterization, an external company specialised in image annotation performed the same tasks on a subset of 100 images.

A second dataset composed of 99 images with 1,308 pieces of waste acquired in September 2019 was used as a *test* set to assess the evolution of the performances.

3.1.2. Waste samples

For the mass data acquisition, five one-minute samples were taken from the monitored conveyor and manually sorted in November 2019. Once sorted, the waste materials were counted and weighed category per category. This dataset was used to calibrate the mass estimation method (section 2.2.3). A second batch of seven other one-minute samples was taken between 22nd June and 2nd July 2020. Unlike the previous five samples, these were precisely timed so as to be compared with PortiK measurements. The waste composition estimated from these 12 samples is detailed in table 1.

3.2. Annotation

3.2.1. How long does it take?

Building datasets with ground truths is key to machine learning based approaches: they are required both for training and evaluation. In this study, a total of 8,712 pieces of waste from 777 images were both located using bounding boxes and labelled as either aluminium cans or contaminants. While drawing the bounding boxes arounds the pieces of waste took, on average, 7.9 seconds per item, assigning them a binary class took less than a second per item (0.95s). This corresponds to approximately 2h40 for 100 images.

Table 1 – *Average waste composition of the 12 one-minute samples taken from the conveyor and manually sorted and weighed.*

Category	Mass		Quantity (count)	
	kg / h	%	number / h	%
Aluminium Cans	315.5	66.1 %	23,260	64.1 %
Non Ferrous Metal (Non Can)	44.15	9.25 %	3,595	9.91 %
Other Prohibited	23.75	4.98 %	2,645	7.29 %
Aluminium Aerosols	21.00	4.40 %	475	1.31 %
Glass	12.80	2.68 %	590	1.63 %
Tetra Pak	12.60	2.64 %	400	1.10 %
Mixed Paper	11.70	2.45 %	1,575	4.34 %
Ferrous Metal (Non Can)	7.15	1.50 %	25	0.07 %
Other Plastics	5.35	1.12 %	545	1.50 %
Plastic Films (Non LDPE)	3.75	0.79 %	970	2.67 %
OCC	2.85	0.60 %	125	0.34 %
WEEE	2.65	0.56 %	15	0.04 %
LDPE Films (Plastic Bags)	2.45	0.51 %	640	1.76 %
Pots, tubs, trays & other plastic packaging	2.30	0.48 %	125	0.34 %
Tissues	2.05	0.43 %	675	1.86 %
News and Pams	1.10	0.23 %	140	0.39 %
NPET Bottles	1.10	0.23 %	40	0.11 %
Food Waste	0.85	0.18 %	80	0.22 %
Wood	0.65	0.14 %	35	0.10 %
Sanitary Ware	0.60	0.13 %	220	0.61 %
Garden Waste	0.35	0.07 %	40	0.11 %
CHDPE Bottles	0.30	0.06 %	15	0.04 %
Steel Cans	0.25	0.05 %	5	0.01 %
CPET Bottles	0.20	0.04 %	10	0.03 %
Other paper & card	0.20	0.04 %	25	0.07 %
Textiles	0.15	0.03 %	10	0.03 %
NHDPE Bottles	0.10	0.02 %	5	0.01 %
PP Bottles	0.05	0.01 %	5	0.01 %
Steel Aerosols	0.00	0.00%	0	0.00 %
PVC Bottles	0.00	0.00%	0	0.00 %
Yoghurt Pots and Margarine Tubs	0.00	0.00%	0	0.00 %
Soil and Rubble	0.00	0.00%	0	0.00 %
Batteries	0.00	0.00%	0	0.00 %
Fines less than 10mm	1.40	0.29 %		-
Total Sample Weight (kg/min)	**477.4**		**36,290**	

3.2.2. Can it be performed by non-specialists?

To assess the feasibility of having the annotations performed by non-specialists of waste characterization, the annotations performed by operators were compared to those done by an external company. When performed by the external company, the annotation tasks took much longer both for bounding box drawing (22 seconds per object) and object labelling (5.3 seconds per object). As for the annotations' quality, results are presented in table Table 2. Annotation errors can come from both the segmentation and labellisation. For the segmentation of 1,063 segmented objects, a precision level of 95% and recall level of 90% was achieved. When taking the labels into consideration as well, the scores dropped for most categories as they require more specific waste knowledge.

3.3. Object recognition for an aluminium stream

3.3.1. From standard detection performance to counting error

Among the 678 annotated images, approximately 80% (535 images with 5,770 objects) were used to train the object detection model (section 2.2.2). The other 143 images were used to evaluate the performances. The inference time was 137 ± 6 ms per image. Object detection performances are summarised in Table 3. Figure 2 shows some sample results. For the aluminium cans, the precision and recall were 91.2% and 90.3% respectively, resulting in an underestimation of the number of aluminium cans by less than 1%. As for the contaminant group, the performances were lower. The precision and recall were 80.2% and 78.4% respectively. Since the precision and recall are close, false positives and false negatives offset each other, resulting in an underestimation of the number of contaminants by 2.2%. All the test images with overlaid detections compared to the ground truths are available in a video in the supplementary materials.

3.3.2. Influence of the number of training images

To assess the influence of the number of training images, the model was also trained on a smaller dataset of 242 images (2,797 objects). In the following p-values are computed with Pearson's chi-squared test.

For the "*aluminium can*" category, the average precision was 93.40%. The precision was 90.7% with a 95% confidence interval of [87.7% - 93.1%]. The recall was 89.0% with a 95% confidence interval of [85.8% - 91.6%]. This corresponds to a counting error of -1.9%. Compared to the model with the full dataset, the performance drop was not significant: p-values of 0.78 and 0.46 for precision and recall respectively.

As for the "*contaminant*" category, the average precision was 75.9%. The precision and recall with their 95% confidence intervals were 76.8% [72.3% - 80.7%] and 79.6% [75.3% - 83.4%] respectively. This corresponds to a counting error of 3.8%. Compared to the model with the full dataset, the performance drop was not significant: p-values of 0.18 and 0.64 for precision and recall respectively.

Table 2 *– Quality of outsourced annotation on 100 images.*

Category	Number of objects	True positives	False Negatives: missing segmentation	False Negatives: wrong label	False positive: improper segmentation	False positive: wrong label	Precision	Recall
Segmentation only	1,063	952	111	-	45	-	95%	90%
Aluminium cans	533	476	24	33	13	25	93%	89%
Contaminants	530	418	87	25	32	33	87%	79%

***Table 3** – Performances after nine months of a trained object detection model. The model was trained t_0. A First evaluation was performed at t_0 and a second one was performed at t_1. Performances are given in terms of precision, recall and counting errors.*

Digital Sample	December 2018 (t_0)		September 2019 (t_1)	
Category	**Aluminium Can**	**Contaminant**	**Aluminium Can**	**Contaminant**
Average Precision	94.6 %	81.2 %	88.5 %	68.2 %
Precision	91.2 %	80.2 %	86.1 %	75.1 %
95% CI*	[88.6% - 93.3%]	[76.4% - 83.6%]	[82.3% - 89.2%]	[70.5% - 79.2%]
Recall	90.3 %	78.4%	83.8%	65.4%
95% CI*	[87.6% - 92.5%]	[74.5% - 81.9%]	[79.9% - 87.0%]	[60.8% - 69.8%]
# True Positives**	704.5	512.5	466.5	385.0
# False Positives**	68.0	126.5	75.5	128.0
# False Negatives**	75.5	141	90.5	203.5
Counting error	-0.96%	-2.2 %	-2.7 %	-12.8 %

* 95% confidence interval not corrected for multiple comparisons.
** objects at the boundary of the image are weighted by 0.5

3.3.3. Influence of the waste distribution on the belt

Object detection approaches are sensitive to occlusion, which happens in this case when objects are overlapping. To assess this dependence, the precision and recall were evaluated as a function of the mean intersection over union (IoU) between objects per image. Results are presented in figure 3. The recall significantly decreased when the mean IoU increased for both categories with a Spearman correlation R coefficient of -97.2% and -94.4% for the "*aluminium can*" and the "*contaminant*" respectively (p-value< 0.001) . While the precision of the contaminant also significantly decreased when the mean IoU increased (R = -88.8%, p-value < 0.001), the precision decrease for the aluminium can category was not significant (R = -10.5%, p-value = 0.25).

Note that when correlating the performance with the number of objects per image, the correlation coefficients were lower. For the "aluminium category" the correlation coefficients were -61.2% with the precision and -46.1% with the recall. As for the contaminant category, the correlation coefficients were 26.6% for the precision and -57% for the recall.

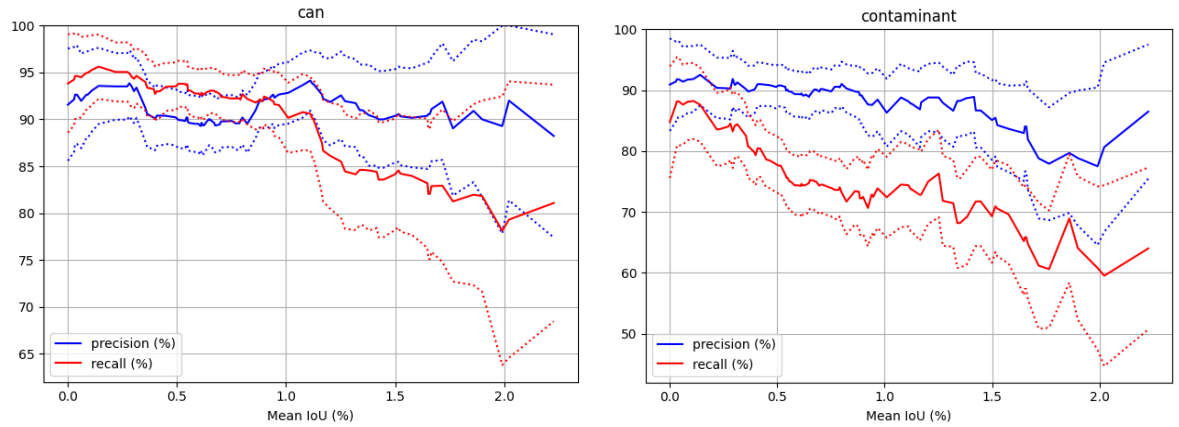


Figure 3 – *Precision and Recall as a function of the mean intersections over unions (IoU) between objects per image. The dotted line represents the 95% confidence interval.*

3.3.4. Performance drift overtime

To assess the performance drift after several months, the model trained on the first dataset was evaluated on the second dataset consisting of 99 images acquired nine months later, in September 2019 (table 3). Let t_0 be the date of the first dataset and t_1 the date of the second dataset.

For the aluminium cans, the precision dropped from 91.2% at t_0 to 86.1% at t_1 (p-value = 0.004). The recall dropped from 90.3% at t_0 to 83.8% at t_1 (p-value < 0.001). This yielded an underestimation of the number of aluminium cans increasing from 1% to 2%. While the performance loss in terms of precision was significant, the impact on the counting error was rather low.

As for the contaminant group, the precision dropped from 80.2% to 75.1% precision (p-value = 0.043) and the recall dropped from 78.4% to 65.4% (p-value < 0.001) at t_1. The precision remained stable while the recall dropped significantly. This yielded a 12.8% underestimation of the contaminants at t_1 while the underestimation at t_0 was 2.2%.

The performance drift could be due to different factors such as (i) change in the waste stream composition, (ii) denser waste stream with more overlapping pieces of waste or (iii) a change in the object appearance with for instance change in packaging.

To assess any change in the waste composition, the contaminants' composition was analysed by classifying them into the following categories: "*glass*", "*aluminium aerosol*", "*Tetra Pack*", *"aluminium foil"*, *"plastic film"*, *"metallized film"*, *or "other"*. The evolution of the composition is presented in table 4. The maximum difference occurred for the "*other contaminant*" with a 5.3% decrease. According to the Chi-square goodness of fit test, the change in distribution was significant (p < 0.001). In a worst case scenario where categories with increasing proportions have null precisions and recalls, such a change in the distribution would have made precision drop from 82.0% to 75.4% and recall drop from 78.4 % to 70.4%. Hence, it is unlikely that the change in distribution alone was the cause of the drop in performance.

As mentioned earlier (section 3.3.3), the amount of overlap also has an impact on performance. As a matter of fact, the average mean intersection over union per image had slightly decreased between t_0 (average: 0.92%; standard deviation: 0.87%) and t_1 (average: 0.88%; standard deviation: 0.72%). The evolution of the overlapping rate between t_0 and t_1 cannot explain the performance drop.

Table 4 *– Stream composition obtained from image annotations. Proportions are given in terms of number of pieces of waste.*

Category	December 2018 (t0)		September 2019 (t1)	
	Overall proportion	**Proportion among contaminants**	**Overall proportion**	**Proportion among contaminants**
Glass	0.9%	2.0%	3.1%	5.9%
Aluminium Aerosol	2.1%	4.6%	1.7%	3.2%
Tetra Pack	2.6%	5.6%	2.1%	4.0%
Aluminium foil	4.4%	9.4%	4.6%	8.8%
Plastic film	4.6%	9.8%	5.6%	10.7%
Metallized film	5.6%	12.0%	8.3%	16.0%
Other contaminant	26.3%	56.7%	26.8%	51.4%
Aluminium Cans	53.6%	-	47.9%	-

3.4. Mass proportion estimation

3.4.1. Comparison with 7 manual sample analyses

PortiK mass purity estimator was calibrated as described in section 2.2.3 with the first batch of five one-minute samples and evaluated on the second batch of seven one-minute samples. Results are presented in figure 4. The true values were always within the estimated confidence interval, which is promising regarding the validity of our approach. Note that the width of the confidence interval depends on different factors including the sample duration. This was confirmed by the aggregation of the seven sample analyses, which was equivalent to a single seven-minute sample. The 95% confidence interval width wass then much tighter.

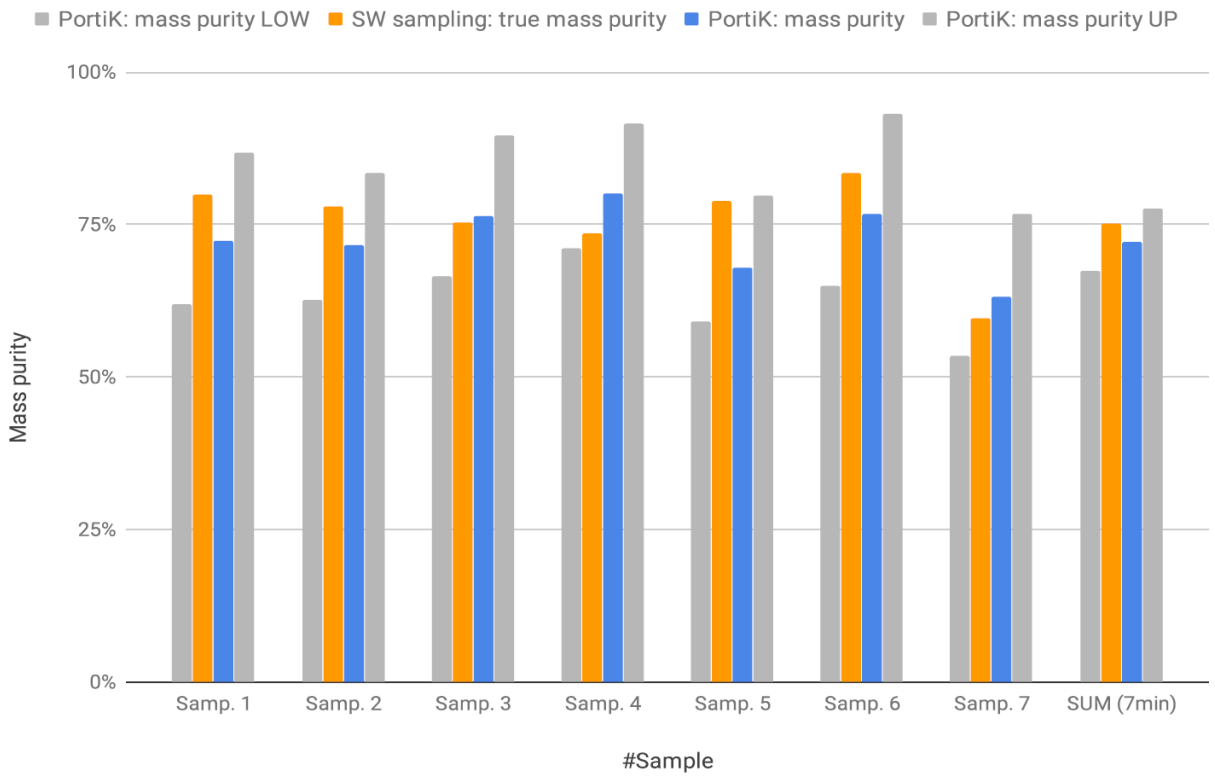


Figure 4 *– Purity estimated by PortiK system (blue) for an aluminium stream upstream of the picking room compared to seven 1 minute samples manually sorted and weighed (orange). Bars in grey indicated the 95% confidence interval estimated by PortiK. The aggregation of the seven samples (SUM) is equivalent to a single sample that would have lasted 7 minutes.*

The width of the 95% confidence interval of the mass purity estimation of aluminium can is computed varying the stream purity and the conveyor analysis duration. Results are presented in figure 5. For instance, for a mass purity around 50%, the result of an analysis with PortiK calibrated with the five one-minute samples would be precise up to ±7% after 5 minutes of monitoring. The result gets more and more precise the longer the analysis lasts: up to ±5% after 8 hours. Note that the confidence of an estimation varies dramatically depending on the stream composition.

Based on the five original samples, sample analyses were simulated to estimate the influence of their numbers and of their sizes. To show the feasibility of such study, results are presented in the supplementary figures S2 and S3 (supplementary materials). Note that the intervals are wider than in figure 5 because of the uncertainty in the sample simulation. It shows that for small numbers, increasing the number of sample analyses dramatically improves the estimation. Similarly, when small, increasing the samples' sizes strongly reduces the width of the confidence interval.

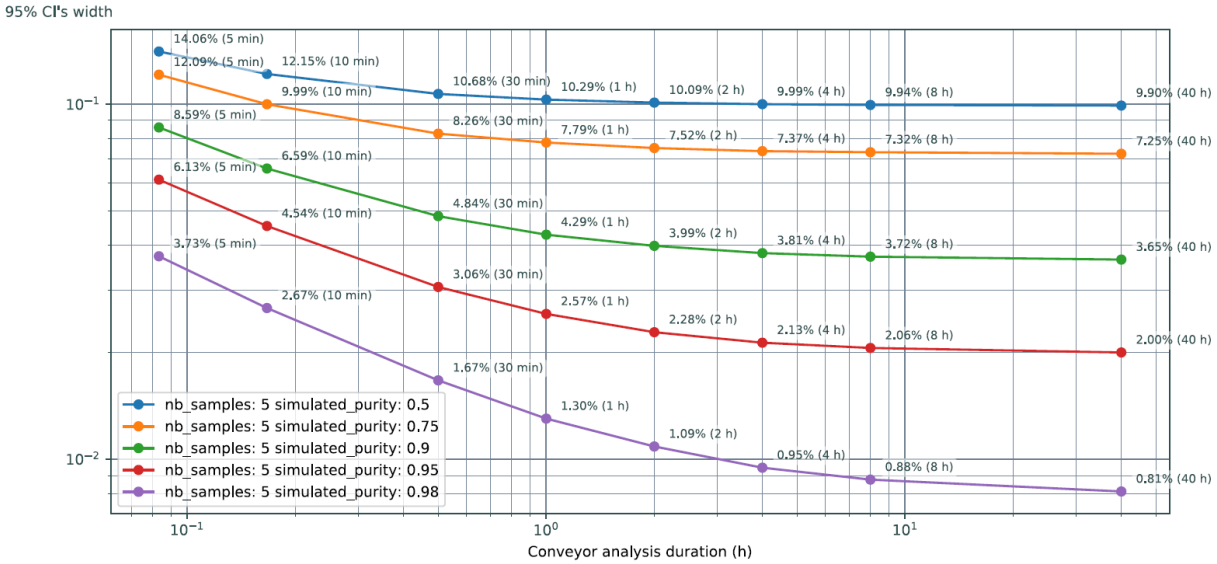


Figure 5 *– 75% quantile of the 95% confidence interval width for the aluminium mass proportion estimation as a function of the analysed time-frame duration for different target purities based on the 5 sample analyses.*

4. Discussion

4.1. Computer vision for waste characterization in MRF

Different studies demonstrated that image information could be used to efficiently discriminate between different types of waste (section 1.2). These convincing proof-of-concepts' results paved the way for the development of computer vision based solutions for SWM.

In this study, an object recognition algorithm was used to monitor an aluminium can stream in an operational environment. Aluminium cans were detected with 91.2% precision and 90.3% recall, respectively, resulting in an underestimation of the number of cans by less than 1%. For contaminants, precision and recall were 80.2% and 78.4%, respectively, giving an 2.2% underestimation.

Based on five sample analyses where pieces of waste were counted and weighed per batch, the detection results were used to estimate the mass purity. A confidence level on the estimation was also given. In a nutshell, the purity was estimated with an error within ±7% after 5 minutes of monitoring and ±5% after 8 hours.

This study showed that *PortiK* allows the operator to get a non-intrusive real time analysis of a waste stream. These performances are promising baselines and are expected to be outperformed by new image analysis solutions to come.

4.2. Factors affecting performance

4.4.1. Category to detect

The estimation accuracy depends on different factors. The proposed solution consists of a supervised object detection followed by some statistical analyses. The performance of the detection is affected by many factors. First of all, the *category to detect* has a major impact on the detection precision and recall. In this study, the "*aluminium cans*" were better detected than the "*contaminants"*. This was probably due to a greater variability in appearance for the latter category.

4.4.2. Waste distribution over the belt

The relation between the number of objects to detect within an image and the detection performance is not straightforward. Nevertheless, this study showed that object detection significantly degrades as the overlap between objects increases. A solution to limit occlusion could be to modify the waste distribution over the belt either using a spreader or by increasing the belt speed. When increasing the belt speed, the image quality can be maintained by decreasing the exposure time and increasing the light power. However, an exhaustive analysis of the stream could then be limited by the image analysis duration. The latter was approximately 150 ms per image.

4.4.3. Size of the training set

Deep learning recognition parameters have a huge amount of parameters to estimate and therefore require a lot of training images when trained from scratch. In this study, a transfer learning strategy was followed by fine-tuning a model pre-trained on MSCOCO Dataset (Lin et al. 2014). Thus, overfitting was avoided even with quite a small number of images. As a matter of fact this was not very different from using the pretrained model as a feature extractor followed by linear classifiers. Nevertheless, a training dataset has to be large and representative enough for a model to capture all the possible variability. Reducing the number of images used for training from 535 to 242 made the performance lower especially for the "*contaminant"* class. However the decrease was not significant.

Besides, with the chosen neural network architecture, more than the number of images, what matters is the number of objects. Indeed, in this study, the background being almost constant, two objects further away than the size of the *receptive field* could be considered as almost independent training samples.

4.4.4. Representativity of the training set and performance drift

Another factor that substantially affects the performance of the object detection is the representativity of the training set. This is often manifested by a performance drift over time. In this study, at the nine-month post-training evaluation, the performances dropped significantly. The performance drift could be due to different factors such as (i) change in the

waste stream composition, (ii) denser waste stream with more overlapping pieces of waste or (iii) a change in the objects' appearances with, for instance, changes in packaging. Hence, following the evolution of the monitored waste could be of interest to detect any potential drop in performance. However such a connection is not straightforward. Therefore, monitoring the performance of the AI model is critical. Note that the representativity of the data set used for calibration is also critical for the statistical analyses part.

As a matter of fact, the evolution of performances strongly depended on the category. While they remain quite stable for the "*aluminium can"* category, the degradation was more important for the "*contaminant"* category. This could be explained by the homogeneity of the aluminium cans category and the fact that it does not change much over time. On the contrary, the contaminants category is more heterogeneous and probably more subject to change in appearance over time.

4.4.5. Monitoring duration

The estimation error of the purity estimation comes from the counting errors, the number of samples analysis but also on the time window of the analysis. The more objects are monitored, the more accurate the purity estimation is.

Other factors such as the quality of the training/calibration data can also affect the performances of the detection.

4.3. A modular and more explainable model

As pointed out by Abdallah et al. (2020), adoption of AI based solutions is partly slowed down by their black box nature. Indeed, the explainability of machine learning models is an active area of research (e.g. Guidotti et al. 2018). Therefore, a modular approach that limits its black box component to the object recognition model was proposed. Object recognition being a classic computer vision problem, one can benefit both from research on new models and on their explainability (e.g. Selvaraju et al. 2017). In practice, splitting the overall approach in two different steps helped the adoption of *PortiK* for two main reasons. First, it helped understand the overall approach. Moreover, by displaying detection results, it allowed the end-users to qualitatively assess the solution performance. Besides acceptability, having a modular and explainable solution, also helped monitor its performance: object detection and mass estimation could thus be evaluated separately. Otherwise, the evaluation procedure would have been cumbersome and not feasible in MRF routine.

4.4. Overall cost

Adoption of new technologies in MRF is also driven by their costs. For the hardware, the main sources of costs are the camera, the lighting system and an industrial computer powerful enough to run convolution network inferences in real time. Adequate lighting must not be neglected since it strongly impacts the image quality. The most difficult costs to

estimate concerns the build, the maintenance and the performances' monitoring of the AI model. It requires building ground-truth datasets (section 2.4) both for the model calibration and for the performance evaluation. Sections 2.4.2 and 3.2 give some estimations of the required number of annotations and their durations. A few hundred images were used, which took a few hours to annotate. The model calibration usually requires a few days on a computer with a dedicated GPU. One should keep in mind that, since waste streams change with time, the calibration cannot be done once for all. Its update frequency depends on the waste streams to analyse. For the aluminium stream, the performances remained rather stable over six months.

4.5. Limitations and future work

4.5.1 Sensitivity to appearances of objects and background

The main limitation of such a computer based approach is that it is, by construction, sensitive to appearances of objects rather than to their chemical composition. Object detection consists in jointly discriminating objects from their background and labelling them based on their appearances. As a result, not only does the objects' appearances affect the performances but also the background. For instance, detecting a can over an empty conveyor is not the same problem as detecting a can in a dense news and pamphlets stream. As a result, a major limitation of such an approach is that, when changing the conveyor or the waste stream to monitor, one must at least re-evaluate the performance of the object detection and must probably retrain it.

4.5.2. Heavy tailed mass distribution

In this study, since the number of pieces of waste was large, the mass distribution of batches of objects was considered to follow a normal distribution. However, the masses of pieces of waste may have a heavy tailed distribution for some categories. This happens when the category of interest contains rare heavier objects. In this case, the convergence to the normal distribution is then very slow. An intuition for this slow convergence is given by the catastrophe principle:

$$P(X_1 + X_2 + \ldots + X_n > x) \sim P\left(max\left(X_1, X_2, \ldots, X_n\right) > x\right) \; as \; x \to \infty \qquad (9)$$

for a sum of n independent random variables X_i with common distribution. Using subexponential one-tailed distributions such as the Log-normal distribution could be of interest.

4.5.3 Dense or Uncountable streams

The proposed approach has been designed for countable streams. It assumed that all pieces of waste on the conveyor could be counted. With some waste categories such as some paper

waste, the notion of pieces of waste are not well defined. For such cases, instead of detecting objects, semantic segmentation (e.g. Long et al. 2015) should be used. It consists in assigning a (waste) category to each pixel. Then the second step would be either convert a surface to a mass or a surface purity to a mass purity. Note that assigning a mass estimation to each pixel of a given category is difficult since it requires knowing how much pixels' information is correlated. This is left for future work.

5. Conclusion

In a nutshell, this paper described an end-to-end solution based on computer vision to automatically provide the MRF operators with real time waste characterization. All the necessary steps to set-up and operate such a system were described. Information about the hardware specifications, the data collection and the automatic analysis were then given. The system was tested on an aluminium can stream. Aluminium cans were detected with 91.2% precision and 90.3% recall, respectively, resulting in an underestimation of the number of cans by less than 1%. Precision and recall obtained on the contaminants category were 80.2% and 78.4%, respectively, giving an 2.2% underestimation.

Furthermore, efforts were made to transform results from off-the-shelf machine learning models into operation indicators dedicated to waste analyses such as mass purity. Based on five sample analyses where pieces of waste were counted and weighed per batch, the purity was estimated with an error within ±7% after 5 minutes of monitoring and ±5% after 8 hours. Finally the overall system and its integration were evaluated on actual data with a prototype running a facility.

References

Abdallah M., Talib M.A., Feroz S., Nasir Q., Abdalla H., Mahfood B. (2020). Artificial intelligence applications in solid waste management: A systematic research review. *Waste Management*, *109*, pp.231-246. http://dx.doi.org/10.1016/j.wasman.2020.04.057

Adedeji O., Wang Z. (2019). Intelligent waste classification system using deep learning convolutional neural network. *Procedia Manufacturing*, *35*, pp.607-612. http://dx.doi.org/10.1016/j.promfg.2019.05.086

Aral R.A., Keskin Ş.R., Kaya M., Hacıömeroğlu M. (2018). Classification of trashnet dataset based on deep learning models. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 2058-2062). IEEE. http://dx.doi.org/10.1109/BigData.2018.8622212

Bottou L., Bousquet O. (2007). The tradeoffs of large scale learning. *Advances in neural information processing systems*, *20*, pp.161-168. https://dl.acm.org/doi/10.5555/2981562.2981583

Cai Z., Vasconcelos N. (2019). Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. http://dx.doi.org/10.1109/TPAMI.2019.2956516

Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. (2020). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Springer, Cham. http://dx.doi.org/10.1007/978-3-030-58452-8_13

Chen J., Mao J., Thiel C., Wang Y. (2020). iWaste: Video-Based Medical Waste Detection and Classification.. In the *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 5794-5797). IEEE. http://dx.doi.org/10.1109/EMBC44109.2020.9175645

Chollet F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE. https://doi.org/10.1109/CVPR.2017.195

Chu Y., Huang C., Xie X., Tan B., Kamal S., Xiong X. (2018). Multilayer hybrid deep-learning method for waste classification and recycling. *Computational Intelligence and Neuroscience*, *2018*. http://dx.doi.org/10.1155/2018/5060857

Cuingnet R. (2021). Bayesian Inference of Normal Distribution Parameters with Aggregate Data. Technical Note – Veolia. http://dx.doi.org/10.13140/RG.2.2.12986.72641

Curtiss J.H. (1941). On the distribution of the quotient of two chance variables. The Annals of Mathematical Statistics,12(4):409–421. http://dx.doi.org/10.1214/aoms/1177731679

De Carolis B., Ladogana F., Macchiarulo N. (2020). YOLO TrashNet: Garbage Detection in Video Streams. In 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1-7). IEEE. http://dx.doi.org/10.1109/EAIS48028.2020.9122693

Devroye L., Györfi L., Lugosi G. (1996). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4612-0711-5

Dhillon A., Verma G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2), 85-112. http://dx.doi.org/10.1007/s13748-019-00203-0

Donati L., Fontanini T., Tagliaferri F., Prati A. (2020). An Energy Saving Road Sweeper Using Deep Vision for Garbage Detection. Applied Sciences, 10(22), 8146. http://dx.doi.org/10.3390/app10228146

Everingham M., Van Gool L., Williams C.K., Winn J., Zisserman A. (2010). The Pascal Visual Object Classes (VOC) challenge. *International journal of computer vision*, *88*(2), pp.303-338. http://dx.doi.org/10.1007/s11263-009-0275-4

Funch O.I., Marhaug R., Kohtala S. and Steinert M. (2020). Detecting glass and metal in consumer trash bags during waste collection using convolutional neural networks. *Waste Management*, *119*, pp.30-38. http://dx.doi.org/10.1016/j.wasman.2020.09.032

Gil-Pelaez J. (1951). Note on the inversion theorem. Biometrika, 38(3-4):481–482. http://dx.doi.org/10.2307/2332598

Goodfellow I., Bengio Y., Courville A. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press. ISBN: 9780262035613

Google AutoML Vision Documentation. (january 2021) https://cloud.google.com/vision/automl/docs/prepare

Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. (2018). *A survey of methods for explaining black box models*. ACM computing surveys (CSUR), 51(5), pp.1-42. http://dx.doi.org/10.1145/3236009

Gundupalli S.P., Hait S., Thakur A. (2017). A review on automated sorting of source-separated municipal solid waste for recycling. *Waste management*, 60, pp.56-74. http://dx.doi.org/10.1016/j.wasman.2016.09.015

Hastie T., Tibshirani R. and Friedman J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media. ISBN: 978-0-387-21606-5

He K., Zhang X., Ren S., Sun J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). http://dx.doi.org/10.1109/CVPR.2016.90

He K., Gkioxari G., Dollár P., Girshick R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969). http://dx.doi.org/10.1109/TPAMI.2018.2844175

Hinkley D.V. (1969). On the ratio of two correlated normal random variables. Biometrika, 56(3):635–639, 1969. http://dx.doi.org/10.2307/2334796

Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. https://doi.org/10.48550/arXiv.1704.04861

Huang G., Liu Z., Van Der Maaten L., Weinberger K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708). http://dx.doi.org/10.1109/CVPR.2017.243

Krizhevsky A., Sutskever I., Hinton GE., (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105). http://dx.doi.org/10.1145/3065386

Lachaize M., Le Hégarat-Mascle S., Aldea E., Maitrot A., Reynaud R. (2016). SVM Classifier fusion using belief functions: application to hyperspectral data classification. In International Conference on Belief Functions (pp. 113-122). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-45559-4_12

Lachaize M., Le Hégarat-Mascle S., Aldea E., Maitrot A., Reynaud R. (2018). Evidential split-and-merge: Application to object-based image analysis. *International Journal of Approximate Reasoning*, 103, 303-319. http://dx.doi.org/10.1016/j.ijar.2018.10.008

LeCun Y., Bottou L., Bengio Y., Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), pp.2278-2324. http://dx.doi.org/10.1109/5.726791

Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-10602-1_48

Lin T.Y., Goyal P., Girshick R., He K., Dollár P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). http://dx.doi.org/10.1109/ICCV.2017.324

Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.Y., Berg A.C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-46448-0_2

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.J., Fei-Fei, L., Yuille, A., Huang, J. and Murphy, K., 2018. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 19-34). http://dx.doi.org/10.1007/978-3-030-01246-5_2

Long J., Shelhamer E., and Darrell T. (2015). *Fully convolutional networks for semantic segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440. http://dx.doi.org/10.1109/CVPR.2015.7298965

Mao W.L., Chen W.C., Wang C.T., Lin Y.H. (2021). Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling*, *164*, p.105132. http://dx.doi.org/10.1016/j.resconrec.2020.105132

Melinte D.O., Dumitriu D., Mărgăritescu M., Ancuţa P.N. (2019). Deep learning computer vision for sorting and size determination of municipal waste. In *International Conference of Mechatronics and Cyber-Mixmechatronics* (pp. 142-152). Springer, Cham. http://dx.doi.org/10.1007/978-3-030-26991-3_14

Newcombe R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, *17*(8), pp.857-872. http://dx.doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C857::AID-SIM777%3E3.0.CO;2-E

Oquab M., Bottou L., Laptev I., Sivic J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1717-1724). http://dx.doi.org/10.1109/CVPR.2014.222

Piessens R., de Doncker-Kapenga E., Überhuber C. W , Kahaner D. K. (2012). QUADPACK: A subroutine package for automatic integration, volume 1. Springer Science & Business Media ISBN 13: 9783540125532

Redmon J., Divvala S., Girshick R., Farhadi A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). http://dx.doi.org/10.1109/CVPR.2016.91

Ren S., He K., Girshick R., Sun J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149. http://dx.doi.org/10.1109/TPAMI.2016.2577031

Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.C. (2018). MobilenetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520). http://dx.doi.org/10.1109/CVPR.2018.00474

Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). http://dx.doi.org/10.1109/ICCV.2017.74

Sheng T.J., Islam M.S., Misran N., Baharuddin M.H., Arshad H., Islam M.R., Chowdhury M.E., Rmili H. and Islam M.T. (2020). An internet of things based smart waste management system using LoRa and tensorflow deep learning model. *IEEE Access*, *8*, pp.148793-148811. http://dx.doi.org/10.1109/ACCESS.2020.3016255

Silberman N., Sontag D.,Fergus R. (2014). Instance segmentation of indoor scenes using a coverage loss. In *European Conference on Computer Vision* (pp. 616-631). Springer, Cham. http://dx.doi.org/10.1007/978-3-319-10590-1_40

Simonyan, K., Zisserman A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. https://doi.org/10.48550/arXiv.1409.1556

Srinilta C., Kanharattanachai S. (2019). Municipal Solid Waste Segregation with CNN. In *2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)* (pp. 1-4). IEEE. http://dx.doi.org/10.1109/ICEAST.2019.8802522

Standley T., Sener O., Chen D., Savarese S. (2017). Image2Mass: Estimating the Mass of an Object from Its Image. In *Conference on Robot Learning* (pp. 324-333). https://proceedings.mlr.press/v78/standley17a.html

Strollo E., Sansonetti G., Mayer M.C., Limongelli C., Micarelli A. (2020), July. An AI-Based Approach to Automatic Waste Sorting. In *International Conference on Human-Computer Interaction* (pp. 662-669). Springer, Cham. http://dx.doi.org/10.1007/978-3-030-50726-8_86

Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). http://dx.doi.org/10.1109/CVPR.2015.7298594

Thanawala D., Sarin A., Verma P. (2020). An Approach to Waste Segregation and Management Using Convolutional Neural Networks. In *International Conference on Advances in Computing and Data Sciences* (pp. 139-150). Springer, Singapore. http://dx.doi.org/10.1007/978-981-15-6634-9_14

Tian Z., Shen C., Chen H., He T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 9627-9636). http://dx.doi.org/10.1109/ICCV.2019.00972

Toğaçar M., Ergen B., Cömert Z. (2020). Waste classification using AutoEncoder network with integrated feature selection method in convolutional neural network models. *Measurement*, *153*, p.107459. https://doi.org/10.1016/j.measurement.2019.107459

Tran D., Wang H., Torresani L., Ray J., LeCun Y., Paluri M. (2018). A closer look at spatiotemporal convolutions for action recognition. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 6450-6459)*. http://dx.doi.org/10.1109/CVPR.2018.00675

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762

White G., Cabrera C., Palade A., Li F., Clarke S. (2020). WasteNet: Waste Classification at the Edge for Smart Bins. *arXiv preprint arXiv:2006.05873*. https://doi.org/10.48550/arXiv.2006.05873

Witkovsky V. (2016). Numerical inversion of a characteristic function: An alternative tool to form the probability distribution of output quantity in linear measurement models. Acta IMEKO, 5(3):32–44, 2016. http://dx.doi.org/10.21014/acta_imeko.v5i3.382

Xiao W., Yang J., Fang H., Zhuang J., Ku Y. (2019). A robust classification algorithm for separation of construction waste using NIR hyperspectral system. Waste Management, 90, pp.1-9. http://dx.doi.org/10.1016/j.wasman.2019.04.036

Yang M., Thung G. (2016). Classification of trash for recyclability status. Stanford CS229 project report.

Yang Z., Luo T., Wang D., Hu Z., Gao J., Wang L. (2018). Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 420-435). http://dx.doi.org/10.1007/978-3-030-01264-9_26

Yosinski J., Clune J., Bengio Y., Lipson H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320-3328). https://doi.org/10.48550/arXiv.1411.1792

Yu Y., Zou S., Yin K. (2020). A novel detection fusion network for solid waste sorting. *International Journal of Advanced Robotic Systems*, *17*(5), p.1729881420941779. http://dx.doi.org/10.1177/1729881420941779

Zheng Y., Bai J., Xu J., Li X., Zhang Y. (2018). A discrimination model in waste plastics sorting using NIR hyperspectral imaging system. Waste Management, 72, pp.87-98. http://dx.doi.org/10.1016/j.wasman.2017.10.0

Supplementary Materials

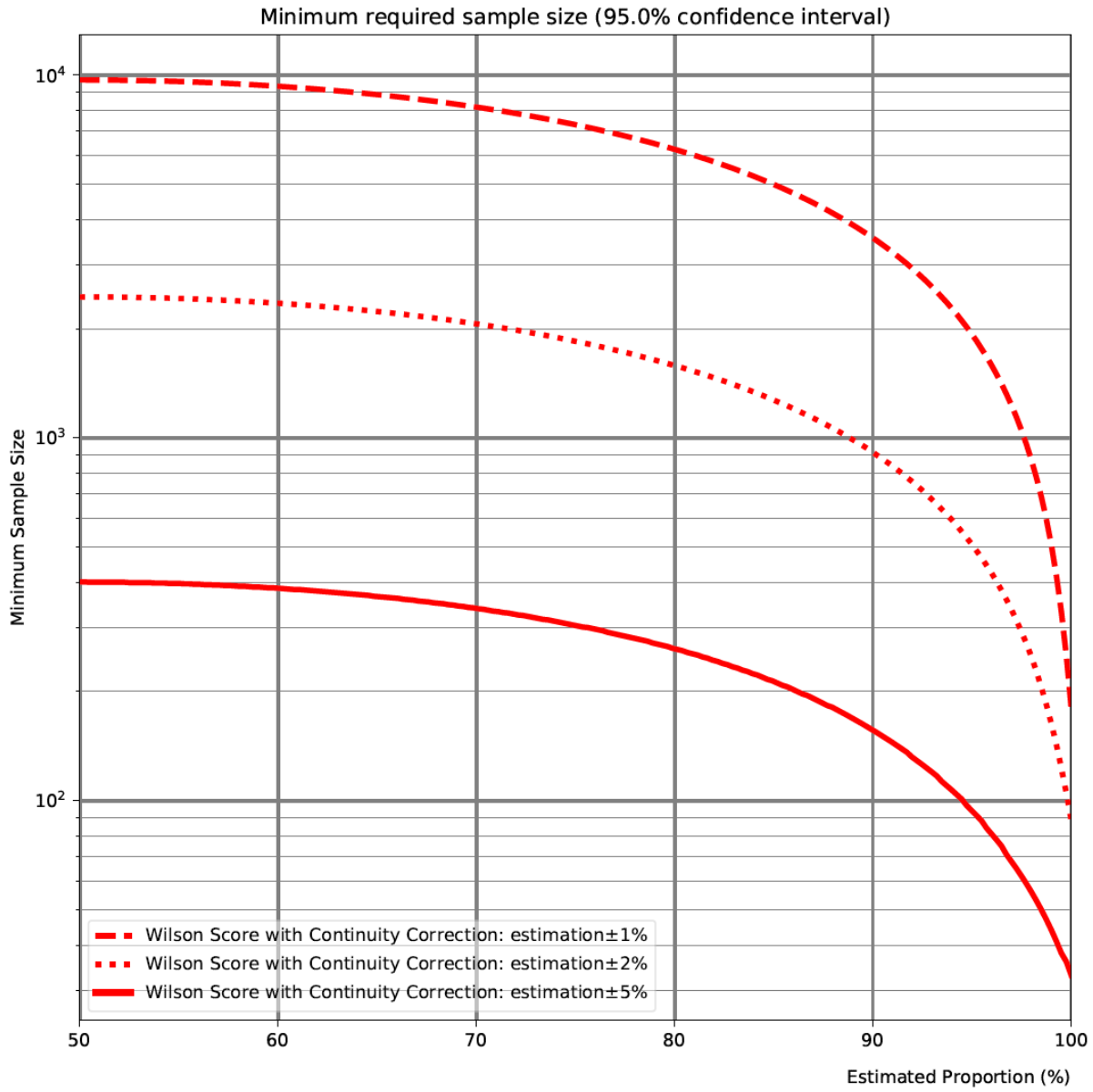


Figure S1 *– Minimum number of sample in a trial to get an estimation with a confidence interval smaller than or equal to ±1% (dashed), ±2% (dotted) and ±5% (plain).*

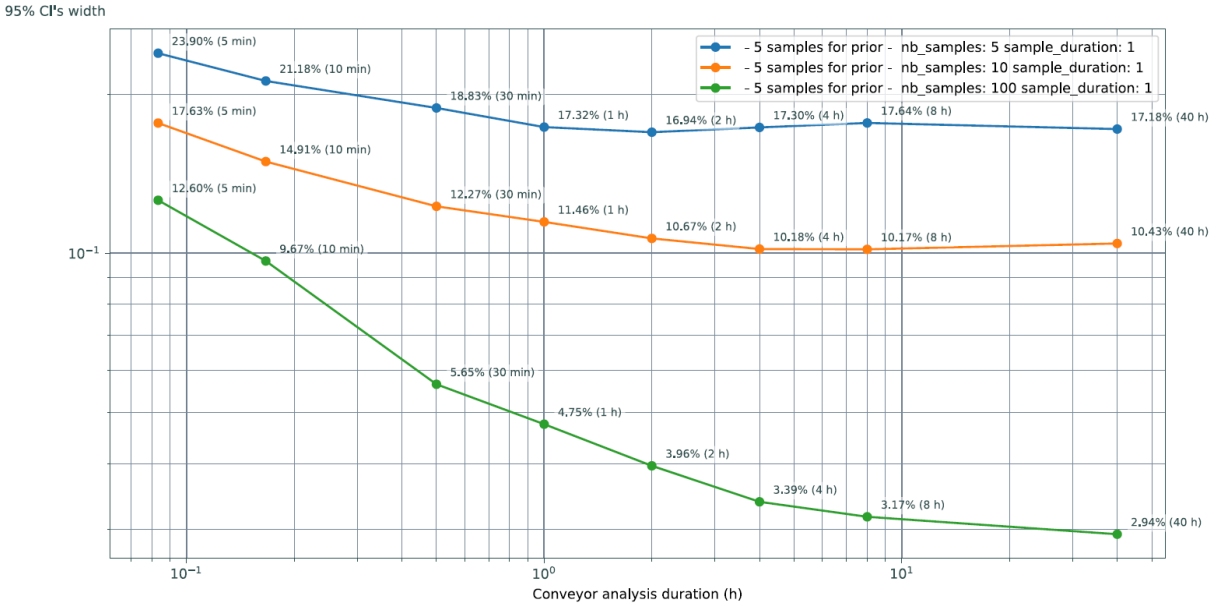


Figure S2 *– 75% quantile of the 95% confidence interval width for the aluminium mass proportion estimation as a function of the analysed time-frame duration for different numbers of simulated samples.*

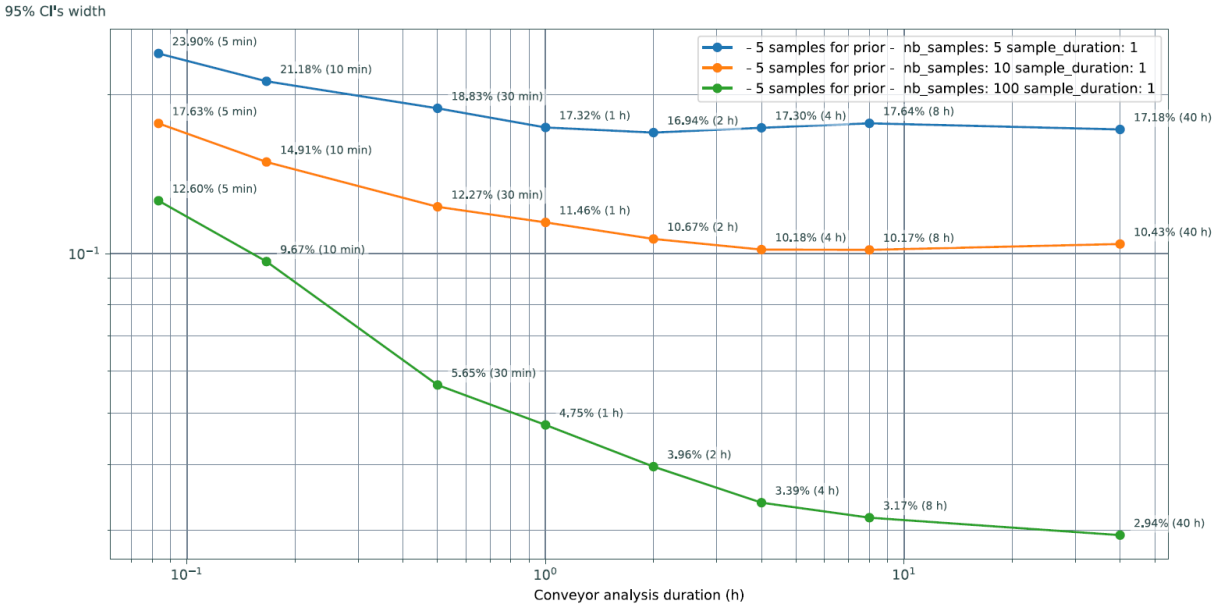


Figure S3 *– 75% quantile of the 95% confidence interval width for the aluminium mass proportion estimation as a function of the analysed time-frame duration for simulated samples of different durations.*