remote sensing

MDPI

Article

Explainable Automatic Detection of Fiber–Cement Roofs in Aerial RGB Images

Davoud Omarzadeh [1,†], Adonis González-Godoy [1,†], Cristina Bustos [1], Kevin Martín-Fernández [1], Carles Scotto [2], César Sánchez [2], Agata Lapedriza [3,4] and Javier Borge-Holthoefer [1,*]

[1] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, 08018 Barcelona, Catalonia, Spain
[2] DetectA, 08009 Barcelona, Catalonia, Spain
[3] e-Health Center, Universitat Oberta de Catalunya, 08018 Barcelona, Catalonia, Spain
[4] Institute for Experiential AI, Northeastern University, Boston, MA 02115, USA
* Correspondence: jborgeh@uoc.edu
† These authors contributed equally to this work.

Abstract: Following European directives, asbestos–cement corrugated roofing tiles must be eliminated by 2025. Therefore, identifying asbestos–cement rooftops is the first necessary step to proceed with their removal. Unfortunately, asbestos detection is a challenging task. Current procedures for identifying asbestos require human exploration, which is costly and slow. This has motivated the interest of governments and companies in developing automatic tools that can help to detect and classify these types of materials that are dangerous to the population. This paper explores multiple computer vision techniques based on Deep Learning to advance the automatic detection of asbestos in aerial images. On the one hand, we trained and tested two classification architectures, obtaining high accuracy levels. On the other, we implemented an explainable AI method to discern what information in an RGB image is relevant for a successful classification, ensuring that our classifiers' learning process is guided by the right variables—color, surface patterns, texture, etc.—observable on asbestos rooftops.

Keywords: asbestos; aerial imagery; deep learning; explainable AI; public health

Citation: Omarzadeh, D.; González-Godoy, A.; Bustos, C.; Martín-Fernández, K.; Scotto, C.; Sánchez, C.; Lapedriza, A.; Borge-Holthoefer, J. Explainable Automatic Detection of Fiber–Cement Roofs in Aerial RGB Images. *Remote Sens.* **2024**, *16*, 1342. https://doi.org/10.3390/rs16081342

Academic Editor: Lefei Zhang

Received: 2 February 2024
Revised: 15 March 2024
Accepted: 2 April 2024
Published: 11 April 2024

1. Introduction

Europe's building stock is aged and heterogeneous. Many of the existing buildings do not provide a healthy environment, and one of the main reasons is because they contain harmful substances, such as asbestos-containing materials. Although the use of asbestos was banned in the European Union (EU) approximately 25 years ago, asbestos can still be found in buildings, as it was widely used in the construction sector from 1970 onward. The safe removal of asbestos from the European building stock is a long-term strategic target, and it is addressed in several EU policy initiatives such as Council Directive 83/477/EEC (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31983L0477, Council Directive 83/477/EEC on the protection of workers from the risks related to exposure to asbestos at work (19 September 1983)) (1983), 2009/148/EC (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32009L0148, Directive 2009/148/EC of the European Parliament and of the Council on the protection of workers from the risks related to exposure to asbestos at work (Codified version, 30 November 2009)) (2009) on the protection of workers from the risks related to exposure to asbestos at work, and 2018/844 (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2018.156.01.0075.01.ENG, Directive (EU) 2018/844 of the European Parliament and of the Council on the energy performance of buildings (30 May 2018)) (2018) on building energy efficiency, which includes healthy environment regulation.

With the identification of severe health issues such as lung and respiratory diseases resulting from exposure to asbestos [1], the necessity for research is obvious, especially

considering potential disparities in exposure to asbestos and its health risks, i.e., environmental inequalities arising from factors such as socioeconomic status, geographical location, or the availability of safe housing. Precisely, the importance of identifying and addressing environmental inequalities has been recognized for many years by governmental and international bodies such as the U.S. Environmental Protection Agency, the European Environmental Agency, and the World Health Organization.

Prior to safe removal, however, the first necessary step is to locate contaminated buildings. Unfortunately, asbestos detection is a challenging task: field inventories via human exploration are labor- and time-intensive. On the other hand, the automatic detection of asbestos-containing rooftops has been carried out in different ways, and often with remarkable success (see Section 2). In all cases, high-resolution, multi-band remote sensing images were used in combination with machine learning algorithms [2]. In this situation, if the quality and resolution of imagery is satisfactory, the accuracy of asbestos rooftop identification should be assured by using modern technology. Unfortunately, access to multi-band imagery is not always possible—or, at least, at an affordable cost. Thus, computer vision methods with multi- and hyperspectral images can only be performed on relatively small areas due to the rapidly increasing costs to obtain such imagery. This explains why, in general, most of the literature is devoted to very specific areas; which leaves open the question for scalable and accessible methods and data that may overcome the mentioned limitations.

In this regard, one of the most economical ways to monitor vast geographical areas is by using remote sensing techniques and satellite/aerial imagery processing. In this work, we leverage Deep Learning (DL) techniques—more advanced than traditional machine learning approaches—with the aim of detecting the presence of asbestos on rooftops, exploiting solely orthorectified aerial imagery, which is often publicly available at high-resolution levels ($\approx$0.2 m/px), enabling the task to be extended to virtually any area in the world.

However, the downside to DL approaches is the lack of explainability. In general, DL models can achieve a high accuracy, but at the expense of high abstraction—eventually hindering the interpretability of their black-box representations [3]. This is particularly relevant in the case of asbestos identification, as the risk of shortcut learning [4] may arise. In machine learning, a shortcut solution occurs when the model relies on a simple characteristic of a dataset to make a decision, rather than learning the underlying relationships between the inputs and outputs. This can lead to models that are able to achieve high accuracy on the training data but perform poorly on new data that do not contain the same simple patterns. In our case of interest, a model might learn to identify images with asbestos rooftops by focusing on the surrounding area, rather than the more complex texture, color, or patterns of the asbestos itself.

While there are several methods to address the problem of explainable artificial intelligence [5], we focus here on the most successful one in the context of computer vision: Class Activation Mapping (CAM) [6]. The idea behind CAM is simple: to visually highlight which areas in an image are more informative for the prediction outcome of a given task given by a Convolutional Neural Network (CNN). CAM methods have been shown to be successful for interpretability tasks in several fields [7–11], including land cover mapping [12]—but not in the context of asbestos remote sensing, to the best of our knowledge.

In this work, we propose a DL pipeline with the aim to detect asbestos in RGB images; see Figure 1. We used aerial imagery information extracted in Catalonia (Spain), pre-processed them, and created a dataset with images manually labeled with *asbestos* and *non-asbestos* rooftops. For the learning process, we evaluated the performance of two CNNs, EfficientNetB0 [13] and ResNet50 [14], to automatically classify asbestos from building rooftop images. To address the interpretability challenges associated with neural network models, we employed CAM using the trained models to understand the classification outputs. We specifically used Gradient-weighted Class Activation Mapping (Grad-CAM) [15]. Finally, we conducted a quantitative analysis of Grad-CAM outputs generated by both

models to assess the localization of the highest heatmap values and determine their correspondence to areas containing asbestos. Our experimental findings indicate that the CNN models exhibit strong performance in the task of asbestos classification, with our best model yielding an accuracy of 92%. Moreover, the analysis of model explainability gives reliability to the learning patterns employed by the models. These results suggest the feasibility of applying this solution to various environments, as long as the aerial imagery is accessible.

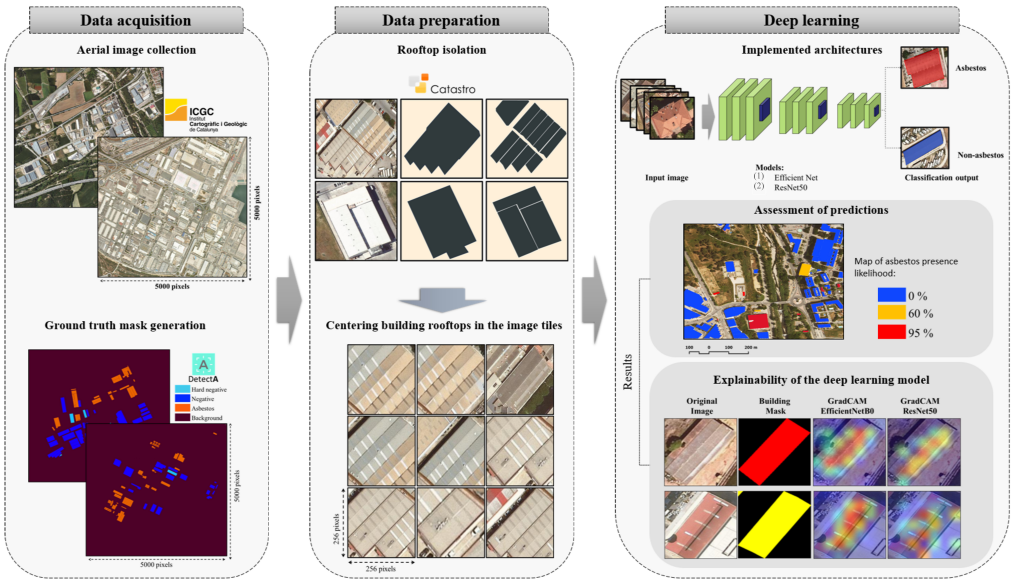

Figure 1. Data pre-processing and inference process overview. The left panel shows the two main data collection steps, while the central panel illustrates how buildings are isolated and centered owing to cadastral data. The right panel exemplifies, first, the classification task, which delivers a number in the range of $[0, 1]$ expressing the likelihood of the presence of asbestos in the image. The classification task undergoes a Grad-CAM analysis, delivering an interpretable heatmap to understand which part of the image is most responsible for the classification score.

The paper is organized as follows: Section 2 reviews the related work in asbestos automated classification, as well as some explainability works involving airborne imagery. Section 3 describes the aerial image dataset acquisition, the asbestos labeling process, the pre-processing of the images, the CNNs employed, and the explainability methodology and analysis. In Section 4, we present our experimental findings: the model's performance in the image classification task and the analysis of the explainability. In Section 5, we discuss the scope and limitations of our proposed pipeline. Finally, we draw our conclusions in Section 6.

2. Related Work

2.1. Remote Sensing in Urbanized Areas

The scholarly literature provides ample evidence that the utilization of computer vision and DL methodologies can facilitate the identification of objects using remote sensing images, both within and outside intricate urban environments. Applications encompass image scene classification [16–18], semantic segmentation [19], and change detection [20].

Many of these works are devoted to the retrieval of natural elements in the urban fabric, from the identification of green spaces, street trees, and other vegetation [21–23], to attempts to detect leaf nitrogen and biomass using red-edge band information [24].

Closer to the interests in this work, the identification of non-natural urban elements, has also been addressed [25,26]. For example, there is extensive work on classification and segmentation tasks to identify street elements, such as roads, sidewalks, or crosswalks [27,28], generally exploiting a computer vision approach from aerial imagery. Even

more specifically, off-street elements—mostly rooftops—have been the object of analysis in several occasions as well, e.g., to estimate photovoltaic (PV) capacity using geographic information data [29–31].

Focusing on the presence of asbestos, optical remote sensing systems have proven to be an efficient input for detecting and mapping asbestos-contaminated roofs. Szabó et al. [32], for example, identified various types of rooftops and assessed the presence of asbestos with the use of high-resolution airborne HyperSpectral Imagery (HSI), with a spatial resolution of 1 m and 126 spectral bands. To assess different types of roofs, the study used several machine learning classification algorithms from the ENVI+IDL 4.8 software to analyze images. The researchers conducted their study in Debrecen, East Hungary, using airborne imagery that covered a 7 km^2 area. The results of the analysis yielded precise maps of the roof types, with asbestos–cement roofs accurately identified at an accuracy rate of over 85%.

Similarly, Cilia et al. [33] employed HSI data with 92 channels between the visible to the shortwave infrared spectrum, as well as 10 bands in the thermal infrared region, with a spatial resolution of 3 m, to map asbestos–cement roofs and their weathering status, which refers to the condition of the roof surface due to long-term exposure to the environment. Such mapping was achieved with an image-based supervised classification method, using the Spectral Angle Mapper (SAM) algorithm [34], which was trained on a set of pixels selected from roofs made of different materials. The research was conducted in five municipalities in Northern Italy, with a combined area of 117.63 km^2. For this area, the task yielded an accuracy of 86%.

Krówczyńska et al. [35] conducted a study where aerial images with a spatial resolution of 25 cm in both natural color (RGB) and color infrared (CIR) compositions were employed. They devised and evaluated a CNN model specifically tailored to mapping asbestos–cement rooftops. The results demonstrated an overall accuracy ranging from 87% to 89%, depending on the used image composition. Furthermore, relying on RGB and CIR compositions, Raczko et al. [36] introduced an innovative CNN architecture for recognizing asbestos roofing, using a feature extraction block based on InceptionNet [37]. Remarkably, the study achieved an overall accuracy of the classification of different scenarios tested ranging from 88.0% to 93.0%.

These and similar studies have shown promising results indeed. However, they suffer from limitations in large-scale rooftop asbestos (or other materials, for that matter) identification due to the high costs and limited availability of HSI, infrared, or thermal bands [38]. As a result, even if those models are highly accurate, they may not be transferable or applicable to large areas [39]. On the other hand, even when free-access satellite images like Sentinel-2 MSI and Landsat series are available [40], their use in varied contexts is limited by spatial resolution constraints. For example, the analysis of compact urban areas, where buildings are typically smaller than 250 m^2, is not possible with publicly accessible imagery, unless high-resolution RGB imagery is used. As described in-depth later on, our work relies exclusively on RGB bands. This choice ensures our methodology's versatility and wide applicability. Furthermore, our pipeline places a strong emphasis on the imagery pre-processing, isolating rooftops prior to the classification stage, enabling us to pinpoint and classify different roofing materials with precision.

2.2. Explainability

Although DL has achieved remarkable success in real-world applications in various engineering fields as well as in remote sensing [36,41–43], it has been proven that the black-box nature of DL algorithms has limited the practical application of the results generated by these models [44]. This compels scholars to delve into understanding the logical reasons behind the DL models producing specific outputs. The visual interpretation of Class Activation Mapping (CAM) [6] and its derived methods, such as Grad-CAM [15], have been used as a supporting result to prove the effectiveness of several neural network proposals in several domains (e.g., [45]), including remote sensing. For example, in the task of understanding remote sensing scene classification, Shi et al. [46] introduced a dual-

branch, multi-level feature, and lightweight CNN. The authors applied their model to four open datasets and employed Grad-CAM to visually display extracted features, highlighting their significance through visual heat maps. In a similar manner, Chen et al. [47] proposed a multi-branch neural network architecture with local attention for remote sensing scene image classification. Utilizing Grad-CAM visualization, they compared the results of their network with baseline models like ResNet. Li et al. [48] presented an end-to-end architecture employing self-supervised contrastive learning for few-shot remote sensing scene classification. They also incorporated Grad-CAM to capture class-discriminative features, with the highest values on the heatmap pinpointing the exact location of the object's most crucial feature for scene classification.

In the task of geospatial object detection in remote sensing images, Li et al. [49] used weakly supervised DL. Their training method involves two stages: learning discriminative convolutional weights based on pairwise scene-level similarity and learning class-specific activation weights using scene-level labels. Then, object detection is achieved by segmenting the result of the CAMs. Notably, the deep networks were trained on an unrelated remote sensing image scene classification dataset, and testing was performed on a multi-class geospatial object detection dataset.

In other uses of explainability, various studies on remote sensing images have aimed to improve visual understanding, often extending the standard CAM formulation. For example, Huang et al. [50] presented a CAM adaptation with a neural network architecture, including encoder, classifier, reconstruction, and CAM modules, enhancing image classification between diverse object categories found in remote sensing images like airplanes, cars, bridges, beaches, residential areas, forests, etc. Their key contribution is the reconstruction module, preserving crucial object location information for improved CAM visualization. Guo et al. [51] proposed Prob-CAM, a CAM variant based on layer weights and a metric that quantifies the probability of occlusion in saliency maps for each convolutional layer. This enables the automatic selection of the optimal layer for visual explanations in the tasks of land-use classification in scenes from aerial orthoimagery. Song et al. [52] presented Bidirectional Gradient Verification (BiGradV) to refine visual explanations produced by Grad-CAM, capitalizing on both positive and negative gradients for class discrimination in remote sensing images. Dutta et al. [53] integrated Grad-CAM with a directed acyclic graph (DAG) generated by a neural network to enhance comprehension, specifically for the classification of remote sensing images of land-use. In the context of aircraft recognition in remote sensing images, Fu et al. [54] proposed Multi-CAM, leveraging predictions from all categories to mitigate errors from a single prediction category. A mask filter strategy further aids in eliminating interference from background areas.

Lastly, other studies have incorporated CAMs as part of their neural network architecture proposal, like Li et al. [55], leveraging the visual information contained in CAM beyond visual explanations through a weakly supervised approach to enhance building extraction in semantic segmentation within the domain of remote sensing imagery.

As discussed, the utilization of CNN explainability techniques, including CAM, Grad-CAM, or customized adaptations, has been prevalent in remote sensing imagery solutions employing DL. This approach ensures confidence in the learned patterns by the neural network or allows for the modification of CAM for enhanced and tailored comprehension. Notably, there is a gap in the literature regarding the exploration of CNN explainability within the specific context of visually analyzing asbestos or hazardous materials on rooftops. Given the critical importance of explainability in our work, where asbestos can be found in diverse environments, ranging from industrial to rural areas, it is crucial to ascertain whether our models are learning intrinsic visual patterns or relying on spurious relationships with objects or contextual factors surrounding the buildings.

3. Materials and Methods

3.1. Aerial Imagery and Asbestos Localization

Our study was conducted in Catalonia, located in the North-East area of the Iberian Peninsula; see Figure 2. Within Catalonia, we focused on the municipalities for which we manually collected asbestos rooftop locations. These municipalities ("Bages" and "Vallès" include more than one municipality, mostly small rural towns, which we aggregated under a single label for convenience; "Zona Franca" is not a municipality but is actually the main industrial cluster in the city of Barcelona) are highlighted in blue, and their names and some features are specified in Table 1. Notably, the vast majority of asbestos rooftop locations were collected in the Barcelona province (green shadow), except for a few asbestos locations which were obtained outside that area.

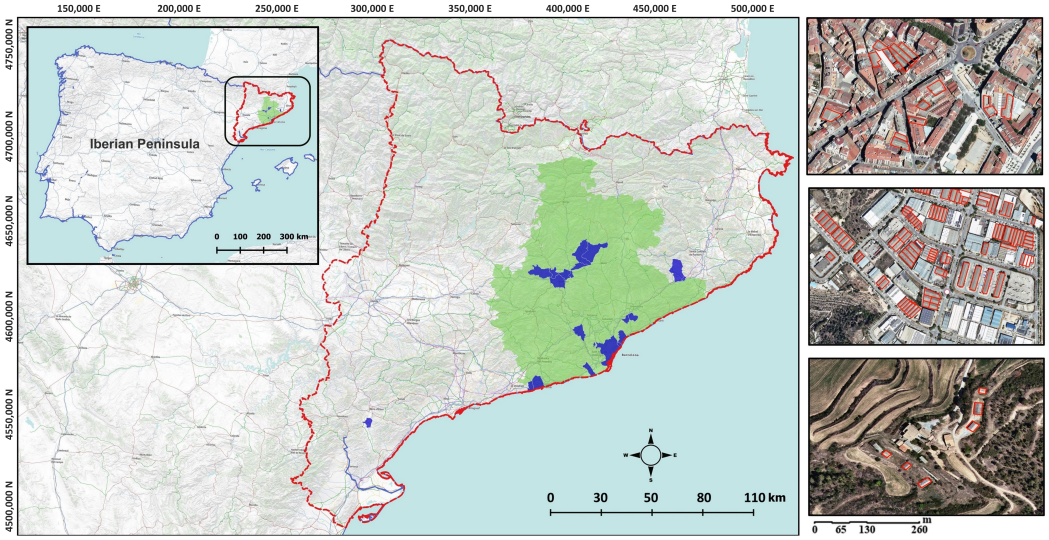

Figure 2. Geographic positioning and ground truth data distribution: the study was centered in the Catalan region, with particular attention to the Barcelona province owing to its landscape diversity, including rural areas, dense urban centers, and important industrial zones, where asbestos constructions are very frequent. Ground truth data were rigorously gathered from various sites (municipalities highlighted in blue). The accompanying magnified images on the right side are illustrative of the ground truth diversity, including urban, industrial, and rural areas.

Table 1. Aerial imagery details table. Municipalities, number of images corresponding to the studied locations, and area specifications.

Area Name	Number of Images	Covered Area (km^2)
Badalona	71	110.9375
Sant Adrià	4	6.25
Bages	285	445.3125
Zona Franca	19	29.6875
Vilanova i la Geltrú	20	31.25
Vallès	32	50
Castellbisbal	44	68.75
Cubelles	15	23.4375
Gavà-Viladecans	16	25
Ginestar	6	9.375
Hostalric	9	14.0625
La Verneda	6	9.375
Total	527	823.4375

This amounts to an area of 823.4375 km^2, which, in turn, corresponds to 527 aerial photographs, each with dimensions of 5000 × 5000 pixels and a spatial resolution of 20 cm (resampled from 25 cm) in RGB composition, acquired from the Institut Cartogràfic i Geològic de Catalunya (Catalan Cartographic and Geologic Institute; ICGC hereafter) (https://www.icgc.cat/, accessed on 1 February 2024). Each photograph represents 1.5625 km^2.

Localization data for the asbestos–cement roofing were obtained from DetectA (https://en.detectamiant.com/, accessed on 1 February 2024), a local company working on dismantling asbestos cement roofs. In particular, the annotated field inventory was collected manually and in situ in the period between April and August, 2022. These data, gathered from the 13 municipalities (Table 1), amount to 4386 buildings in total in areas that may be classified as urban, rural, and industrial (see Figure 2; magnified images on the right side). Indeed, the Barcelona province is highly heterogeneous in population distribution, with highly dense urban and industrial areas on the coast, and a slow progression to smaller peri-urban and rural areas as we move inland. The manual classification included two categories, namely, the expected *asbestos* (2420 instances) and *non-asbestos* (1966 instances). Worth noting, the negative class includes *hard non-asbestos* instances (525), i.e., negative cases where certain colors and textures can be confused with the presence of asbestos material on a rooftop.

3.1.1. Aerial Imagery GIS Pre-Processing

Aerial images must undergo some transformation for them to be later processed by the DL architecture (Section 3.2.1). To start with, the classification task operates only on rooftops. Accordingly, individual buildings need to be presented as isolated as possible and centered in the image.

To accurately isolate buildings from aerial images, we first obtained the shapefile delineating the separation of buildings from the INSPIRE Services of Cadastral Cartography (http://www.catastro.minhap.gob.es/webinspire/index_eng.html, accessed on 1 February 2024). This file classifies buildings based on their adjacency, even allowing for the differentiation of attached structures.

The outcomes of this stage successfully isolated sizeable buildings and even small, scattered ones (see Figure 3). With some further processing, it also allowed for the distinction of small and large buildings that are merged together in large blocks, as it is often the case in cities (see Figure 3, second column). The resulting raster is a two-class layer containing *every* building in the image, with class 0 representing the background and class 1 representing the footprint of the rooftops. This binary mask dataset is useful later on during the explainability quantification process (see Section 3.2.3).

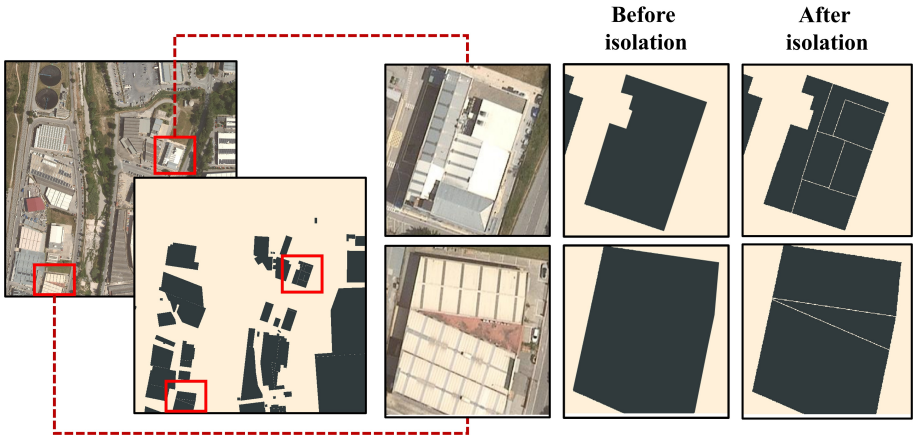

Figure 3. Rooftop isolation and centering. Catastral data enable a precise isolation and centering of individual properties, even when these are embedded in blocks which appear to be single facilities.

3.1.2. Ground-Truth Training Dataset Construction

Taking a subset of the previous general collection (every building represented), we generated a training dataset from the available binary masks of rooftops. To achieve this objective, the ground truth data collected by DetectA were the main inputs. First, the geometry features were analyzed to avoid any possible geographical misalignment between vector layers and the aerial images. Subsequently, the data were converted into a three-class raster, with an assigned code for asbestos, non-asbestos, and "background" (referring to anything other than the building's rooftop). That is, the dedicated masks, with the same dimensions as each image frame (5000 × 5000 pixels), were generated in the form of a four-valued matrix (Figure 4).

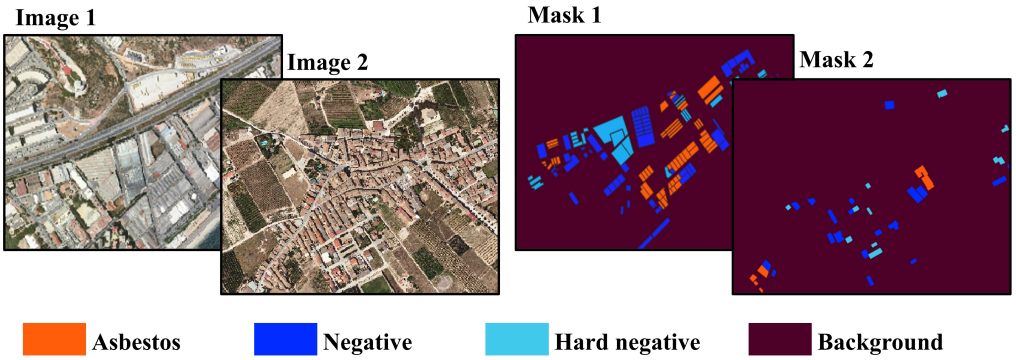

Figure 4. The pair of images and their corresponding masks as ground truth data during the training phase. The variety of buildings and structures in both rustic and urban areas was considered during the ground truth data collection. The figure includes examples of the sub-class *hard non-asbestos* for the sake of illustration, although it was not used in the training stage.

Images were then adapted to satisfy the input requirements of the CNN, i.e., 224 × 224 image input with three channels. To do so, images were cropped into smaller tiles so as to keep the buildings' roofs in the center of the image. Worth remarking, the centering method created image tiles of different sizes, so each building image was re-scaled to the same 224 × 224 dimensions.

We split the data into train, validation, and test sets. In total, there were 2244 images in the training set, with 1168 instances belonging to the positive class (asbestos class) and 1076 instances in the negative class, which also included hard negatives samples (hard negatives are still non-asbestos samples, but they are very similar to the positive class). In the validation set, there were 448 instances, which constituted 20% of the training set. The testing set consisted of 559 instances. This test set included 291 instances from the positive class and 268 instances from the non-asbestos class. Examples of the final input images given to the CNN are offered later.

3.2. Classification with Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a subtype of Artificial Neural Networks. Their ability to automatically learn discriminative features from raw pixel data makes them a powerful image classification method. In this section, we detail the specific CNN we propose to classify asbestos roofs.

3.2.1. CNN Architectures

In this work, we compared two popular DL architectures widely used in computer vision: EfficientNetB0 [13] and ResNet50 [14,56].

EfficientNetB0 is a convolutional neural network that combines efficiency and performance. Its core components include depth-wise separable convolutions, mobile inverted bottleneck blocks inspired by MobileNetV2 [57], and squeeze-and-excitation blocks [13]. This network uses a scalable approach, adjusting the depth, width, and resolution simul-

taneously to accommodate a wide range of resource constraints. Training techniques like dropouts and batch normalization enhance its performance. The benefits of this architecture include its ability to efficiently process images while maintaining competitive performance.

On the other hand, ResNet50 is a deep convolutional neural network architecture, part of the ResNet (Residual Network) family. The network uses residual blocks with skip connections, enabling very deep networks. It employs pre-activated residual blocks and global average pooling, and it is composed of 50 layers, making it effective for image classification tasks. Its state-of-the-art performance on various benchmarks, along with its suitability for transfer learning, has established it as a prominent choice in computer vision.

We selected these CNN models instead of state-of-the-art architectures such as transformers because of their high interpretability capability (through techniques such as Class Activation Mapping [6]) without compromising performance. Additionally, these models offer feature representations from pre-trained datasets that can be effectively fine-tuned for classification tasks in various domains. In order to obtain a more comprehensive overview of widely adopted DL architectures, various layer types, and loss functions within computer vision, we direct the reader to [58].

We trained both the EfficientNetB0 and ResNet50 architectures for the binary task of *asbestos* vs. *non-asbestos* classification. We added a dropout layer and fully connected layer of two neurons to each model at the end (right after the global average pooling). We used Softmax as the activation function at the end of both models, where the values of the Softmax output represent the probabilities of the input belonging to each class. The loss function used for this purpose was categorical cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \cdot \log(\hat{y}_i) \tag{1}$$

where C is the total number of classes, y_i is the true class label, and $\hat{y}_i$ is the predicted class probability. This loss function is commonly used in multiclass classification tasks, and it quantifies the dissimilarity between predicted class probabilities and true class labels. While binary classification problems can be addressed through binary cross-entropy loss, categorical cross-entropy can still be applied with a slight modification. During evaluation, we interpreted the predicted probabilities as the probability of belonging to the *asbestos* class by applying a threshold (0.5) to convert probabilities into binary predictions.

3.2.2. Training Details

In this section, we describe the training process in detail. The models were pre-trained with ImageNet [59], which is one of the largest publicly available labeled image datasets. ImageNet is a significant dataset in computer vision, known for its extensive collection of labeled images covering a diverse range of categories. It has advanced the field by facilitating the training of deep neural networks for tasks such as image classification. Models trained on ImageNet have become the starting point for a wide range of computer vision applications, making it a valuable resource. This pre-training allows the model to learn an initial visual feature representation that can be transferred to other visual classification tasks by fine-tuning the model.

Once the models were initialized with pre-trained weights, we used the collected dataset to fine-tune the models for the asbestos classification task. We used the training set for the learning process and the validation set for checking the model' hyperparameters.

To prevent overfitting, the dropout rate was set to 0.5, which means that during each training iteration, 50% of the units in the previous layer (after the global average pooling and batch normalization) will be randomly set to zero. We used a batch size of 32 samples, and we used Adam Optimizer as a gradient optimization algorithm with a learning rate of 10^{-3}. During training, we monitored the model's performance on the validation set based on the categorical cross entropy loss function (defined in Equation (1)) and used it as an early stop criterion. Additionally, we implemented a learning rate schedule that gradually reduced the learning rate to a fraction of 10^{-1} when the loss function fell into a plateau.

Finally, for model evaluation, we used Accuracy and F1 score metrics, defined as follows:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{3}$$

where TP is the amount of True Positives, TN the amount of True Negatives, FP the amount of False Positives, and FN the amount of False Negatives.

3.2.3. Explainability of CNNs with Class Activation Maps

We used a post hoc explainability technique to gain information on how the CNNs make their prediction. The technique is called Class Activation Maps (CAM) [6]. Concretely, given an input image and a specific output class (e.g., asbestos), CAM computes a heatmap highlighting the image regions that played the most significant role in determining the classification score for the specific output class.

In this work, we specifically used Grad-CAM [15], which is an extension of the original CAM technique that can be applied to CNNs that contain fully-connected layers before the output layer (notice that the CNNs we experimented with in this work—EfficientNetB0 and ResNet50—have both fully-connected layers before the output layer). Specifically, Grad-CAM computes importance scores per pixel of the input image by utilizing the gradients of the class score, denoted as y^c, with respect to the feature maps A^t_{ij} of the final convolutional layer. These scores are calculated as follows:

$$\alpha^c_t = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^t_{ij}} \tag{4}$$

In this equation, Z represents the total number of elements in a feature map. This setup leverages the feature maps derived from the gradient values of the class score with respect to the feature maps, leading to the following equation:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}(\sum_t \alpha^c_t A^t) \tag{5}$$

Grad-CAM integrates the feature maps A^t with their corresponding importance weights α^c_t, spotlighting the input image areas most pertinent for predicting the target class c. Specifically, it computes the gradients of the class score relative to the activations from the last convolutional layer, multiplies these by the activations, and applies a Rectified Linear Unit (ReLU) function. This process generates a heatmap, emphasizing the critical regions impacting the class prediction. The final map of Grad-CAM is then generated by summing up these target class weights and then re-sizing the heatmap to the input image size.

The goal of the explainability study presented in Section 4.3 is to assess if the crucial information to recognize a true asbestos rooftop actually comes from the rooftop area withing the image, as expected. Alternatively, there might exist spurious correlations between the asbestos rooftops and the surroundings of the building. For example, it might happen that there are more asbestos rooftops in industrial areas, where the streets follow a certain prototypical layout. In the presence of this specific spurious correlation, the model could be learning discriminant features about the surroundings of the rooftop rather than the rooftop itself. If the CNNs are actually relying on the rooftop characteristics to recognize the presence of asbestos, the CNNs should be making the decision based on the region of the image containing the rooftop, leading the CAM to highlight the rooftop region as the most informative. In contrast, if there are spurious correlations with the surrounding of the buildings, then the CNNs might focus on these surrounding areas, and CAM would indicate that the most discriminant information is outside the rooftop area.

In Section 4.3, we first present a qualitative assessment with CAM, and then a quantitative evaluation. For the quantitative evaluation, we calculated the ratio of the overlapping of the highest Grad-CAM values with the rooftop area within the image. First, we generated the Grad-CAM heatmaps for each building image belonging to the test set. The Grad-CAM heatmaps underwent a min-max normalization process between 0 and 1. Second, we took the binary mask for each building, which indicates, within the image, pixels corresponding to the building's rooftop and pixels that do not. These two data allowed us to combine the Grad-CAM heatmaps with rooftop masks and compare them pixel-wise.

More formally, we use G^k to denote the number of pixels belonging to the top $k = \{5, 10, \ldots, 100\}$ highest values of the Grad-CAM output; G^k_{rt} the is number of pixels belonging to the top k highest values of the Grad-CAM output that lie within the rooftop area; and g_k is the percentage of the k highest values that lie over the rooftop area. Then, we compute the following ratio:

$$g_k = \frac{G^k_{rt}}{G^k} \times 100 \tag{6}$$

This ratio, which ranges from 0 to 100, quantifies the fraction of rooftop pixels that overlap with the mask in the positive asbestos class: high values of g_k indicate that the model is actually relying on the rooftop area.

4. Results

In this section, we initially assess the performance of both models with two well-known techniques—random test set and k-folds cross validation. In both cases, we report the results using Accuracy, F1-score, and confusion matrix as metrics. Furthermore, we provide qualitative examples of the asbestos prediction map in different context areas. Finally, we analyze the explainability behind the decision-making of both models using Grad-CAM. We present insights from the qualitative visualizations of the Grad-CAM heatmaps and a quantitative analysis of the image regions the models are fixating on.

4.1. Random Test Set

Our first assessment was based on randomly separating the test data, a straightforward and computationally less intensive method. We left 20% of the dataset images for testing, and the remaining 80% were used for training and validation. Specifically, for the test set, we obtained 291 instances for the *asbestos* class and 268 instances for the *non-asbestos* class. For a fair comparison, the learning process was performed with the same training and testing sets for both networks. Table 2 presents the accuracy (Equation (2)) and F1-score (Equation (3)) of the predicted outputs given by each model's inference on the test set.

Table 2. Results of both CNN models' inference on the random test set.

Models	Accuracy	F1-Score	Asbestos Samples	Non-Asbestos Samples
EfficientNetB0	0.92	0.92	291	268
ResNet50	0.81	0.80	291	268

Tables 3 and 4 are the confusion matrix for EfficientNetB0 and ResNet50, respectively. As it can be observed, the EfficientNetB0 model has better performance than ResNet50, presenting a lower number of false positives and false negatives. Figure 5 shows three examples of the asbestos prediction map using the predictions given by EfficientNetB0 for each rooftop building in different context areas. The maps represent the model's prediction (after Softmax) for the *asbestos* class, using a color gradient to visualize the output. In this representation, rooftops with a lower Softmax likelihood are shaded in blue, while those with a higher likelihood are colored in bright red.

Table 3. Confusion matrix for the prediction outputs of the model EffiecientNetB0. "GT" stands for Ground Truth and "Pred" for Prediction.

		Pred	
		Asbestos	Non-asbestos
GT	Asbestos	275	16
	Non-asbestos	31	237

Table 4. Confusion Matrix for the prediction outputs of the model ResNet50. "GT" stands for Ground Truth and "Pred" for Prediction.

		Pred	
		Asbestos	Non-asbestos
GT	Asbestos	242	49
	Non-asbestos	60	208

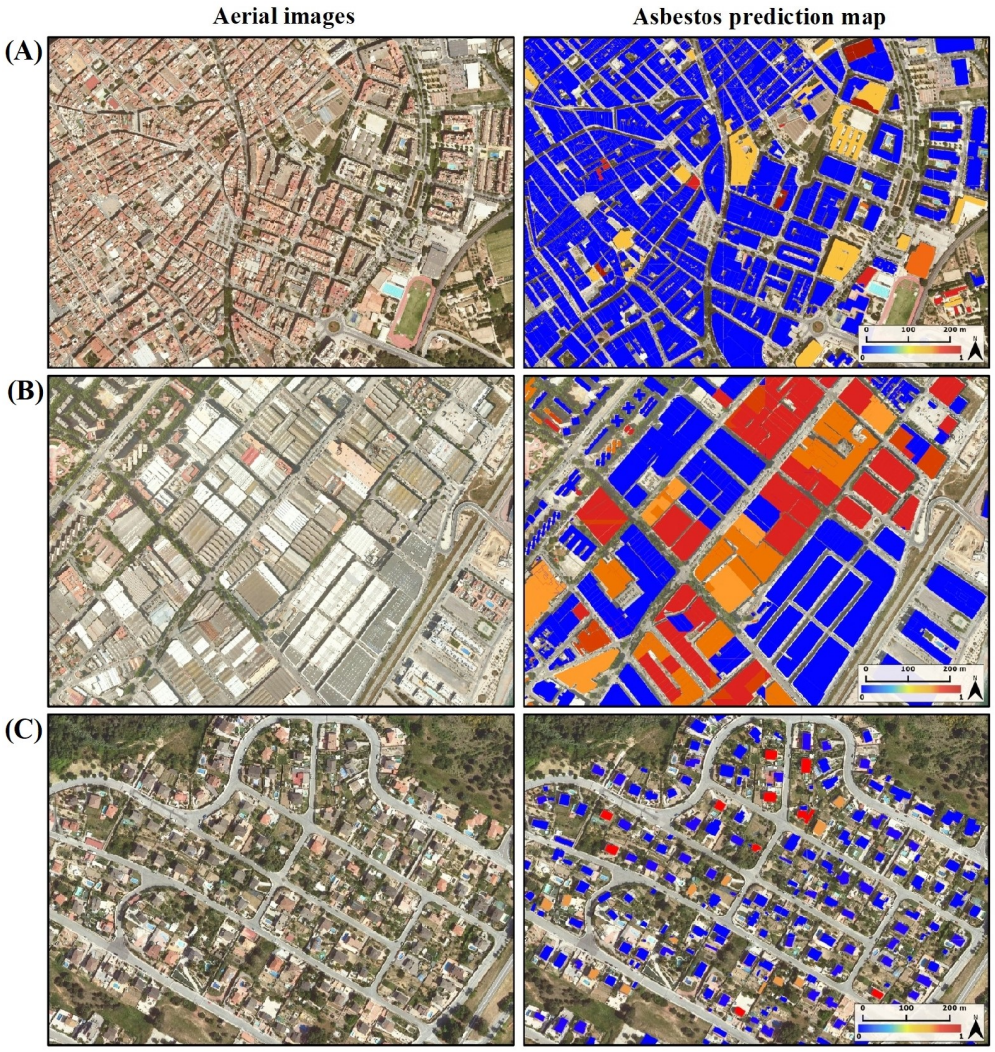

Figure 5. Asbestos prediction maps in various contexts including (**A**) dense urban center; (**B**) industrial area; and (**C**) rural/peri-urban area. As expected, the predicted level of asbestos presence is much lower in urban and rural areas, compared to industrial clusters.

4.2. k-Fold Cross Validation

Widely adopted in machine learning, *k*-fold cross-validation is a technique for statistically evaluating a model's accuracy and robustness. This approach ensures that the robustness of the implemented CNN models remains independent of the specific samples assigned to the training and test sets. In this experiment, we used *k*-fold cross validation with *k* set to 5. We trained both CNN models several times using different training and test sets. Specifically, we randomly partitioned our dataset into five groups, using one group for testing and the remaining four groups collectively for training.

The results of the 5 folds, reported in Table 5, show consistency, ruling out the possibility of a spurious impact of the randomness in a single split.

Table 5. Results of *k*-folds.

Networks	k_1-Fold	k_2-Fold	k_3-Fold	k_4-Fold	k_5-Fold	Avg Accuracy
EfficientNetB0	0.81	0.88	0.89	0.86	0.85	0.86
ResNet50	0.78	0.81	0.81	0.83	0.75	0.79

4.3. Explainability Results

We produced Grad-CAM heatmaps for each rooftop image in the test set, targeting both classes in our classification problem. In order to generate the heatmaps, we specifically used the trained models presented in Section 4.1. The Grad-CAM was generated based on the image's classification output; if the image was classified as *asbestos*, the corresponding Grad-CAM for asbestos was generated, and likewise for the *non-asbestos* class.

4.3.1. Qualitative insights

Figure 6 shows some examples of the original building image, the binary mask indicating where the building is located, and the Grad-CAM heatmaps for the two models tested: EfficientNetB0 and ResNet50. Rows (A) and (B) contain images belonging to the *asbestos* class whilst rows (C) and (D) contain images from the *non-asbestos* class.

As it can be observed in Figure 6, the highest values in the Grad-CAM heatmaps are notably localized within the building rooftop for both models. This observation suggests that the models effectively capture intricate rooftop patterns, encompassing elements such as texture, lines, and colors or the presence of those patterns together.

Upon comparing the heatmaps of both models, it becomes evident that the EfficientNetB0 heatmaps exhibit a tendency to cover a more expansive area compared to the ResNet50 maps. This discrepancy may be attributed to the architectural differences between the models, where EfficientNetB0, with its 4 million parameters, contrasts with ResNet50, which incorporates 23 million parameters. Consequently, ResNet50, due to its greater parameter count, may exhibit a tendency to overfit and concentrate on specific details within the building rooftop, while EfficientNetB0 demonstrates a broader focus on the entirety of the rooftop.

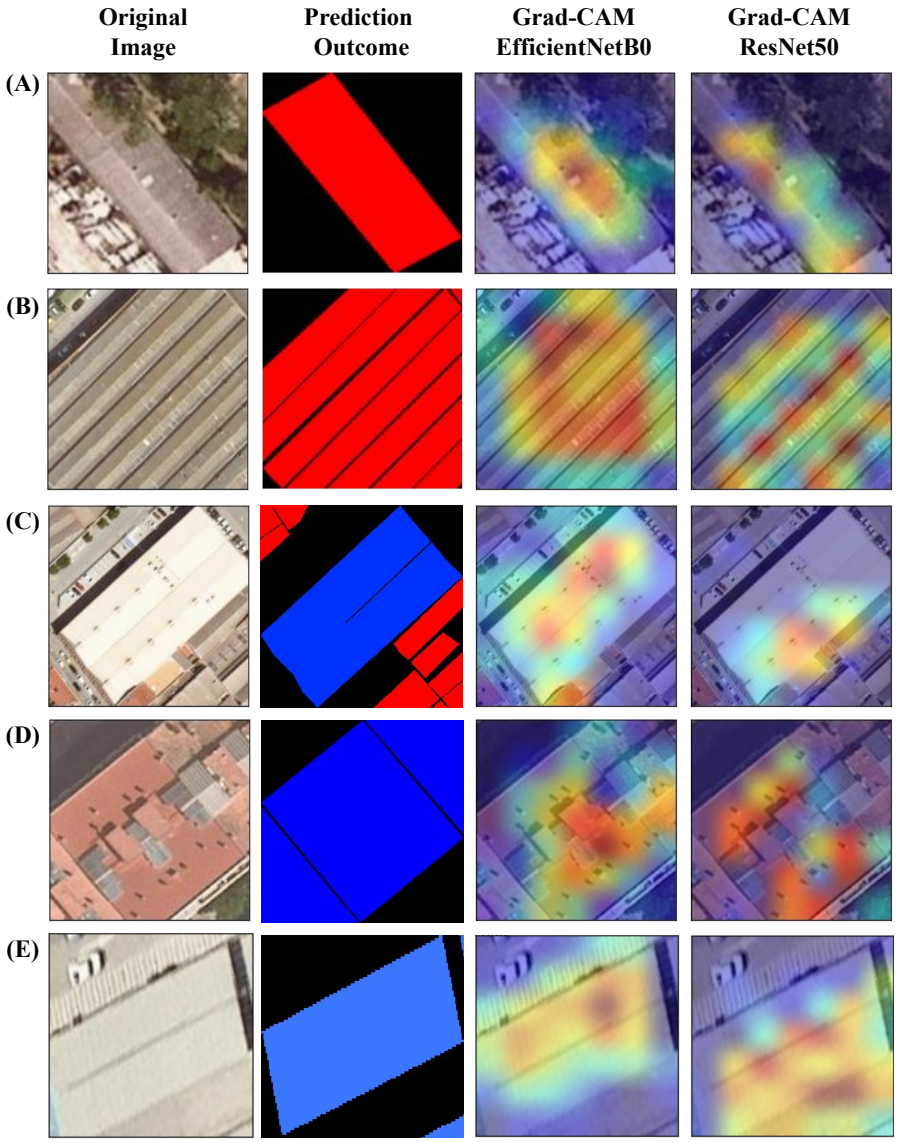

Figure 6. Grad-CAM results for selected *asbestos* (**A**,**B**), *non-asbestos* (**C**,**D**), and *hard non-asbestos* (**E**) images. The second column of the figure represents the EfficientNetB0 prediction outcome, while third and fourth columns show the Grad-Cam results in both models: a heatmap suggesting which parts of an image played a relevant role during the classification task.

4.3.2. Quantitative Analysis

After generating Grad-CAM heatmaps for all the images in the test set, we specifically analyzed them for the class of interest: *asbestos*. The objective of this analysis was to systematically check the locations with the highest Grad-CAM values within the original input images. Essentially, we investigated whether these values were concentrated within the building rooftop area or if they extended into areas outside the building structure, following the methodology described in Section 3.2.3. In particular, if the highest Grad-CAM values aligned with the rooftop area, we could conclude that our models successfully learned the intricate nuances associated with asbestos-containing patterns in our dataset.

For every image classified as *asbestos*, we generated a bar plot to illustrate the outcomes of g_k (refer to Equation (6)), ranging from $k = 5$ to $k = 100$. In this graphical representation, the x-axis corresponds to the top-k values, while the y-axis indicates the corresponding values of g_k. This visual presentation allows for a clear depiction of the relationship between the varying values of k and the resulting g_k values for each image. Figure 7 illustrates some examples of the original input image from the test set classified as asbestos

(first column), its corresponding EfficientNetB0 Grad-CAM heatmap overlapped with the building mask (second column), the g_k vs. k histogram of EfficientNetB0 Grad-CAM (third column), its corresponding ResNet50 Grad-CAM overlapped with the building mask (fourth column), and the g_k vs. k histogram of ResNet50 Grad-CAM (last column). Furthermore, in Figure 7, rows (A) and (B) display images from an urban context, rows (C) and (D) present images from industrial areas, and lastly, rows (E) and (F) display images from rural/peri-urban areas.

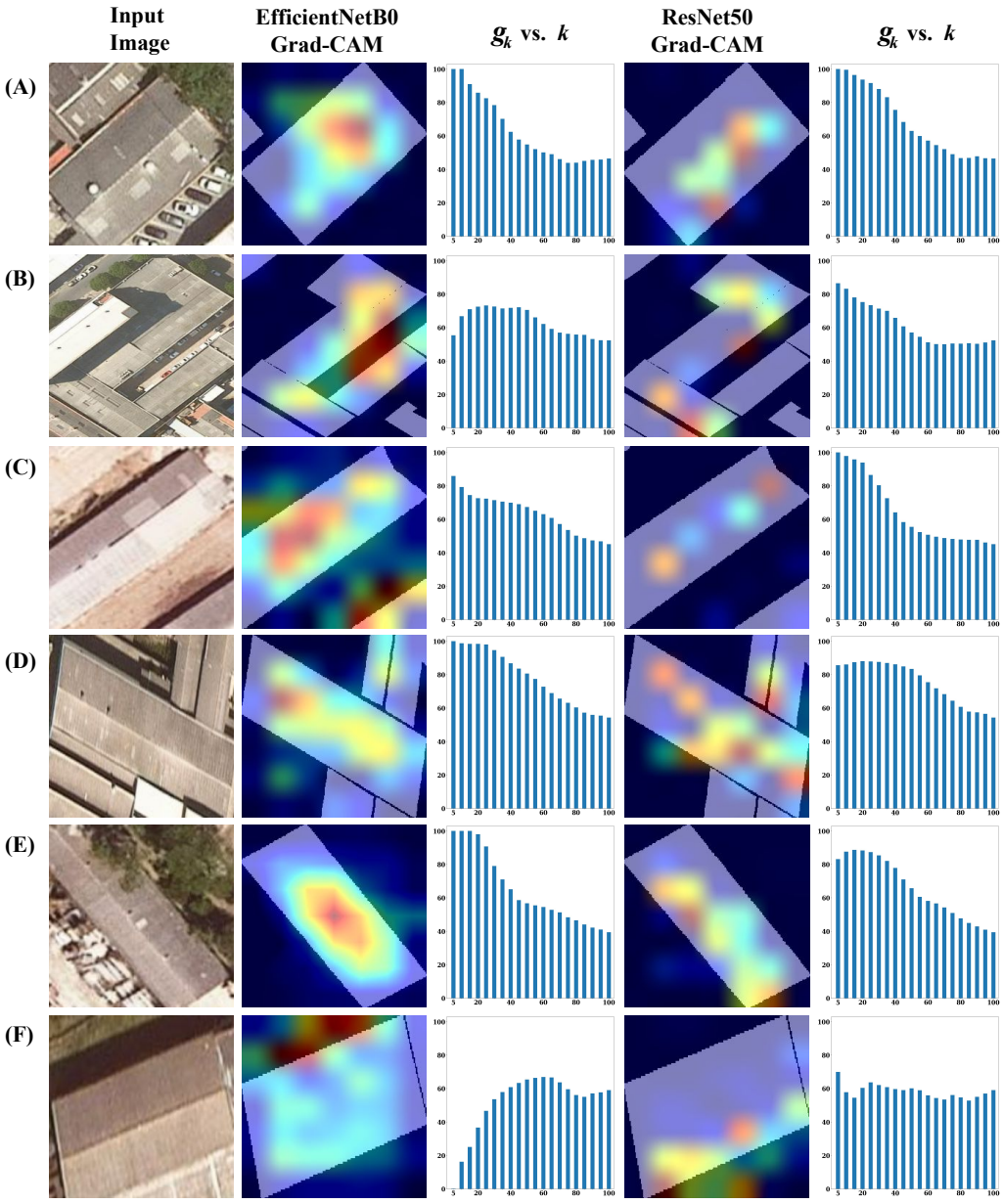

Figure 7. Grad-CAM values of rooftop overlapping g_k for individual buildings. For a set of facilities with a positive classification (first column), the Grad-CAM heatmap against the building mask is shown for both EfficientNetB0 and ResNet50 (second and fourth columns). Further, the g_k histograms (third and fifth columns) represent how Grad-CAM values are distributed on the image, with respect to the rooftop surface. (**A**,**B**) are samples from a dense urban context. (**C**,**D**) are samples from an industrial context and (**E**,**F**) are from rural/peri-urban contexts.

In Figure 7, row (A) shows an instance where 100% of the top five ($k = 5$) highest Grad-CAM values are concentrated within the rooftop area for both models Grad-CAMs, both for

EfficientNetB0 and ResNet50. This observation suggests that the models are predominantly focusing on the rooftop region. As the value of k increases, corresponding to less relevant Grad-CAM values, g_k exhibits a decreasing trend. This reduction in g_k continues until it reaches the ratio percentage of the building area concerning the total image size at $k = 100$.

Row (B) illustrates an instance where the highest activation values from the Grad-CAM, generated by EfficientNetB0, are mostly situated outside the building's perimeter. This deviation could be attributed to the building's atypical shape. On the other hand, the Grad-CAM heatmap produced by ResNet50 tends to be more concentrated; hence, approximately 80% of the top five ($k = 5$) activations are accurately located within the rooftop area of the building.

In rows (C), (D), and (E), we observe instances where approximately 80% of the top five highest activations in the Grad-CAM heatmaps, for both models, are predominantly located within the rooftop. While there are subtle differences in the distribution of g_k between the models, the heatmaps from EfficientNetB0 are generally wider but remain mostly within the building's structure. Conversely, despite ResNet50's good performance in terms of accuracy, its g_k distribution tends to be marginally less focused compared to EfficientNetB0. This variance in focus and spread is mirrored in the respective Grad-CAMs generated by each model, highlighting the nuanced differences in how each model perceives and processes the features for classification. In rows (C) and (D), where neighboring buildings are present alongside the main building, we note that certain Grad-CAM values extend beyond the boundaries of the primary building. Notably, the adjacent buildings are also labeled and classified as *asbestos*. This observation suggests that the model's attention is not solely fixated on the specific rooftop of interest but is influenced by the presence of additional rooftop structures within the image, contributing to the distribution of the Grad-CAM values.

Finally, row (F) presents a diverging example with respect to the previous. Instead, the concentration occurs in the borders or shadows of the building. In the case of EfficientNetB0, the highest values are distributed outside the rooftop's upper area, and along its borders to some extent. In the case of ResNet50, while some top values are within the rooftop, a considerable number are positioned outside the rooftop's lower boundaries. Examples like this one demand a more in-depth analysis. One plausible hypothesis for the misalignment between the Grad-CAM's focus and the actual rooftop could be attributed to the output of the model, specifically from the Softmax likelihood. It is conceivable that the Softmax likelihood values for these examples are not as high compared to other instances presented in Figure 7. This discrepancy in model confidence might cause the Grad-CAM attention to be more influenced by subtle features such as borders or shadows rather than the main rooftop area.

Figure 8A shows the average of the Grad-CAM quantification for all the images in the test set classified as *asbestos*. A comparative analysis is provided for the Grad-CAM values generated by EfficientNetB0 and ResNet50. Notably, for EfficientNetB0, on average, approximately 80% of the top five highest values are situated within the building's rooftop area, whereas for ResNet50, this percentage is slightly lower at 75%. This observation indicates that, on average, EfficientNetB0's Grad-CAMs exhibit a more precise localization within the rooftop area compared to ResNet50, persisting even up to $k = 95$ (note that for $k = 100$, the proportion is trivially equal to the percentage of areas occupied by the building under scrutiny; that is, when $k = 100$, g_k is the same no matter which classification model is used). This alignment with the accuracy results reported in Section 4.1 further supports the notion that EfficientNetB0 demonstrates a slightly superior performance in the classification task, reflected not only in accuracy but also in the spatial localization accuracy as indicated by the Grad-CAM analysis.

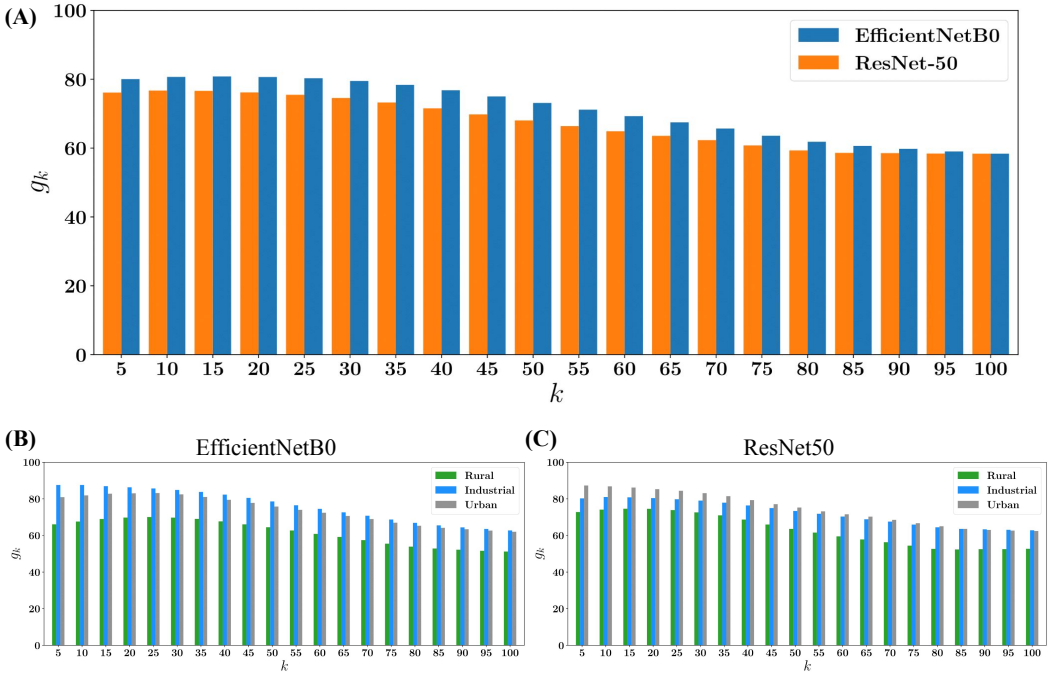

Figure 8. (**A**) Average Grad-CAM g_k for all the images classified as *asbestos* in the training set with EfficientnetB0 (blue) and ResNet50 (orange). (**B**,**C**) Average Grad-CAM g_k for all the images classified as *asbestos* divided by area context for each model: EfficientNetB0 (**B**) and ResNet-50 (**C**).

On the other hand, panels (B) and (C) of Figure 8 illustrate the average of the Grad-CAM quantification for each context area—rural, industrial, and urban—in our dataset, as analyzed for EfficientNetB0 and ResNet50, respectively. Images of rooftops from rural environments exhibit the weakest performance compared to the other two contexts. Our hypothesis is that it is common to see lichen-covered asbestos rooftops, which deviate significantly from the usual and expected grayish color. Additionally, we observe that the Grad-CAM outputs from EfficientNetB0 demonstrate superior performance in industrial settings, where, on average, almost 90% of the top five highest activations are correctly localized within the rooftop area. Conversely, ResNet50 exhibits better performance with urban images. These plots provide a comprehensive overview of how each model generally responds to different environmental settings, facilitating a deeper understanding of their contextual differences.

Despite EfficientNetB0's superiority, both models exhibit a similar trend, revealing a notable concentration of the highest Grad-CAM values within the rooftop area. This trend reaches its peak at $k = 15$, suggesting a strong fixation of the models on the visual patterns inherent in the building rooftops. The clarity of this pattern provides confidence in asserting that the models are specifically focusing on the visual features within the buildings' rooftops, rather than being influenced by objects or regions surrounding the buildings.

Figure 9 shows a collage that highlights the specific regions within randomly selected original RGB images from the test set, where the highest activation values from the EfficientNetB0 Grad-CAM are concentrated: (A) the areas of the highest activation for the Grad-CAM corresponding to the *asbestos* class and, conversely, (B) for the *non-asbestos* class. These collages effectively illustrate the areas within the rooftop that the EfficientNetB0 model identifies as most critical in determining the presence or absence of asbestos. Qualitatively, we can observe distinct patterns in the classification process. For images classified as *asbestos*, the model identifies rooftops with a grayish hue and a sort of scratched texture. On the other hand, for images classified as *non-asbestos*, the model frequently highlights rooftops that exhibit a reddish or lighter color palette and smoother textures. This differ-

entiation in color and texture underscores the model's learned criteria for distinguishing between asbestos and non-asbestos rooftops.

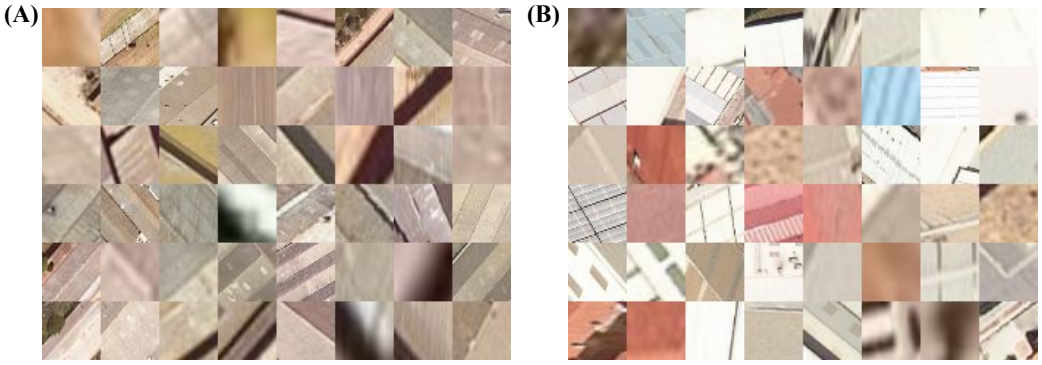

Figure 9. Collage of regions within randomly selected original RGB images from the test set, where the highest activation values from the EfficientNetB0 Grad-CAM are concentrated: (**A**) corresponds to images classified as *asbestos*. (**B**) corresponds to images classified as *non-asbestos*.

5. Discussion

The Deep Learning pipeline described in this work was conceived for identifying building rooftops with asbestos presence. The classification was performed through the processing of publicly available aerial images obtained from the ICGC. Such images were pre-processed, combining them with catastral data, so as to guarantee their suitability for subsequent classification. Remarkably, and unlike many previous studies on asbestos classification reported in the literature, the dataset in this study only contained RGB images, without data such as infrared, hyperspectral, or LiDAR.

A thoroughly curated dataset ensuring that the models take full advantage of the information present in high-resolution imagery was key to achieving high levels of accuracy, irrespective of the architecture used for classification (EfficientNetB0 and ResNet50). The outcomes from both classification processes underwent a double validation process, which consisted of a random test and a *k*-fold test. Both validations showed high levels of accuracy, with low numbers of false positives and false negatives, as seen in the corresponding confusion matrix.

Given the public health and even legal implications of the presence of asbestos in buildings, it is important to move beyond the deployment of black-box solutions—as it frequently occurs in AI applications. Seeking an interpretable outcome, we applied Grad-CAM as a visual means to interpret the output of the classification methods. In doing so, we obtained qualitative and quantitative evidence to ensure that successful classifications were not relying on spurious information. On the contrary, our results clearly point at the fact that the classification tasks learned the relevant variables to decide on the presence or absence of asbestos—those of the visual features of the material, as seen on aerial images: rugosity, color, texture, and so on.

Admittedly, this work leaves room for improvement and future research directions. To start with, the training dataset was not as general as one would wish. In the area under study (mostly the Barcelona province), asbestos was used for construction for at least three decades before it was banned, and in very diverse environments. The material has, thus, undergone different aging processes, which have resulted in notable differences, mostly in coloring and texture. In rural environments, it is common to see lichen-covered asbestos rooftops, which deviate significantly from the usual and expected grayish color. This explains why the classification task was extremely successful in industrial and urban contexts, but the performance decreased in rural areas. Such problems, which are hard to estimate quantitatively but are easily identified through visual inspection, call for a larger effort in the manual sampling of asbestos, with particular care for underrepresented areas.

Furthermore, we only considered EfficientNetB0 and ResNet50 architectures. Other choices (e.g., VGG [60]) could have been researched as well. However, the choice of these architectures is grounded in the efficiency and performance attributes inherent in the ResNet and EfficientNet family of model architectures. EfficientNet models, known for achieving comparable performance with fewer parameters in classification benchmarks, emerged as a fitting choice. Meanwhile, ResNet models, with their skip connections architecture, have demonstrated robust performance and serve as a foundational model in various computer vision tasks. We prioritized simplicity in terms of model parameters, mainly because of our dataset's limited size, leading us to select ResNet50 and EfficientNetB0 as they represent the simplest models within their respective groups in terms of parameters.

Another possible research direction might be moving from mere detection to asbestos area quantification. Such enhancement would imply the use of image segmentation instead of classification. We opted for image classification provided that our target was focused on the elaboration of a census of presence/absence of asbestos. On top of that, the usage of segmentation increased the complexity of the task. Classification tasks employ simpler loss functions such as cross-entropy, which compares the predicted label against the true label for each image. This is straightforward when dealing with binary outcomes per rooftop. In contrast, segmentation requires a more complex computation, typically involving loss functions that assess accuracy in the aggregation of pixel-wise predictions for each image. This means that the loss calculation would involve $N \times M$ (image size) comparisons, where each predicted pixel label is compared against the corresponding true pixel label across each entire image. Although segmentation may offer detailed insights into asbestos distribution in rooftops, the increased complexity and potential for higher error rates made classification the more viable option for our scope. So, we acknowledge segmentation's value for future exploration.

Finally, the efforts towards explainability could be enhanced in different ways. One possibility would be the addition of an image segmentation task to the pipeline. As it is now, we have a binary mask (*rooftop*, *off-rooftop*) directly from cadastral data. Expanding these categories via automatic segmentation, the explainability analysis would go beyond the binary quantification of Grad-CAM values to a collection of labels as identified in the image. In doing so, secondary elements enhancing or damaging the prediction quality could be identified.

The pipeline proposed in this study can guide the process of an automatic census of asbestos presence. This is of particular importance both in large cities and smaller towns, given the human and economic cost of manual monitoring to meet the legal mandates that emanate from European and national directives. In Spain, for example, municipalities should have created a census of facilities and locations with asbestos, along with a scheduled plan for their removal, by April 2023. In many cases, such mandate is still to be fulfilled. Along this line, an asbestos-detection pipeline should integrate the temporal dimension in future research endeavors. As the census of facilities with asbestos is already challenging, tracking the compliance of asbestos rooftop removal over time emerges as a critical consideration.

6. Conclusions

In this work, we proposed a competitive Deep Learning pipeline for the identification of asbestos materials on rooftops. The term "competitive" has two implications in our work: on the one hand, the task was performed at a low cost, relying exclusively on RGB high-resolution aerial images, which are publicly available in most European and worldwide countries from national and regional geographic institutes. As opposed to methods dependent on infrared or hyperspectral imagery, our choice ensures our methodology's versatility and applicability. On the other hand, the classification task was performed on two state-of-the-art models, with very high accuracies in both cases—although showing that EfficientNetB0 is superior to ResNet50 in this particular task. Such good results in two different architectures highlight the importance of the data curation process, including its

representativeness in industrial, urban, and peri-urban/rural contexts, and the use of the cadastre–most often available as open data—to ensure that the model's learning process is centered and taking full advantage of the images resolution.

Beyond the classification performance itself, the application of Grad-CAM to support the models' explainability reveals that, indeed, both EfficientNetB0 and ResNet50 classifiers extract the most useful information from rooftops themselves, i.e., their color, texture, shape, etc., and not any other spurious surrounding elements which could deceptively increase or decrease the accuracy of the proposed methods.

Author Contributions: Conceptualization, A.L. and J.B.-H.; methodology, C.B., A.L. and J.B.-H.; software, D.O., A.G.-G. and K.M.-F.; validation, A.G.-G., C.B. and A.L.; data collection, C.S. (Carles Scotto) and C.S. (César Sánchez); data curation, D.O.; writing—original draft preparation, all authors; writing—review and editing, C.B., A.L. and J.B.-H.; visualization, D.O., A.G.-G. and C.B.; supervision, C.B., A.L. and J.B.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by PID2021-128966NB-I00 (JBH) and PID2022-138721NB-I00 (to AL) grants from the Spanish Ministry of Science, Research National Agency and FEDER (UE). JBH acknowledges financial support from the Ramón y Cajal program through the grant RYC2020-030609-I. DO and CB were supported by a PhD grant from the Universitat Oberta de Catalunya (UOC).

Data Availability Statement: Aerial images and cadastre information are available from ICGC and the INSPIRE Services of Cadastral Cartography, respectively. The manually-sampled asbestos data presented in this study are available on request from the corresponding author due to commercial nature. Requests to access asbestos location data should be directed to detectamiant@gmail.com, DetectA.

Conflicts of Interest: Authors C. Scotto and C. Sánchez were employed by the company DetectA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EU	European Union
DL	Deep Learning
CAM	Class Activation Mapping
Grad-CAM	Gradient-weighted Class Activation Mapping
CNN	Convolutional Neural Networks
HSI	HyperSpectral Imagery
ICGC	Catalan Cartographic and Geologic Institute

References

1. Nielsen, L.S.; Baelum, J.; Rasmussen, J.; Dahl, S.; Olsen, K.E.; Albin, M.; Hansen, N.C.; Sherson, D. Occupational asbestos exposure and lung cancer—A systematic review of the literature. *Arch. Environ. Occup. Health* **2014**, *69*, 191–206. [CrossRef]
2. Abbasi, M.; Mostafa, S.; Vieira, A.S.; Patorniti, N.; Stewart, R.A. Mapping Roofing with Asbestos-Containing Material by Using Remote Sensing Imagery and Machine Learning-Based Image Classification: A State-of-the-Art Review. *Sustainability* **2022**, *14*, 8068. [CrossRef]
3. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
4. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [CrossRef]
5. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 93. [CrossRef]
6. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.

7. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10705–10714.
8. Wagner, J.; Kohler, J.M.; Gindele, T.; Hetzel, L.; Wiedemer, J.T.; Behnke, S. Interpretable and fine-grained visual explanations for convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9097–9107.
9. Desai, S.; Ramaswamy, H.G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 972–980.
10. Patro, B.N.; Lunayach, M.; Patel, S.; Namboodiri, V.P. U-cam: Visual explanation using uncertainty based class activation maps. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7444–7453.
11. Bustos, C.; Rhoads, D.; Solé-Ribalta, A.; Masip, D.; Arenas, A.; Lapedriza, A.; Borge-Holthoefer, J. Explainable, automated urban interventions to improve pedestrian and vehicle safety. *Transp. Res. Part Emerg. Technol.* **2021**, *125*, 103018. [CrossRef]
12. Charuchinda, P.; Kasetkasem, T.; Kumazawa, I.; Chanwimaluang, T. On the use of class activation map for land cover mapping. In Proceedings of the 2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Pattaya, Thailand, 10–13 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 653–656.
13. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
16. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 3735–3756. [CrossRef]
17. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
18. Kumari, M.; Kaul, A. Deep learning techniques for remote sensing image scene classification: A comprehensive review, current challenges, and future directions. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7733. [CrossRef]
19. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
20. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote. Sens.* **2020**, *12*, 1688. [CrossRef]
21. Men, G.; He, G.; Wang, G. Concatenated Residual Attention UNet for Semantic Segmentation of Urban Green Space. *Forests* **2021**, *12*, 1441. [CrossRef]
22. Kabisch, N.; Strohbach, M.; Haase, D.; Kronenberg, J. Urban green space availability in European cities. *Ecol. Indic.* **2016**, *70*, 586–596. [CrossRef]
23. Wolch, J.R.; Byrne, J.; Newell, J.P. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landsc. Urban Plan.* **2014**, *125*, 234–244. [CrossRef]
24. Ramoelo, A.; Cho, M.A.; Mathieu, R.; Madonsela, S.; Van De Kerchove, R.; Kaszta, Z.; Wolff, E. Monitoring grass nutrients and biomass as indicators of rangeland quality and quantity using random forest modelling and WorldView-2 data. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *43*, 43–54. [CrossRef]
25. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8311–8320.
26. Omarzadeh, D.; Karimzadeh, S.; Matsuoka, M.; Feizizadeh, B. Earthquake Aftermath from Very High-Resolution WorldView-2 Image and Semi-Automated Object-Based Image Analysis (Case Study: Kermanshah, Sarpol-e Zahab, Iran). *Remote. Sens.* **2021**, *13*, 4272. [CrossRef]
27. Bastani, F.; He, S.; Abbar, S.; Alizadeh, M.; Balakrishnan, H.; Chawla, S.; Madden, S.; DeWitt, D. Roadtracer: Automatic extraction of road networks from aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4720–4728.
28. Hosseini, M.; Sevtsuk, A.; Miranda, F.; Cesar, R.M., Jr.; Silva, C.T. Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery. *Comput. Environ. Urban Syst.* **2023**, *101*, 101950. [CrossRef]
29. Izquierdo, S.; Rodrigues, M.; Fueyo, N. A method for estimating the geographical distribution of the available roof surface area for large-scale photovoltaic energy-potential evaluations. *Sol. Energy* **2008**, *82*, 929–939. [CrossRef]
30. Mainzer, K.; Fath, K.; McKenna, R.; Stengel, J.; Fichtner, W.; Schultmann, F. A high-resolution determination of the technical potential for residential-roof-mounted photovoltaic systems in Germany. *Sol. Energy* **2014**, *105*, 715–731. [CrossRef]
31. Schallenberg-Rodríguez, J. Photovoltaic techno-economical potential on roofs in regions and islands: The case of the Canary Islands. Methodological review and methodology proposal. *Renew. Sustain. Energy Rev.* **2013**, *20*, 219–239. [CrossRef]

32. Szabó, S.; Burai, P.; Kovács, Z.; Szabó, G.; Kerényi, A.; Fazekas, I.; Paládi, M.; Buday, T.; Szabó, G. Testing algorithms for the identification of asbestos roofing based on hyperspectral data. *Environ. Eng. Manag. J.* **2014**, *143*, 2875–2880. [CrossRef]
33. Cilia, C.; Panigada, C.; Rossini, M.; Candiani, G.; Pepe, M.; Colombo, R. Mapping of asbestos cement roofs and their weathering status using hyperspectral aerial images. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 928–941. [CrossRef]
34. Kruse, F.A.; Lefkoff, A.; Boardman, J.; Heidebrecht, K.; Shapiro, A.; Barloon, P.; Goetz, A. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote. Sens. Environ.* **1993**, *44*, 145–163. [CrossRef]
35. Krówczyńska, M.; Raczko, E.; Staniszewska, N.; Wilk, E. Asbestos—cement roofing identification using remote sensing and convolutional neural networks (CNNs). *Remote. Sens.* **2020**, *12*, 408. [CrossRef]
36. Raczko, E.; Krówczyńska, M.; Wilk, E. Asbestos roofing recognition by use of convolutional neural networks and high-resolution aerial imagery. Testing different scenarios. *Build. Environ.* **2022**, *217*, 109092. [CrossRef]
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
38. Kaplan, G.; Gašparović, M.; Kaplan, O.; Adjiski, V.; Comert, R.; Mobariz, M.A. Machine learning-based classification of asbestos-containing roofs using airborne RGB and thermal imagery. *Sustainability* **2023**, *15*, 6067. [CrossRef]
39. Baek, S.C.; Lee, K.H.; Kim, I.H.; Seo, D.M.; Park, K. Construction of Asbestos Slate Deep-Learning Training-Data Model Based on Drone Images. *Sensors* **2023**, *23*, 8021. [CrossRef] [PubMed]
40. Hikuwai, M.V.; Patorniti, N.; Vieira, A.S.; Frangioudakis Khatib, G.; Stewart, R.A. Artificial Intelligence for the Detection of Asbestos Cement Roofing: An Investigation of Multi-Spectral Satellite Imagery and High-Resolution Aerial Imagery. *Sustainability* **2023**, *15*, 4276. [CrossRef]
41. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote. Sens.* **2019**, *152*, 166–177. [CrossRef]
42. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote. Sens.* **2018**, *56*, 2811–2821. [CrossRef]
43. Zhang, C.; Yue, P.; Tapete, D.; Shangguan, B.; Wang, M.; Wu, Z. A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *88*, 102086. [CrossRef]
44. Kakogeorgiou, I.; Karantzalos, K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102520. [CrossRef]
45. Zhang, Y.; Hong, D.; McClement, D.; Oladosu, O.; Pridham, G.; Slaney, G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J. Neurosci. Methods* **2021**, *353*, 109098. [CrossRef] [PubMed]
46. Shi, C.; Zhang, X.; Sun, J.; Wang, L. Remote sensing scene image classification based on dense fusion of multi-level features. *Remote. Sens.* **2021**, *13*, 4379. [CrossRef]
47. Chen, S.B.; Wei, Q.S.; Wang, W.Z.; Tang, J.; Luo, B.; Wang, Z.Y. Remote sensing scene classification via multi-branch local attention network. *IEEE Trans. Image Process.* **2021**, *31*, 99–109. [CrossRef] [PubMed]
48. Li, X.; Shi, D.; Diao, X.; Xu, H. SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 5801112. [CrossRef]
49. Li, Y.; Zhang, Y.; Huang, X.; Yuille, A.L. Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *146*, 182–196. [CrossRef]
50. Huang, X.; Sun, Y.; Feng, S.; Ye, Y.; Li, X. Better visual interpretation for remote sensing scene classification. *IEEE Geosci. Remote. Sens. Lett.* **2021**, *19*, 6504305. [CrossRef]
51. Guo, X.; Hou, B.; Wu, Z.; Ren, B.; Wang, S.; Jiao, L. Prob-POS: A Framework for Improving Visual Explanations from Convolutional Neural Networks for Remote Sensing Image Classification. *Remote. Sens.* **2022**, *14*, 3042. [CrossRef]
52. Song, W.; Dai, S.; Wang, J.; Huang, D.; Liotta, A.; Di Fatta, G. Bi-gradient verification for grad-CAM towards accurate visual explanation for remote sensing images. In Proceedings of the 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 8–11 November 2019; IEEE Piscataway, NJ, USA, 2019; pp. 473–479.
53. Dutta, S.; Das, M.; Maulik, U. Towards Causality-Based Explanation of Aerial Scene Classifiers. *IEEE Geosci. Remote. Sens. Lett.* **2023**, *21*, 8000405. [CrossRef]
54. Fu, K.; Dai, W.; Zhang, Y.; Wang, Z.; Yan, M.; Sun, X. Multicam: Multiple class activation mapping for aircraft recognition in remote sensing images. *Remote. Sens.* **2019**, *11*, 544. [CrossRef]
55. Li, Z.; Zhang, X.; Xiao, P.; Zheng, Z. On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 3266–3281. [CrossRef]
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016, Springer: Berlin/Heidelberg, Germany, , 2016; pp. 630–645.
57. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

58. Chai, J.; Zeng, H.; Li, A.; Ngai, E.W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Mach. Learn. Appl.* **2021**, *6*, 100134. [CrossRef]
59. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE Piscataway, NJ, USA, 2009; pp. 248–255.
60. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.