(Article begins on next page)

Spotting insects from satellites: modeling the presence of *Culicoides imicola* through Deep CNNs

Stefano Vincenzi*, Angelo Porrello*, Pietro Buzzega*, Annamaria Conte†, Carla Ippoliti†, Luca Candeloro†, Alessio Di Lorenzo†, Andrea Capobianco Dondona†, Simone Calderara*

*AImageLab, University of Modena and Reggio Emilia, Modena, Italy
†Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise 'G.Caporale', Teramo, Italy

Abstract—Nowadays, Vector-Borne Diseases (VBDs) raise a severe threat for public health, accounting for a considerable amount of human illnesses. Recently, several surveillance plans have been put in place for limiting the spread of such diseases, typically involving on-field measurements. Such a systematic and effective plan still misses, due to the high costs and efforts required for implementing it. Ideally, any attempt in this field should consider the triangle vectors-host-pathogen, which is strictly linked to the environmental and climatic conditions. In this paper, we exploit satellite imagery from Sentinel-2 mission, as we believe they encode the environmental factors responsible for the vector's spread. Our analysis - conducted in a data-driver fashion - couples spectral images with ground-truth information on the abundance of *Culicoides imicola*. In this respect, we frame our task as a binary classification problem, underpinning Convolutional Neural Networks (CNNs) as being able to learn useful representation from multi-band images. Additionally, we provide a multi-instance variant, aimed at extracting temporal patterns from a short sequence of spectral images. Experiments show promising results, providing the foundations for novel supportive tools, which could depict where surveillance and prevention measures could be prioritized.

Index Terms—Remote-sensing, vector-borne diseases, entomological surveillance, *Culicoides imicola*, bluetongue, Infectious disease, satellite imagery, Sentinels, Deep Learning, Convolutional Neural Networks, Europe

I. INTRODUCTION

Recent incursions into Europe of some vector-borne diseases (Chikungunya and Malaria), the persistence and spread of others (West Nile Disease, Bluetongue and Lyme disease), as well as the risk of introduction of new exotic diseases (Rift Valley Fever, Crimean Congo Hemorrhagic Fever, African Horse Sickness) prompt an integrated surveillance action. On the one hand, such a solution may identify both the presence and the abundance of the vectors (e.g., ticks, culicoides, mosquitoes); on the other hand, it may lead to a deeper knowledge of the territory and the potential habitats of different arthropod species. Over the years, public institutions and organizations implemented various surveillance plans for monitoring the spread of vectors in Europe [1]. However, a systematic and intensive plan would require extremely high resources and funds. In this respect, the identification of specific monitoring areas could bring to a cost reduction and fine-grained entomological surveillance strategies.

The survival and spread of vectors and pathogens in a given area rely on several factors, such as climate, vegetation, landscape composition, and hosts' availability. In the recent past, various studies relate the presence and the abundance of vectors with different climatic and environmental factors. As an example, in [2], the authors defined a series of liable variables, such as temperature, aridity index, elevation, and vegetation index. The analysis of these factors capture the distribution of different species of *Culicoides* and the environment favorable to them. Furthermore, the authors of [3] pointed out the climate change as responsible for the spread of certain pathogens, which in turn are involved in the spread of the above-mentioned diseases.

To sum up, the environmental and climatic conditions represent a crucial aspect for modeling the distribution and abundance of vectors. Nowadays, dozens of satellites provide images at very high resolutions and with unprecedented rhythms. In this work, we explore the analysis of these images for predicting the presence of the *Culicoides imicola*, the latter being among the species responsible for the Bluetongue disease. In more details, we conduct all the analyses on the satellite imagery coming from the Sentinel 2 mission, handled by the European Space Agency (ESA) [4], [5] within the Copernicus program. Copernicus builds on a global network of thousands of sensors for land, ocean and atmospheric monitoring of Earth, resulting in an impressive number of daily observations. The Sentinel-2A and Sentinel-2B satellites are equipped with a multi-spectral device (MSI) capable of acquiring 13 spectral bands, ranging from visible (RGB) to short-wave infrared (SWIR), including near-infrared (NIR).

These multi-band images could provide an enormous amount of information about the landscape composition, fragmentation and configuration, which relate to the presence and abundance of vectors. Aiming to capture these kind of relationships among data, we gathered a new dataset, which pairs spectral bands coming from the Sentinel mission with ground-truth measures of presence/absence for the *Culicoides imicola* in a certain area.

However, analyzing these data with shallow machine learning models would require intense efforts during the features extraction stage, which could be enhanced using deep learning techniques. These techniques have already been used in various researches related to the remote sensing field, such as soil and crop classification. In our work, we frame the problem of estimating the presence of *C. imicola* as a binary

classification task, and take advantage of Deep Convolutional Neural Networks (DCNNs) [6], which provide meaningful and hierarchical representations from multi-band images. To the best of our knowledge, their application in the field of vector-borne diseases (VBDs) is still very limited.

We evaluate two kinds of approaches: on the one side, we adopt a single multi-band image for predicting the given binary target; on the other side, we underpin a sliding window approach over time, where the network takes as input a sequence of multi-band images. Those are acquired during the month preceding the field catch of the vector: this way, we aim to assess whether a temporal series of inputs could be more informative. In the above-mentioned approaches, we rely on ResNet [7] as a backbone network for the feature extraction stage, coupled with an aggregation module for merging representations within the input window. For assessing the merits of different inputs, we test our solution with: i) RGB channels ii) the NDVI and NDWI indices; iii) all the 13 raw bands captured by the Sentinel satellites. Briefly, experiments show two insightful remarks: firstly, the spectral bands encode important features for the task at hand; secondly, the multi-instance approach outperforms the single-instance one.

The manuscript is organized as follows: related works are discussed in Section II. We describe the datasets we used in Section III. In Section IV, we introduce the proposed methods. We report the experiments and ablation studies we conducted in Section V. We finally draw several conclusions in Section VI.

II. RELATED WORKS

A. Vector-borne diseases and machine learning approaches

Recently, several studies exploit machine learning algorithms for analyzing the spread of VBDs in a given area. In this respect, the authors of [8] took advantage of eco-regions, namely areas' within which there are associations of interacting biotic and abiotic features' [9]. The underlying idea relies on the observation that areas with similar conditions are potentially prone to similar disease risk, even though they are distant from each other. Basically, they proposed a map of eco-climatic Italian regions, the latter identified through a data-driven approach. Formally, their clustering method relies on seven environmental and climatic variables, relevant for the spread of various VBDs. Afterward, they related the resulting clustering map with ground-truth information about the Bluetongue vectors and West Nile Disease (WND) outbreaks in Italy. They showed that the habitat of *C. imicola* was represented by two out of the twenty-two eco-regions, thus indicating high-risk areas and therefore where surveillance measures could be prioritized.

Differently, [10] predicted both the presence and abundance of five different vectors (*C. imicola*, *C. newsteadii*, *Pulicaris complex*, *C. punctatus* and *C. obsoletus*) across the Spanish territory. The authors pointed out the abundance as an important factor, essential for mathematical models for estimating both the potential of the transmitted disease and its future behaviour. Previous approaches considered the abundance linearly related

to the probability of presence. Instead, they compute the former by combining the predicted probability with other features, such as temperature or precipitations. To this aim, they used various dataset (e.g. GTOPO30, CORINE etc.) providing meteorological and environmental features (e.g. land cover, vegetation, ground water etc.), mainly derived from satellite imagery. Similarly to us, they gather ground-truth information from traps located near the farms.

In technical details, they fed these data to an ensemble model, exploiting a classification and regression forest for estimating the presence and the abundance respectively. Results show particular factors being fundamental for the spread of specific vectors. In particular, focusing on the *C. imicola*, the authors found summer rainfall and dryness as the most significant factors for its distribution.

We consider these works as a starting point for our efforts. However, we highlight several methodological differences: i) while they identified *a priori* risk areas, we conduct a *a posteriori* analysis on top of them; ii) we shift the paradigm from an unsupervised to a supervised one, modeling the presence of a vector as a binary classification problem; iii) we do not rely on hand-crafted features, as we leverage satellite images for predicting the target directly, allowing the convolutional model to learn the proper features for the task at hand.

B. Crop and Land classification

In the remote sensing field, numerous researches focus on crop and land cover classification. A widely used approach relies on various indices, such as the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI). These two indices arise directly from raw spectral bands and reflect a clear physical meaning: while the former quantifies the green biomass, the latter depicts the water-body mapping. However, they show some limitations: as an example, the denser the vegetation coverage becomes, the more they will saturate, and the more performance will degrade [11].

In such a case, the Enhanced Vegetation Index (EVI) represents a valid alternative: [12] used the EVI time-series as input for a deep learning architecture, aiming to classify summer crops (i.e., Rice, Corn, Safflower, etc.). In this respect, the motivations behind the adoption of time-series emerge from seasonal patterns in the dynamics of vegetation. Moreover, the authors showed that traditional machine learning approaches (XGBoost, Random Forest, and Support Vector Machine) lead to lower performance with respect to the proposed deep model, which exploits one-dimensional convolutions for capturing both seasonal patterns and small scale temporal variations.

Differently, the authors of [13] described a three-dimensional Convolutional Neural Network (3D-CNN), inferring crops categories (i.e. Corn, Rice, Soyb and Tree) from satellite images acquired by the Gaofen-2 satellite [14]. Since vegetation indices rely on few bands among the available ones, some crucial information could be discarded, thus resulting in lower generalization capabilities in the crop classification task. On this latter point, the authors propose the 3D-CNN, as it can
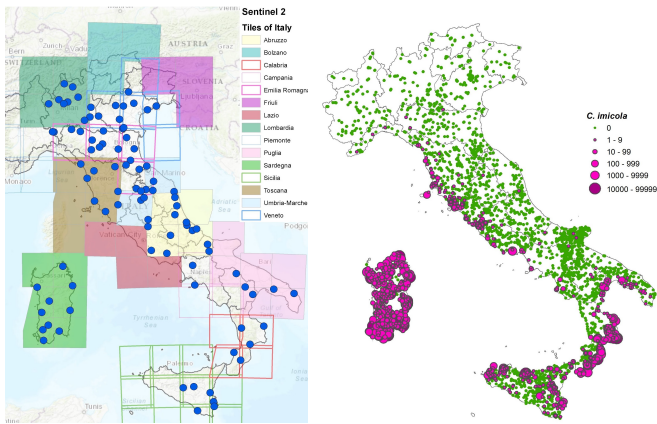
Fig. 1. Left: Italian locations of 94 capture farms (2018). Right: spatial distribution of the *C. imicola*, ranging from August 2000 to August 2017 [8].

	Bands	Wavelength (μm)	Res (m)
B_1	Coastal Aerosol	0.443	60
$B_{2,3,4}$	BGR channels	0.490	10
B_5	Vegetation Red Edge	0.705	20
B_6	Vegetation Red Edge	0.740	20
B_7	Vegetation Red Edge	0.783	20
B_8	NIR	0.842	10
B_{8A}	Vegetation Red Edge	0.865	20
B_9	Water Vapour	0.945	60
B_{10}	SWIR (Cirrus)	1.375	60
B_{11}	SWIR	1.610	20
B_{12}	SWIR	2.190	20

look for temporal patterns and dynamics within the raw bands sequence directly.

Similarly, the authors of [15] proposed a land cover classification method, with four different classes (crop, tree, water, and road), underpinning Sentinel bands as input to a Support Vector Machine (SVM) classifier. In this respect, they conduct experiments on three different types of features: vegetation indices, hand-selected bands, and, finally, all bands. As observed in [13], the authors claimed that exploiting all bands leads to better results.

III. DATASETS

A. Bluetongue vectors Dataset

Aiming to predict the presence of *C. imicola*, we gathered a dataset coupling satellite patches depicting the Italian territory with ground-truth information regarding vectors' catches. In more details, we extracted the data regarding the presence/absence of the *C. imicola* species from the 'Culicoides collections'. Such information have been stored in a centralized database, hosted by the IZSAM Institute[1] as part of the Bluetongue National Surveillance Plan [16]. This plan takes place in Italy since 2000 [17], and defines field and laboratory protocols for the collection of *Culicoides* and their identification. The catches of these insects are conducted near the farms through various traps, activated in the evening and collected the next morning. Eventually, an entomological laboratory examines the traps for identifying the *Culicoides* species and their actual amount.

B. Satellite Imagery

As mentioned before, the multi-spectral images we rely on are acquired through optical devices onboard the Sentinel-2A and 2B satellites. These images depict the Earth's surface with a high spatial, temporal, and spectral resolutions.

The two satellites simultaneously orbit around the Earth on the same trajectory, at the height of about 780 km, staggered between them by 180 degrees. They provide a multispectral

"photograph" of the territory that they fly over every five days. The spectral bands (see Table I) are acquired under different resolutions, in particular at 10, 20 and 60 meters per pixel. The ESA makes these images available through an online platform [2].

In the following list, we sum up the main characteristics for each band, arranged in ascending order with respect to their wavelength:

- B_1: it identifies any type of aerosol;
- B_2: the blue color, useful to discriminate between the different types of vegetation. Compared to lower frequencies, it provides a better clear water penetration and a superior illumination of the materials in shadow;
- B_3: the green color, which excels in highlighting muddy water, and helps in finding oil on the water surface and vegetation;
- B_4: the red color, suitable for finding urban and soil features; its good response to dead foliage allows to discriminate between the vegetation types;
- $B_{5,6,7,8A,8}$: all these bands are designed specifically for analyzing and classifying the vegetation. In particular, the last one enables shorelines mapping and extracting information about biomass content;
- $B_{9,10}$: belonging to the infra-red region, they both aim at identifying water vapour. Notably, the second one is specific for cirrus clouds;
- $B_{11,12}$: the last two bands distinguish areas covered by ice and snow from clouds.

C. Datasets integration

We combined the ground-truth information regarding *C. imicola* catches with the ones coming from Sentinel-2 satellites. With no loss of generality, we restrict our analysis to the year 2018, as Sentinel-2B data have been made available since March 2017 only.

For each capture, we collect the corresponding satellite image depicting its site location. To avoid misleading associations

---

[1]Namely, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise.

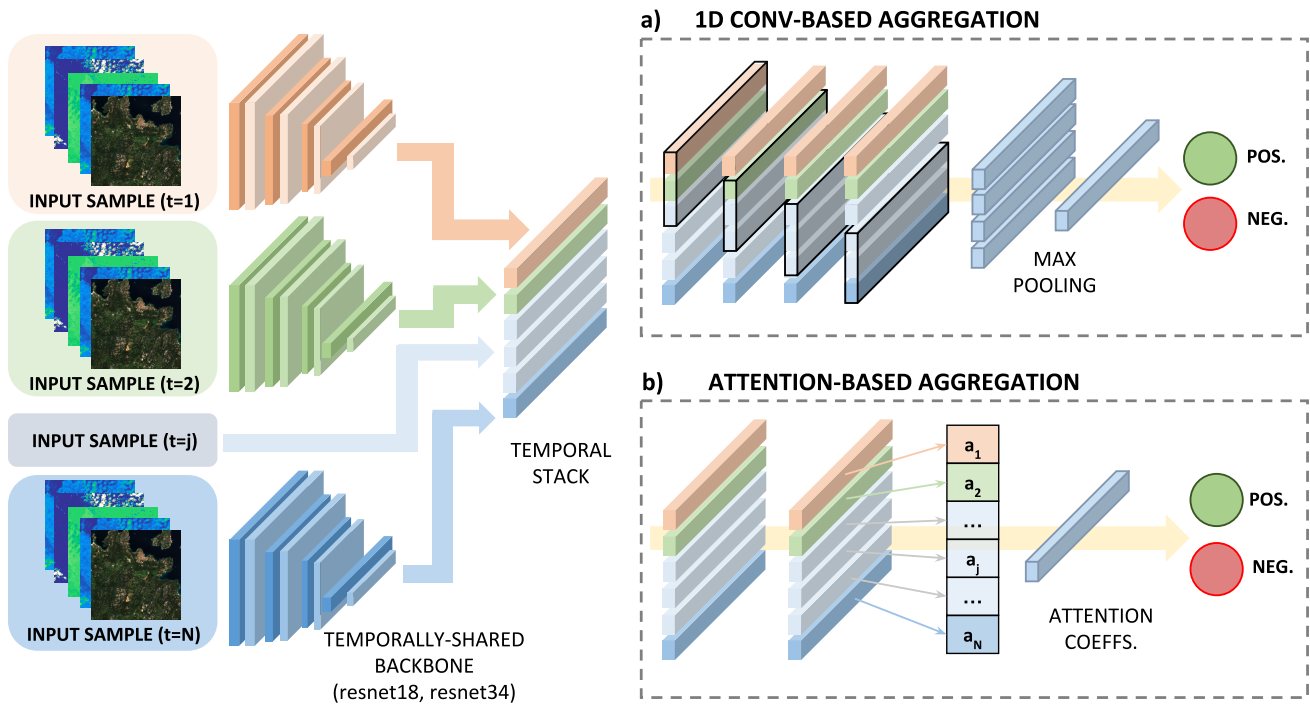[2]Please refer to https://scihub.copernicus.eu for additional details.

Fig. 2. The proposed framework, in its multi-instance variant: it takes multiple samples as input, each of which is independently processed by a shared CNN, the latter acting as the feature extraprinter (e.g. resnet18). Afterward, we stack those representations and form the input for the next neural aggregation module, which fulfills one of the following two options: (a) a 1D-Convolution based schema, leveraging the underlying time ordering; (b) an attention-based module, which estimates the importance of each representation during the overall computation. Best viewed in color.

Regions	Farms	Catches	Images	Regions	Farms	Catches	Images
Abruzzo	6	196	491	Molise	1	26	69
Basilicata	2	102	142	Apulia	4	196	280
Calabria	5	192	495	Sardinia	10	419	778
E. Romagna	13	489	999	Sicily	6	188	423
Lazio	3	112	216	Tuscany	8	374	702
Lombardy	13	431	984	Veneto	10	384	792
Marche	5	229	429	Umbria	6	262	503

between input and targets, we focus on farms with an accurate georeferencing only. Importantly, we create each example so that the image's acquisition date is as close as possible to the time at which the trap was placed. Among the different spatial resolutions available, we adopt 20 meters as the default input resolution. We empirically observe that it represents a good compromise between preserving images' details and the overall memory footprint. This way, for each capture site, we obtain a 224x224 pixel image (resulting in a cutout of 4480x4480 m), which captures environmental information near the farm. These catches take place periodically, at variable intervals.

Given a capture, the presence or absence of the *C. imicola* represents the ground truth we adopt during our experiments. The dataset is unbalanced in favor of the absence of the vector. On the one hand, the positive catches constitute approximately one-seventh of the total. On the other hand, the distribution of

positive labels is not uniform among different regions: in this respect, Figure 1 (right) shows the distribution of the catches over the Italian territory, as well as the tiles of the satellite images (left).

To sum up, the final dimension of the dataset equals the number of catches, in our case 3671, distributed among 94 farms. For our purposes, we leverage 7514 satellite images. As mentioned before, 3182 samples refer to the negative class, thus indicating the absence of the *C. imicola*, while 489 samples belong to the positive one. We propose in Table II a region-wise summary regarding the dataset.

IV. MODELS

A. Features Extraction

Our proposal heavily relies on ResNet [7] as feature extractor for the overall architecture. In more details, ResNet is one of the most used networks for classification and object detection tasks. The underlying idea - which makes this architecture so powerful - regards the residual blocks, which aim to simplify the training of neural networks characterized by many layers. In this regards, several researches [18], [19] highlight networks' depth as a crucial factor: indeed, deeper models allow to learn richer representations. On the other hand, the deeper the networks become, the more the vanishing gradient issue [20] emerges. A careful initialization has addressed this problem primarily [20], as well as placing intermediate normalization layers [21]. Differently, ResNet and its variants tackle the degradation problem by means of skip connections,

as the network is equipped with an identity mapping between each block and the following one. This way, the gradients can be back-propagated without suffering from numerical issues. Moreover, identity shortcut connections add neither additional parameters nor computational complexity.

Since we rely on a few labeled data, our proposal exploits a ResNet pre-trained model on the Imagenet dataset [22], and then fine-tune it on ours. Since training a model from scratch could be hard, a standard guideline envisions a pre-training stage on a large dataset. Therefore, one could take such an achieved optimum as an initialization point for the next optimization phase, the latter guided by the task at hand.

B. Attention-based Features Aggregation

As we rely - for both the training and the inference stages - on multiple images at a time, we designed two solutions for aggregating representations coming from each sample. Let us start from the first one: we discard information arising from the temporal order and focus on features of each image solely.

Such an approach is based on the work described in [23], where the authors proposed a novel neural module for handling Multiple Instance Learning (MIL) problems. MIL represents a variation of the supervised learning paradigm, where a single label refers to a bag of instances. Considering $X = \{x_1, x_2, \ldots, x_K\}$ as a bag of samples exhibiting neither dependency nor ordering among each other, the labels of each individual sample are $y_1, y_2, \ldots, y_K$, with $y_k \in \{0, 1\}$; however, they remain unknown during the training phase. In this respect, the assumption underlining the MIL framework can be summarized as follows:

$$Y = \begin{cases} 0, & \text{iff } \sum_k y_k = 0, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

The authors' proposal consists in training a MIL model by optimizing the log-likelihood function. Importantly, the probability assigned to a given bag must be permutation invariant: this way, the solution to the MIL problem arises naturally from the Fundamental Theorem of Symmetric Functions [24].

We inherit from [23] a general three-steps approach for classifying a bag of instances: i) to apply a non-linear transformation to instances trough a function f; ii) to exploit a symmetric function σ to aggregate these features; iii) to transform such a comprehensive feature vector again by means of a function g. In our work, we adopt the embedding level approach for designing the function f, the latter mapping instances to a low dimensional embedding through a deep neural network. This way, the whole approach results flexible and trainable end-to-end. For the second step - the one implementing the aggregation phase - we rely on the attention mechanism. In this regards, we compute a weighted average of instances' latent representations, where a multi-layered perceptron outputs the coefficients. Given a bag of k embeddings $H = \{h_1, \ldots, h_K\}$ and parameters W and V, we combine them as follows:

$$z = \sum_{k=1}^{K} a_k h_k, \quad (2)$$

where:

$$a_k = \frac{\exp\{W^\top \tanh\left(V h_k^\top\right)\}}{\sum_{j=1}^{K} \exp\{W^\top \tanh\left(V h_j^\top\right)\}}. \quad (3)$$

Our assumptions deviate from the MIL framework ones, where the bag would be labeled as positive whether at least one of its sample is positive (see Eq. 1). In our problem, differently, we evaluate how past images influence the prediction for the current one, regardless of their order.

C. Convolution-based Features Aggregation

As a second step, we aimed to assess whether the temporal information could bring meaningful features for the task at hand. In this regards, we compare the above-described proposal with a different one, which exploits the images' acquisition time. Starting from a query image, this alternative takes as input five images from the temporal neighborhood of the query, restricting to the ones acquired before. We then arrange the images according to their temporal order. Afterward, the ResNet backbone independently processes each sample within the sequence, so resulting in a list of 512-dimensional embeddings. We then apply a one-dimensional convolution, followed by a max-pooling stage, on the temporal axis (see Fig. 2 a)). Eventually, a linear layer outputs the binary predictions (the same holds for the attention-based architecture).

During the preliminary stage, we intentionally discard recurrent networks for designing such a time-aware approach. Indeed, we would not benefit from the capabilities of these networks, as we focus on short and fixed-length input sequence. In support of this, [25] presented an empirical evaluation of convolutional and recurrent architectures. From their remarks, convolution arises as a natural starting point for sequence modeling problems; this is due to its low memory requirements, its stable gradient and its suitability for parallelism. Eventually, as shown in [12], the one-dimensional convolutions are able to extract features from satellite imagery, thus being viable for classifying crops.

V. EXPERIMENTS

A. Evaluation protocol

For avoiding overfitting and misleading conclusions, we adopt the Stratified K-Fold cross-validation to evaluate the performance of our models (with K = 5). Stratified K-Fold is defined as a simple variation of the K-fold, in a way that each fold preserves approximately the probability distribution of the classes (the one that can be computed on the complete dataset). Moreover, before splitting the dataset, we make sure each fold holds samples from all regions under consideration; thus guaranteeing the model to inspect images with different environmental and vegetative characteristics.

B. Metrics

We propose several metrics to evaluate the performance among various scenarios and architectures. Since the dataset at hand is actually unbalanced, an high accuracy value does not represent a symptom of satisfying performance nor generalization capabilities. Differently, metrics as precision and recall are suitable, as well as the F1-score, the latter rewarding solutions showing high values for both of the former. Eventually, we assess the models computing the Area Under the Precision-Recall Curve (AUPRC) for the positive class, thus obtaining a threshold-invariant summary.

C. Implementation details

Before being fed into the network, we normalize each spectral band independently, by computing the 2nd and 98th percentile values. We use these ones as boundaries for clipping each pixel. We empirically found that percentiles gain robustness when compared to the minimum and maximum, which suffer in presence of noisy acquisitions. After clipping the pixels' values accordingly, we apply the min-max normalization technique [26].

During each experiment, we exploit the pre-trained weights for finetuning ResNet. We apply data augmentation to the input images, as horizontal flip, vertical flip, and rotation. In our experiments, we train the model for 150 epochs, setting the learning rate fixed at 0.001. The batch size equals to 16. For regularization purpose, we apply the dropout technique, with a drop probability of 0.2.

For mitigating the above-discussed classes unbalance, we exploit the weighted counterpart of the cross-entropy loss. In so doing, the weight of the positive class is seven times greater the weight of the negative one.

D. Baselines

Firstly, we consider as baseline classifier the one implementing a naive strategy, as generating random predictions by sampling from the class distribution of the training set. Indeed, such a naive choice leads to a measure of 'baseline' performance, namely the expected success rate when simply guessing. This way, we are able to evaluate whether the model effectively learns from the data.

Secondly, we assess a single-instance binary classifier, equipped with resnet18 as feature extractor. Moreover, we tested it on different input settings, such as RGB, the Normalized Difference Vegetation Index (NDVI) coupled with the Normalized Difference Water Index (NDWI) (see Eq. 4), and all 13 bands. Moreover - as previously done in [15] - we test the B_4, B_{8A} and B_{11} bands, the ones involved during the NDVI and NDWI computations. The reasons behind these analyses are twofold: on the one hand, we can evaluate whether more acquisition will lead to significant improvements; on the other hand, we get a clear view on the contribution of spectral bands when compared to their hand-designed derivatives.

$$\text{NDVI} = \frac{B_{8A} - B_4}{B_{8A} + B_4}; \quad \text{NDWI} = \frac{B_{8A} - B_{11}}{B_{8A} + B_{11}}. \quad (4)$$

TABLE III
RESULTS OBTAINED IN THE SINGLE-INSTANCE SETTING.

Model [pre-trained yes/no]	Acc.	Pr.	Rc.	F1	AUPRC
Random classifier [✗]	.819	.169	.140	.153	.116
RGB [✗]	.922	.702	.584	.637	.606
RGB [✓]	.931	.764	.616	.679	.653
NDVI + NDWI [✗]	.926	.708	.640	.668	.656
NDVI + NDWI [✓]	.928	.714	.616	.659	.668
$B_4 + B_{8A} + B_{11}$ [✗]	.908	.612	.607	.609	.632
$B_4 + B_{8A} + B_{11}$ [✓]	.912	.697	.602	.643	.641
Spectral bands [✗]	.913	.619	.705	.656	.668
Spectral bands [✓]	.924	.667	.721	**.689**	**.682**

E. Single-instance Classification

We report in Table III the results for the single-instance setting. Looking at the first row, the one showing the random classifier's performance, we can draw the following considerations: firstly, the problem unbalance makes the accuracy unsuitable for judging the model's capabilities; secondly, as all variants achieve higher F1 and AUPRC scores, our model effectively learns useful patterns from the data.

Looking at Table III, we draw the following remarks:

- The raw bands lead to the highest performance for our problem. Such a result strengths the underlying intuition of our work, which exploits the environmental patterns arising from raw satellite images for predicting the presence of *C. imicola*;
- The NDVI and NDWI indices yield better results than the B_4, B_{8A} and B_{11} bands, namely the ones from which the former are computed: such a finding highlights their relevance for the field. However, it is worth noting that exploiting all the channels brings to even better results; thus indicating the occurrence of significant features ignored by the NDVI and NDWI;
- We remark the beneficial effects arising from the ImageNet pre-training: despite the ours being a completely different task, the convolutional weights can be inherited and fine-tuned for working on spectral bands. This may be useful when facing remote sensing applications, especially the ones characterised by few labeled data.

For additional notes on the single-istance setting, please refer to Subsection V-G .

F. Multi-instance Classification

In the multiple-instance setting we evaluate three variants: i) Attention-based (Section IV-B); ii) 1D-Convolution (Section IV-C); iii) a Mean-based approach, the latter computing the mean of ResNet's activations along the temporal dimension. We design such a baseline for investigating the impact of the proposed aggregation modules w.r.t. a naive one. Indeed, averaging the representations does not consider the samples'

TABLE IV
RESULTS OBTAINED IN THE MULTI-INSTANCE SETTING.

Model	Acc.	Pr.	Rc.	F1	AUPRC
Single-istance	.924	.667	.721	.689	.682
Mean-based	.957	.783	.877	**.826**	.804
Attention-based	.955	.786	.853	.817	.812
Conv-based	.956	.786	.860	.821	**.821**

ordering, nor it is able to identify the key instances. It is worth noting that the three variants compared in this section takes all spectral bands as input. We apply the same data augmentation and hyperparameters adopted per the single-instance setting, except for the number of epochs, which is lowered to 120 due to faster convergence.

We report the results in Table IV and observe that:

- The single-instance classifier earns worse results w.r.t. the multi-instance one. Indeed, the latter can capture temporal patterns, appearing in more than one image. Additionally, in the single-instance approach, the prediction could suffer in the presence of noisy acquisitions. As an example, we mention the cloudiness, which affects a large part of the available data. Differently, under the multi-instance setting, it is unlikely that all five images are cloudy;
- The Conv-based model leads to slightly better results w.r.t. the attention-based one; thus indicating that the order contains some valuable information for the task;
- Both the Attention-based and the Conv-based models exhibit a better AUPRC value compared to the Mean-based one. Nevertheless, F1-score does not support this result. We will conduct further investigations to assess which of the two metrics is more suitable for the problem at hand.

G. Limitations and Future Works

Even though the results being suitable, we found that the spectral bands perform just slightly better than the RGB ones (especially when looking at the F1-score, see Table III). This indicates the *curse of dimensionality* when dealing with complex data as the multi-spectral images. In future works, we will explore this aspect, trying to increase the gap between spectral bands and other hand-crafted representations. In this regards - further than gathering more data - we envision a network pre-training on crop and land classification directly, for which large datasets [27] are available currently. This could result in a proper initialization, thus improving the generalization capabilities of our model.

As discussed in Subsection V-A, we assess our proposal through Stratified K-Fold. In doing so, each of the K runs has been conducted by splitting entomological information about catches in two non-overlapping sets (namely, the training and test set). As a consequence, the same capture site could appear in both the training and test set, even though under

different time periods. Because of this, we cannot claim for sure whether the model generalizes to unseen sites. We will remove such a limitation in future works, defining a clear cross-site evaluation protocol.

As can be appreciated in Fig. 1 (right), the *C. imicola* shows some subtle patterns in its geographical distribution. In particular, it follows a sort of locality principle, stating that neighboring sites are likely to share the same levels of abundance. Thus, one could infer the presence of the *C. imicola* in a certain area just looking to its neighboring nodes, modeling the overall environment as a graph connecting sites and regions. This way, we could frame the problem under the graph-based semi-supervised framework [28], the latter assuming the label information to be smoothed over the graph. In this respect, we will profit from recent advances on Graph Convolutional Neural Networks (GCNNs) [29], [30] and propose a novel inference engine, which would exploit the local dynamics arising from the spread of *C. imicola*.

VI. CONCLUSION

In this work, we focus on predicting whether the *C. imicola* vector is present in the area of interest. We achieve that through a novel deep-learning-based approach, which extracts meaningful features from a sequence of satellite imagery depicting that area. Even though a bigger dataset may enhance its performance, our model already provides suitable generalization capabilities than both the usual hand-crafted features and the RGB bands only; therefore, all the spectral bands must carry fundamental features about the spread of the vector. Furthermore, we propose a multi-instance model leveraging temporal patterns, which leads to better results than single-instance one. Future work aims to improve the overall performance through a targeted pre-training phase, over a large dataset specific for satellite imagery. Moreover, as we observe locality playing a significant role in the geographical spreading of *C. imicola*, we will frame the here-discussed problem in terms of semi-supervised learning over graphs.

REFERENCES

[1] J. C. Semenza and J. E. Suk, "Vector-borne diseases and climate change: a european perspective," *FEMS microbiology letters*, vol. 365, no. 2, p. fnx244, 2017.

[2] A. Conte, M. Goffredo, C. Ippoliti, and R. Meiswinkel, "Influence of biotic and abiotic factors on the distribution and abundance of culicoides imicola and the obsoletus complex in italy," *Veterinary parasitology*, vol. 150, no. 4, pp. 333–344, 2007.

[3] K. M. McIntyre, C. Setzkorn, P. J. Hepworth, S. Morand, A. P. Morse, and M. Baylis, "Systematic assessment of the climate sensitivity of important human and domestic animals pathogens in europe," *Scientific reports*, vol. 7, no. 1, p. 7134, 2017.

[4] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort *et al.*, "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[5] M. Berger, J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen, "Esa's sentinel missions in support of earth system science," *Remote Sensing of Environment*, vol. 120, pp. 84–90, 2012.

[6] I. Sutskever, G. E. Hinton, and A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] C. Ippoliti, L. Candeloro, M. Gilbert, M. Goffredo, G. Mancini, G. Curci, S. Falasca, S. Tora, A. Di Lorenzo, M. Quaglia *et al.*, "Defining ecological regions in italy based on a multivariate clustering approach: A first step towards a targeted vector borne disease surveillance," *PloS one*, vol. 14, no. 7, p. e0219072, 2019.

[9] R. G. Bailey, "Identifying ecoregion boundaries," *Environmental management*, vol. 34, no. 1, pp. S14–S26, 2004.

[10] E. Ducheyne, M. A. M. Chueca, J. Lucientes, C. Calvete, R. Estrada, G.-J. Boender, E. Goossens, E. M. De Clercq, and G. Hendrickx, "Abundance modelling of invasive and indigenous culicoides species in spain," *Geospatial Health*, pp. 241–254, 2013.

[11] Y. Gao, J. P. Walker, M. Allahmoradi, A. Monerris, D. Ryu, and T. J. Jackson, "Optical sensing of vegetation water content: A synthesis study," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 4, pp. 1456–1464, 2015.

[12] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sensing of Environment*, vol. 221, pp. 430–443, 12 2018.

[13] S. Ji, Z. Chi, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, p. 75, 01 2018.

[14] H. G. H. L. J. G. Ming Li, Teng Pan, "Gaofen-2 mission introduction and characteristics," in *Proceedings of the 66th International Astronautical Congress (IAC)*, 2015.

[15] T. Zhang, J. Su, C. Liu, W.-H. Chen, H. Liu, and G. Liu, "Band selection in sentinel-2 satellite for agriculture applications," in *2017 23rd International Conference on Automation and Computing (ICAC)*. IEEE, 2017, pp. 1–6.

[16] M. Goffredo and R. Meiswinkel, "Entomological surveillance of bluetongue in italy: methods of capture, catch analysis and identification of culicoides biting midges," *Vet Ital*, vol. 40, no. 3, pp. 260–265, 2004.

[17] A. Giovannini, C. Paladini, P. Calistri, A. Conte, P. Colangeli, U. Santucci, D. Nannini, and V. Caporale, "Surveillance system of bluetongue in italy," *Vet Ital*, vol. 40, no. 3, pp. 369–84, 2004.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[23] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[24] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Neural Information Processing Systems*, 2017, pp. 3391–3401.

[25] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[26] G. Prathap and I. Afanasyev, "Deep learning approach for building detection in satellite multispectral imagery," in *2018 International Conference on Intelligent Systems (IS)*. IEEE, 2018, pp. 461–465.

[27] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," *arXiv preprint arXiv:1902.06148*, 2019.

[28] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on Machine Learning*, 2003, pp. 912–919.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[30] A. Porrello, D. Abati, S. Calderara, and R. Cucchiara, "Classifying signals on irregular domains via convolutional cluster pooling," in *International Conference on Artificial Intelligence and Statistics*, 2019, pp. 1388–1397.