Fusion of Satellite Images and Weather Data With Transformer Networks for Downy Mildew Disease Detection

WILLIAM MAILLET®, MARYAM OUHAMİ®, AND ADEL HAFIANE®, (Member, IEEE)
PRISME Laboratory, INSA CVL, University of Orléans, 45100 Bourges, France
Corresponding author: Maryam Ouhami (maryam.ouhami1@insa-cvl.fr)

This work was supported in part by the European Consortium ERA-NET ICT-AGRI-FOOD, in part by the French National Research Agency (ANR) under Grant ANR-21-ICAF-0002-01, in part by the European Project MERIA VINO, and in part by ICT-AGRIFOOD ERA-NET.

ABSTRACT Crop diseases significantly affect the quantity and quality of agricultural production. In a context where the goal of precision agriculture is to minimize or even avoid the use of pesticides, weather and remote sensing data with deep learning can play a pivotal role in detecting crop diseases, allowing localized treatment of crops. However, combining heterogeneous data such as weather and images remains a hot topic and challenging task. Recent developments in transformer architectures have shown the possibility of fusion of data from different domains, such as text-image. The current trend is to custom only one transformer to create a multimodal fusion model. Conversely, we propose a new approach to realize data fusion using three transformers. In this paper, we first solved the missing satellite images problem, by interpolating them with a ConvLSTM model. Then, we proposed a multimodal fusion architecture that jointly learns to process visual and weather information. The architecture is built from three main components, a Vision Transformer and two transformer-encoders, allowing to fuse both image and weather modalities. The results of the proposed method are promising achieving an overall accuracy of 97%.

INDEX TERMS Remote sensing, image processing, deep learning, data fusion, vegetation indices, crop monitoring, agriculture.

I. INTRODUCTION
Agricultural chemicals such as fungicides and pesticides are increasingly used to avoid and minimize disease damage. Nonetheless, the overuse of these chemicals has become problematic from an ecological and ethical point of view, while the European Union even aims to halve the use of chemicals by 2030 [1]. The challenge for farmers in the coming decades is, therefore, to find a way to better monitor their lands in order to maximize localized treatments for the infected crops instead of spraying chemicals on a large scale. Remote sensing is one of the popular tools in precision agriculture as it helps monitoring crops problems such as diseases, weed infestation, lack of water, etc [2].

Satellite imagery and deep learning have been widely used in crop monitoring as they provides applications in multiple areas: lands classification [3], yields predictions [4], wildfires management [5], disease detection [6] and various detection tasks [7]. In the recent years, multiple technologies have been applied to accomplish these tasks from CNNs [8] to RNNs [9] and recently transformers [10] which are becoming very popular state-of-the-art models, particularly for multimodal data such as images, text, etc.

Along with the satellite imagery, the sensor-based weather monitoring devices become an essential part of precision agriculture. Nowadays, most of the farmers have weather conditions monitoring facilities. Furthermore, combining weather data with the satellite images is becoming more and more attractive to prevent crop diseases. In this paper, we propose a fusion architecture combining the weather measurements...
with the satellites images in order to predict downy mildew in the crops. This architecture uses the Vision Transformer (ViT) [11] as image features extractor, the fusion with weather data is performed in bottleneck mode. The architecture was trained on images and weather data over a period of two years, and tested on the same type of data over two years, in order to evaluate the performance of the model in predicting the presence of downy mildew. The contributions of this paper are mainly:

- A new deep learning-based approach for temporal satellite image generation to handle missing information;
- A new multimodal ViT fusion architecture using three-encoder components to integrate heterogeneous data;
- An application of disease detection and identification on vine crops using weather and satellite images.

The paper is organized as follows. Section II provides a review of related works, section III describes the proposed approach. Section IV presents the experiments conducted in this study and the results. The discussion and conclusion are presented in Sections V and VI respectively.

II. RELATED WORK

Promising approaches for detecting diseases were proposed in recent years using machine learning and deep learning on weather data [12], [13]. In addition, Convolution Neural Networks (CNNs) have shown interesting precision results on remote sensing images for disease detection [14], [15], [16], [17]. However, CNN-based models require a large dataset and the training process takes a lot of time due to the models’ complexity. Recently, vision transformers were introduced as a better-performing solution for computer vision problems.

A. VISION TRANSFORMERS

Transformers were first proposed by [18], and they soon revolutionized the Natural Language Processing (NLP) domain. They replaced recurrent neural networks thanks to their shorter training time and parallelism. Yet, multiple architectures using the attention mechanism of the NLP transformers were created shortly after. ViT [11] on the other hand, is the most popular transformer used in image classification. It is known for its simple architecture allowing to process images with very few modifications and exploiting the full power of transformers. Transformers are becoming popular because they have proven to be more efficient than traditional CNNs. In fact, CNNs can recognize relationships between distant objects because of the convolutional operations whereas the attention mechanisms of transformers allow getting a better global understanding of the images and their intrinsic relationships.

Transformers have been used for multiple computer vision tasks, namely scene classification [19], [20], change detection [21], and image segmentation [22], [23]. ViTs were also used for various tasks in satellite imagery, such as change detection [24] and deforestation monitoring [25]. The ViTs results were convincing, they even outperformed the classical convolutional architectures.

B. MULTIMODAL FUSION

Since the agricultural field provides a rich variety of data from multiple sources, it seems more judicious to merge the variety of data in one model to achieve better performance. In disease detection, multimodal fusion is still an ongoing area of study [26], [27]. Multimodal fusion based on machine learning is able to jointly learn to process different modalities. Depending on the level of data abstraction, different fusion architectures are possible, such as input data level fusion, feature fusion allowing data integration using feature vectors, etc. We can divide deep learning approaches for multimodal fusion into two categories, those based on CNNs and transformer-based architectures.

1) CNN BASED MULTIMODAL FUSION

Using multiple modalities to learn is one of the most important topics in machine learning as it is the closest to human learning. Various multimodal fusion techniques presented in the literature were based on CNNs to extract feature maps from the input images. Those techniques have different applications, namely in the medical field [28], in mechanics [29] and in agriculture specifically data fusion for yield prediction [30], land monitoring [31], crop identification [32] and disease detection [33].

2) MULTIMODAL TRANSFORMERS ARCHITECTURES

Multimodal learning with transformers has been tested in multiple areas, especially in the audiovisual field to join video, language and audio features [34], [35], [36], [37] but also in deepfake detection [38], medical imagery synthesis [39], etc. Most of these proposed models extract embeddings from the modalities without transformers to make the fusion in a single custom transformer. In fact, they rarely use multiple transformers to extract features. [40] proposed a multimodal architecture using a custom bottleneck transformer to combine features. Nonetheless, this bottleneck architecture is within the transformer and not as a single transformer. In the satellite imagery areas, [41] proposed a multimodal fusion architecture using multiple image sources. The modalities features are extracted using LSTM cells and a modified ViT transformer. [42] proposed a multimodal architecture for object detection, using ViT as a feature extractor. Recently, [43] also proposed modifications of the ViT transformer allowing to use LIDAR data along with multispectral images. So far, very little attention has been paid to the fusion of satellite images with weather data for crop disease detection.

III. METHODOLOGY

In this section we describe the proposed method that combines Sentinel-2 satellite images as 2D information, with ground weather data as 1D information. First, we detail
The proposed method is split into two parts: first, we generate the training images, we use linear interpolation and deep learning procedures. Sec-
ond, we describe the fusion architecture between the two heterogeneous data. An overview of the entire process flowchart is depicted in Fig. 1.

A. SATELLITE IMAGES GENERATION

The Sentinel-2 satellite provides multispectral images of a given location, usually on a 5-day cycle. In cloudy conditions, the temporal resolution between exploitable images can be much lower. As the weather data is daily, it is essential to have one image per day to enable a consistent fusion of 2D/1D information. To overcome this limitation, we propose a data generation method that combines linear interpolation with noise injection and Convolutional Long Short-Term Memory (ConvLSTM). The goal is to train a ConvLSTM model to output the missing intermediate images between two dates. We believe that the ConvLSTM ability to capture temporal and spatial features would give more interesting predictions, compared to only interpolation based on an analytical model. The proposed method is split into two parts: first, we generate linearly interpolated images with gaussian noise, and then we use those images to train a ConvLSTM model outputting the final images that we use for our fusion model (see Fig. 2).

1) TEMPORAL INTERPOLATION OF IMAGES

To generate the training images, we use linear interpolation and gaussian noise. Let $I_n$ and $I_{n+k}$ be two real images (matrices) captured $k$ days apart. To generate all the missing images from day $n+1$ to day $n+k-1$, we use the Equation 1, $i$ is the day where an intermediate image is needed ($i \in ]n; n+k[$). Once the image is computed, we add gaussian noise to the image $I_{n+i}$ to create another image $I_{G,(n+i)}$ including small variations as indicated in Equation 2. This noise is added to integrate randomness to the linear interpolation and avoid the ConvLSTM model to overfit the linear interpolation. Note that Equation 1 and Equation 2 are element-wise operations.

\[ I_{n+i} = I_n \times (k - i) + I_{n+k} \times \frac{i}{k}, \quad i \in ]n; n+k[ \quad (1) \]

\[ I_{G,(n+i)} = I_{n+i} + \eta(\mu, \sigma) \quad (2) \]

2) ConvLSTM IMAGE GENERATION

The next phase consists in training the ConvLSTM model on the interpolated and real images, to generate a final dataset of daily artificial images. The ConvLSTM model is composed of four consecutive blocks, each one containing a ConvLSTM layer, a BatchNorm layer, and a leakyReLU layer. The model takes as input a set of daily ordered images: $\{D_{n-3}, D_{n-2}, D_{n-1}\}$ and outputs the next image, $D_n$. We train the model using the Adam optimizer and the Root Mean Square Error (RMSE) loss function. Once the model is trained, it is used to generate all the missing images and create a dataset for subsequent use.

\[ i_t = \sigma(W_{D_i} \ast D_t + W_{hi} \ast h_{t-1} + W_{ci} \circ C_{t-1} + b_i) \]

\[ f_t = \sigma(W_{D_f} \ast D_t + W_{hf} \ast h_{t-1} + W_{cf} \circ C_{t-1} + b_f) \]

\[ C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{Dc} \ast D_t + W_{hc} \ast h_{t-1} + b_c) \]

\[ o_t = \sigma(W_{D_o} \ast D_t + W_{ho} \ast h_{t-1} + W_{co} \circ C_t + b_o) \]

\[ h_t = o_t \circ \tanh(C_t) \quad (3) \]

Equation 3 illustrates the mathematical aspect of the ConvLSTM model where, $W_{D_i}, W_{hi}, W_{cf}, W_{D_f}, W_{hf}, W_{Dc}, W_{hc}, W_{D_o}, W_{ho}, W_{ci}, W_{co}$ denote the weights; $b_i, b_f, b_c$ and $b_o$ are the biases. The operation $\circ$ denotes element-wise product, $D_t$ denotes the current inputs and $h_{t-1}$ denotes the output of LSTM unit at the previous moment, and $\sigma$ as well as tanh($\cdot$) are nonlinear activation functions.

The ConvLSTM provides intermediate images to fill the gap of the missing information. Using deep learning to create these images will enhance basic linear interpolation since the ConvLSTM can extract both temporal and spatial features within images.

B. 2D-1D FUSION MODEL

To combine satellite imagery with weather data, we propose a fusion model based on transformers-encoders. Its architecture can be split into three main components: the ViT, the weather attention encoder, and a bottleneck transformer-encoder. The proposed architecture uses transformers to generate embeddings from the data, and then these embeddings are combined through the fusion bottleneck encoder. The output layer indicates whether crops are infected or not with a specific disease (here the downy mildew). Fig. 1 illustrates the proposed approach.
1) VISION TRANSFORMER (ViT)

The ViT introduced by [11] is based on a transformer-encoder. It becomes a popular architecture for its simplicity, and efficiency in terms of computation and performance. The ViT splits images in small patches ordered as a sequence and uses only the encoder part of the traditional encoder-decoder transformer architecture, which makes it light. The transformer learns by measuring the relationship between image patches. This relationship can be learned by providing attention in the network. It is also possible to extract attention maps, which are useful for understanding the output of the model and the parts of the image on which the focus is. This is a key concept for anomaly detection, as we could interpret certain areas of the attention maps as infected crops.

Multi-Head Attention is basically a linear projection of the queries, keys, and values. ViT is based on three main components, patch embedding Equation 6, feature extraction via stacked transformer encoders Equation 7 and classification head Equation 8 and Equation 9.

\[
\begin{align*}
    z_0 &= [x_{\text{class}}; x_p^1E; x_p^2E; \ldots; x_p^N E] + E_{\text{pos}}, \quad E \in \mathbb{R}^{(P^2,C) \times D}, E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \\
    z'_l &= \text{MSA}(LN(z_{l-1})) + z_{l-1}, \quad l = 1 \ldots L \\
    z_l &= \text{MLP}(LN(z'_l)) + z'_l, \quad l = 1 \ldots L \\
    y &= LN(z_{224}^L)
\end{align*}
\]

In the proposed architecture, the ViT takes inputs \( I_V = \{x_1, x_2, x_3\} \), where \( x_i \) has a dimension of \( (3, 224, 224) \) which represents a vegetation index encoded with RGB image. Three different RGB vegetation indices, resulting in a tensor of size \( (9, 224, 224) \). It outputs extracted features \( O_V \) of dimension \( (14, 14) \). Fig. 3 illustrates the ViT encoder architecture used.

2) WEATHER ATTENTION ENCODER

The Weather Attention Encoder is a simple Transformer Encoder that has the role of embedding the weather data. It takes as inputs \( I_W = \{w_1, \ldots, w_n\} \) a vector of weather features, such as rain duration and quantity, potential evaporation, sunlight amount, etc. First, the input \( I_W \) is expanded to a high dimension using linear and dropout layers Equation 10. Then, this expanded vector is passed to the first encoder block (out of 12 encoder blocks) Equation 11. After these blocks, the output representation \( O_W \) is produced using a network composed of linear, flatten, and dropout layers Equation 12. We call this output \( O_W \) the weather embeddings of dimension \( (14, 14) \).

\[
\begin{align*}
    A_W &= \text{Linear}((\text{Dropout}(I_W))) \\
    A'_W &= \text{EncoderBlock}(A_W) \\
    O_W &= \text{Linear}((\text{Flatten}(\text{Dropout}(A'_W))))
\end{align*}
\]

The encoder uses the architecture proposed in [18]. The encoder block can be decomposed into two sub-parts. The first part consists of multi-head attention and residual connections on the input vector \( I_e \) to produce an intermediate
feature $H_c$ as indicated in Equation 13, on which layer normalization is applied (LN). The same process is made using a feed-forward network in the second sub-part of the encoder block described in Equation 14.

\[
H_e = LN(I_e + Dropout(MultiHeadAtt(I_e))) \quad (13)
\]

\[
O_e = LN(H_e + Dropout(FFN(H_e))) \quad (14)
\]

3) FUSION BOTTLENECK ENCODER
The Fusion Bottleneck Encoder takes as input $I_F$, the concatenation of the two previous embedded vectors that represent the visual and weather information, as described in Equation 15. The concatenation ends up being a (28, 14) tensor, since the $O_V$ and $O_W$ are of dimension (14, 14) each. This concatenated vector $I_F$ is then passed to the three parts of the encoder, using the same process as the weather attention module. First, the vector is expanded Equation 16, then passed to the encoder Equation 17. The resulted $A_F'$ is then mapped to a vector using a feed-forward network, made of linear, flatten and dropout layers Equation 18. The last layer $O_F$ consists of 2 outputs to predict whether downy mildew disease exists or not (yes or no output).

\[
I_F = O_W \oplus O_V \quad (15)
\]

\[
A_F = Linear(Dropout(I_F)) \quad (16)
\]

\[
A_F' = Encoder(A_F) \quad (17)
\]

\[
O_F = Linear(Flatten(Dropout(A_F'))) \quad (18)
\]

IV. EXPERIMENTS AND RESULTS
This section presents the different experiments conducted in this research work. It includes data collection, implementation details, evaluation results of the image generation, and the fusion method.

A. DATA COLLECTION
Two types of data were collected, satellite images and weather conditions. To evaluate the presence of downy mildew within the change in weather conditions for the same periods in each year, we selected images and weather data of June, July and August for the period 2018 until 2021 from the plots of Lycée Agricole d’Amboise in Centre Val de Loire region, France.

1) IMAGE COLLECTION
Remote sensing is an important data source for crop monitoring. The images of the vine crops from the studied site were collected using the Sentinel-2 Level-2A product. Sentinel-2 is a satellite mission of the European earth surveillance program Copernicus. It provides multispectral images with different spatial resolutions 10, 20, and 60 m [44]. The orthoimages were provided with atmospheric correction which is known to improve the images for subsequent use. In addition, Level-2A images have useful features such as cloud detection.

Sentinel-2 images provide rich sources of data related to vegetation. The spectral bands allow the calculation of vegetation indices that are useful for measuring crop conditions such as vigor, biomass, chlorophyll content, and disease. The Difference Vegetation Index (NDVI) is a widely used vegetation index to measure the health status of vegetation, based on visible and near-infrared light reflected by vegetation. However, other types of indices can be useful for extracting more relevant information on vegetation or land condition. The Normalized Difference Chlorophyll Index (NDCI) introduced by [45] is an index that describes the chlorophyll concentration, designed for water regions. The Normalized Difference Moisture Index (NDMI) measures the moisture levels of the crops and helps monitor droughts.

In this study we used the three indices individually and in combination, all extracted from the SentinelHub API coded on RGB images. Table 1 lists the Sentinel-2 bands of the three indices with their spatial resolution. Fig. 5 shows the indices NDVI, NDCI, and NDMI extracted from the studied vineyards zone.

| Index        | Bands  | Spatial Resolution |
|--------------|--------|--------------------|
| NDVI         | B04, B08 | 10m                |
| NDCI         | B04, B05 | 10m                |
| NDMI         | B8A, B11 | 20m                |

2) IMAGES PROCESSING
Even if the images are requested with a maximum cloud percentage of 10%, there are still many images with large parts of clouds that slip through the filtering process. We deleted some of these remaining images using a cloud detection algorithm with pixel values and we cleaned the rest by hand. The satellite images cover a zone of four different vine plots. Since this zone contains buildings, trees, etc. We cropped the images to only keep the appropriate vine parcels (see Fig. 5-a). In total, The dataset is composed of 1472 images (including the generated ones). We have 92 images per year per plot from June to August in four plots over four years from 2018 to 2021, where the numbers of real images per year are 38, 33, 40, and 29 respectively, more details are presented in Table 2. Each image is labeled positive or negative to mildew depending on the ground truth. For our dataset, July and August of 2018 and 2021 were labeled as positive, and the rest of the dataset was labeled negative.

3) WEATHER DATA
Research indicates that downy mildew infections are due necessary to an accumulation of days with favorable weather
conditions [46]. In France, the downy mildew of the vine generally starts to appear at the end of spring and begins to disappear with the fall of the leaves. In spring, after the maturation of vine mildew ascospores, they germinate in rainwater from an average temperature of 11°C and cause primary contamination. After an incubation period of 10 to 20 days depending on the temperature, the secondary contaminations phase arrives in the presence of rain [46], [47].

The weather data used in this study was collected from the local meteorological station implemented on the studied site which tracks daily weather conditions. The data contains precipitation (in mm), rainfall variability, rainfall duration, minimum, average, and maximum temperature (in °C), humidity, potential evaporation, and on-site sun exposure (min, max, mean). These 11 features form a vector of shape [1], [11] which is passed to the weather-dedicated transformer.

B. IMPLEMENTATION DETAILS AND SETUP

The experiments were performed on a computer with an Intel®Xeon(R) W-2123 CPU and an Nvidia GeForce GTX 1080 Ti, on Ubuntu 20.04 LTS. The model has been trained using the Adam optimizer with a learning rate of \(1 \times 10^{-6}\). We used a cosine warmup learning rate scheduler from the “transformers” library [48], with a warmup of 100 epochs of 600 total epochs. The model early stopped at the 120/130th epoch.

The used hyperparameters for the ViT are 12 encoder-blocks layers with 8 heads each, a patch size of 16 pixels, an embedding size of dimension 768, and images of size 224 × 224. For the other encoders of the model, we also use 12 encoder-blocks layers using 8 heads each and a feed-forward network size of 128. The model embedding dimension is 64.

The next section presents the evaluation results of the temporal image generation described in section III-A, and then the performance of the 2D-1D fusion architecture (see section III-B) for downy mildew detection. We carried out the experimental study with the variation of several parameters of the proposed methods.

C. EVALUATION OF THE IMAGE GENERATION METHOD

The performance was evaluated with the one-leave-out cross-validation procedure. Each fold contains images of one year, so we have four folds for these experiments. The ConvLSTM network was trained on three folds and tested on one fold. Note that only the real images are involved in the test. Three previous images were used to predict the actual image. The RMSE metric measures the error between the predicted image and the real one. The ConvLSTM network was trained with different noise values added to the interpolated images Equation 2. The noise tested is of Gaussian type with mean \(\mu = 0\) and different standard deviation \(\sigma\), varying from 0.04 to 0.2, with a step of 0.02. We can observe in Table 3 that the RMSE is of the range of \(10^{-3}\) for all indices, which represents a low prediction error. The NDCI error is the smallest, followed by the NDVI and NDMI. Fig. 6 shows the RMSE error for the three indices as a function of the noise added to the interpolated images. Similarly, the RMSE error remains low (range of \(10^{-3}\)) for the different values of the noise variance. We can notice some spikes in the curves, NDVI and NDMI. The RMSE of NDCI is more stable and smaller than the other two indices. Nevertheless, small noise values seem to work better for all indices.

### TABLE 3. Average of RMSE value for each indice and for each years over sigma values.

| Year          | NDCI  | NDVI  | NDMI  |
|---------------|-------|-------|-------|
| RMSE 2018     | 1.94  | 2.23  | 2.15  |
| RMSE 2019     | 2.95  | 3.071 | 3.86  |
| RMSE 2020     | 2.18  | 2.67  | 2.42  |
| RMSE 2021     | 2.47  | 3.07  | 2.64  |
| Total         | **2.38** | 2.75  | 2.76  |
D. EVALUATION OF THE 2D-1D FUSION ARCHITECTURE

The aim is to evaluate the performance of the proposed fusion method with different setups. We investigated the influence of combining different vegetation indices, ablation of some components, pruning of connections and changes in model layers. For this purpose, we used real images and those generated by the proposed method, as well as weather data, for training and testing. Data from 2018 (presence of downy mildew) and 2019 (no downy mildew) were considered for training. For the test, we used data from 2020 (no downy mildew), 2021 (downy mildew). The performance measurement was performed by the well-known accuracy measure and the F1 score.

1) MODEL COMPARISON

Transformers were primarily designed to process sequential input data, encoding this information is a native capability of the transformer. The critical part of our architecture is the encoding of the generated images, which needs to be consistent in embedding space with the weather data. Hence, in addition to ViT we tested two different popular architectures to compare their capabilities to embed the generated image information into the fusion module. To assess the performances of different models, we replaced the ViT model by the Microsoft CvT model [49] and a classical Fully Convolutional Network. The CvT model is pre-trained on ImageNet-1k, and the FCN model uses a ResNet50 backbone, and it is pre-trained with a subset of COCO-train2017 containing 20 categories present in the Pascal VOC dataset. All models have been adapted to the inputs data, where the first embedding is changed to use 9 channels (NDCI, NDVI, NDMI) instead of 3 and the classifiers of the model have been replaced with a simple output network to suit the attention maps used with the ViT model. The current results presented in Table 4 show that the ViT model seems the best performing among the other options. The analysis of the results shows that the ViT model provides the best performance among the other models, indicating that it is better suited for our problem.

| Table 4. Comparison of the image encoding models in the fusion architecture. For each evaluation, the “Vision Transformer” module was replaced by the ViT, CvT or FCN model. |
|---------------------------------|-----------------|-----------------|
|                                 | ViT             | CvT             | FCN             |
| Accuracy                       | 0.970           | 0.946           | 0.810           |
| F1-Score                        | 0.975           | 0.954           | 0.770           |

2) VEGETATION INDICES COMBINATIONS

Multiple vegetation indices combinations were tested, and the best accuracy and F1-Score were obtained with the combination of all three indices. As we can observe in Table 5, NDCI might be the one playing the most important role as it produces a very good accuracy on its own and performs well with both NDVI and NDMI. Even though this vegetation index seems to be the most important, we cannot neglect others as the combination of all three outperforms by more than 2%.

Fig. 7 shows vegetation indices images used as an input for the ViT part of the model (see section III-B1). We can observe that the attention map looks similar to the vegetation indices images, as it highlights the shape of some parts of the image that may be related to downy mildew. This also indicates that the model seems to learn to recognize patterns in vegetation indices and that it’s not highlighting uninteresting parts of the image.

3) ABLATION STUDY

To evaluate whether the network architecture is biased by the weather features or the image features, we tested the network with an ablation method to cut off parts of the network. This consists in cutting one of the parts of the network, either the ViT encoder or the Weather Attention Encoder. We replace the concatenation of weather and image features of dimension (28, 14) by a (14, 14) features map as if the other part of the model did not exist, and we adapt the Fusion Bottleneck
Transformer to use these new dimensions as input by tweaking the input parameters. We then train this new model using the same process as for the global model, ignoring the missing parts.

### TABLE 6. Results of the ablation study of the fusion model.

|                  | Image only | Weather Only | Fusion Model |
|------------------|------------|--------------|--------------|
| Accuracy         | 0.481      | 0.662        | 0.970        |
| F1-Score         | 0.630      | 0.790        | 0.975        |

After the training, we obtained the results described in Table 6. When cutting off the weather data, the model achieves 48.1% accuracy, which is fairly bad, and when cutting the images part, the accuracy goes up to 66.2%; which is better but still not good. This study shows that the model does not appear to be biased by weather or images and that it needs both inputs for correct learning.

### 4) PRUNING STUDY

Network pruning has become common to study models, as it can enlighten a potential overparameterization. We tested pruning by taking the parameters that give the best accuracy (97% accuracy and 97.5% F1-Score), and we prune parameters of either part of the network (only the Weather Attention Encoder for instance) or the whole network.

### TABLE 7. Results of the fusion model pruning.

| % pruned | W.A.E. | Vision T. | Fusion B. | Global |
|---------|--------|-----------|-----------|--------|
| 1%      | 0.733  | 0.982     | 0.555     | 0.516  |
| 5%      | 0.647  | 0.976     | 0.54      | 0.478  |
| 7.5%    | 0.635  | 0.969     | 0.461     | 0.508  |
| 10%     | 0.612  | 0.962     | 0.543     | 0.501  |

W.A.E. denotes Weather Attention Encoder, ViT denotes Vision Transformers and Fusion B. denotes Fusion Bottleneck, and Global means pruning global.

Table 7 presents the accuracy of the model when pruning 1%, 5%, 7.5% and 10% of each model part. We can observe that the ViT is more resilient to these prunings, and it can be explained by its number of parameters; the ViT has approximately 86 million parameters which makes it robust, while the other encoders of the model have almost three times fewer parameters. As we saw in the ablation study, keeping only the weather gives better results than keeping only the images, and here we can see that removing parts of the ViT does not affect the accuracy by a lot. This means our model gives the best results when using images with the weather, but the weather module is still the most important part of the network.

### 5) ENCODER LAYERS

The number of layers in the encoder has an impact on the accuracy of the model, as shown in Figure 8. The performances remain correct for a number of layers between 6 and 16. Otherwise, we observe a degradation in performances beyond these values. The optimal number of layers providing the best performances in these experiments is 12.

![Figure 8. Model precision depending on the number of encoder layers.](image)

### V. DISCUSSIONS

The first hypothesis of this research states that a deep neural network trained with data obtained by linear interpolation will be able to generate artificial images intermediate between two temporally distant acquisitions. Thus allowing to match and merge of heterogeneous data acquired with different frequencies. In this research, it was found that interpolation combined with a deep network (ConvLSTM) provides good results to fill the information gap in satellite images. The error is small when the generated images are compared to the real ones. Another important result is that perturbing the linear interpolation with Gaussian noise helps to improve the training and convergence of the ConvLSTM network. Therefore, we can assume that this method could be an interesting tool to complete the data and allow better performance of deep networks.

Even though ConvLSTM achieves good RMSE, further research should be undertaken to study the effectiveness of this model using more data. Real satellite images are not very frequent and there is no guarantee that the intermediate images (interpolated images) are a good representation of the ground truth. Nevertheless, given the results obtained,
we can expect that the images generated will be close to the real images, and this will not significantly affect the machine learning process. On the other hand, the artificial generation of intermediate images could also induce noise in the images or pixels invisible to the naked eye, which could bias the network. This is certainly not the case because we can observe in Table 6 that when the ViT is cut, the model still seems able to learn, even if it is not very accurate.

The second working hypothesis consisted of showing the effectiveness of transformers for the implementation of information fusion with the aim of crop disease detection. The difficulty lies in the heterogeneity of the data and their dimensions, as well as the difference in the acquisition frequencies. The proposed method was able to address these problems. The obtained scores showed that the fusion of multispectral satellite images (2D) and weather data (1D) offers better performances for the detection and identification of downy mildew. Indeed, for the test set, our model outputs predictions with a 97% accuracy and a 97.5% F1-Score. These results seem promising, and the qualitative analysis of attention over images seems to confirm that the network learned how to recognize downy mildew on vegetation indices of satellite images.

The weather has an impact on the development of downy mildew, the vegetation indices include indications of the health status of a crop. The results showed that the combination of the two parts gives better scores than a single type of data. Comparing the results of ablation experiments shows that each module (weather data, image), affects the fusion results, as depicted in Table 6. From the results obtained, we can assume that each module provides complementary information which increases significantly the performance. It was also found that using the three vegetation indices (NDVI, NDCI, NDMI) is more effective than using only one or two. Nevertheless, it seems that NDCI has a few more discriminating elements compared to the other two (see Table 5).

The proposed fusion model performs well, but there is room for improvement in several aspects of the methodology. For instance, improvements could be made to the fusion model. It seems that the model gains a little bit in accuracy when a small percentage of its parameters are cut, which possibly indicates redundancies in the parameters. We believe it is possible to modify the Weather Attention Encoder and Fusion Bottleneck Encoder to enhance them for this task, using a custom loss function or small modifications in the architecture.

VI. CONCLUSION
In this research work, the objective was to take advantage of the complementarity of heterogeneous data to detect downy mildew disease in vine plots. For this purpose, we have proposed a new framework where the first part consists in generating intermediate satellite images, to create daily images and pairing them with the weather data. The second part consists of a new data fusion architecture based on transformer networks that combine weather data and satellite images. We used two transformers (an encoder and ViT) to extract modalities features/embeddings (weather and satellite images), and a third bottleneck transformer-encoder outputs the prediction based on the embeddings/features. We achieve promising results which validates the hypothesis that fusion using multiple transformers instead of a single custom one is possible and can achieve good results. We believe these results are interesting for agriculture challenges as they could lead to better crop monitoring. In addition, the proposed architecture can be used for other application domains.

ACKNOWLEDGMENT
The authors would like to thank the Institut Français de la Vigne (IFV) for providing ground truth information and Lycée Agricole d’Amboise for providing studied plots.

REFERENCES
[1] Farm to Fork Targets Progress. Accessed: Jul. 18, 2022. [Online]. Available: https://food.ec.europa.eu/plants/pesticides/sustainable-use-pesticides/farm-fork-targets-progress_en
[2] M. Weiss, F. Jacob, and G. Duveiller, “Remote sensing for agricultural applications: A meta-review,” Remote Sens. Environ., vol. 236, Jan. 2020, Art. no. 111402.
[3] M. Persson, E. Lindberg, and H. Reese, “Tree species classification with multi-temporal Sentinel-2 data,” Remote Sens., vol. 10, no. 11, p. 1794, Nov. 2018.
[4] M. L. Hunt, G. A. Blackburn, L. Carrasco, J. W. Redhead, and C. S. Rowland, “High resolution wheat yield mapping using Sentinel-2,” Remote Sens. Environ., vol. 233, Nov. 2019, Art. no. 111410.
[5] E. B. Castillo, E. T. Cayo, C. de Almeida, R. S. López, N. R. Bricenó, j. S. López, M. B. Gurbillón, M. Oliva, and R. Espinoza-Villar, “Monitoring wildfires in the northeastern peruvian Amazon using Landsat-8 and Sentinel-2 imagery in the GEE platform,” ISPRS Int. J. Geo-Inf., vol. 9, no. 10, p. 564, Sep. 2020.
[6] H. Santos, T. Gunawan, R. H. Jatnika, W. Darmosarkoro, and B. Minasny, “Mapping and identifying basal stem rot disease in oil palms in north Sumatra with QuickBird imagery,” Precis. Agricul., vol. 12, no. 2, pp. 233–248, Apr. 2011.
[7] K. Themistocleous, C. Papoutsas, S. Michaelides, and D. Hadjimitsis, “Investigating detection of floating plastic litter from space using Sentinel-2 imagery,” Remote Sens., vol. 12, no. 16, p. 2648, Aug. 2020.
[8] J. Rosentretter, H. Hagensieker, and B. Wasse, “Towards large-scale mapping of local climate zones using multitemporal sentinel 2 data and convolutional neural networks,” Remote Sens. Environ., vol. 237, Feb. 2020, Art. no. 111472.
[9] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, “Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., Jul. 2019, pp. 214–217.
[10] Y. Yuan, L. Lin, Q. Liu, R. Hang, and Z.-G. Zhou, “SITS-former: A pre-trained spatio-spectral-temporal representation model for Sentinel-2 time series classification,” Int. J. Appl. Earth Observ. Geoinf., vol. 106, Feb. 2022, Art. no. 102651.
[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16x16 words: Transformers for image recognition at scale,” 2020, arXiv:2010.11929.
[12] A. Khattab, S. E. Habib, H. Ismail, S. Zayan, Y. Fahmy, and M. M. Khairy, “An IoT-based cognitive monitoring system for early plant disease forecast,” Comput. Electron. Agricul., vol. 166, Nov. 2019, Art. no. 105028. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0166169919311949
[13] S. Trilles, J. Torres-Sospedra, Ó. Belmonte, F. J. Zarazaga-Soria, A. González-Pérez, and J. Huerta, “Development of an open sensorized platform in a smart agriculture context: A vineyard support system for monitoring mildew disease,” Sustain. Comput., Informat. Syst., vol. 28, Dec. 2020, Art. no. 100309.
[14] M. Kerkech, A. Hafiane, and R. Canals, “VddNet: Vine disease detection network based on multispectral images and depth map,” Remote Sens., vol. 12, no. 20, p. 3305, Oct. 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/20/3305

[15] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, and B. F. Birstein, “Soybean yield prediction from UAV using multimodal data fusion and deep learning,” Remote Sens. Environ., vol. 237, Feb. 2020, Art. no. 111599.

[16] J. Abdulridha, Y. Ampatzidis, P. Roberts, and S. C. Kakarla, “Detecting powdery mildew disease in squash at different stages using UAV-based hyperspectral imaging and artificial intelligence,” Biosyst. Eng., vol. 197, pp. 135–148, Sep. 2020.

[17] T. Poblete, C. Camino, P. S. A. Beck, A. Hornero, T. Kattenborn, M. Saponari, D. Boscia, J. A. Navas-Cortes, and P. J. Zarco-Tejada, “Detection of Xylella fastidiosa infection symptoms with airborne multispectral and thermal imagery: Assessing bandset reduction performance from hyperspectral analysis,” ISPRS J. Photogramm. Remote Sens., vol. 162, pp. 27–40, Apr. 2020.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–12. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee24357de9f1b5df03e1c4e8454-aaa-Paper.pdf

[19] Z. Xu, W. Maillet et al.: Fusion of Satellite Images and Weather Data With Transformer Networks

[20] L. Jing, T. Wang, M. Zhao, and P. Wang, “An adaptive multi-sensor data fusion method based on deep convolutional neural networks for fault diagnosis of planetary gearbox,” Sensors, vol. 17, no. 2, p. 414, Feb. 2017.

[21] Z. Chu and J. Yu, “An end-to-end model for Rice yield prediction using deep learning fusion,” Comput. Electron. Agric., vol. 174, Jul. 2020, Art. no. 105471.
WILLIAM MAILLET received the Baccalaureate degree from the Montesquieu High School, Le Mans, France. He joined the PRISME Laboratory as an Intern, from April 2022 to August 2022, in the field of artificial intelligence and computer vision. He has been specialized in big data, since 2018. He is currently a Student Engineer at the INSA Centre Val-de-Loire.

MARYAM OUHAMI received the B.S. degree in mathematics and informatics from the University of Hassan 2, Morocco, in 2016, and the M.S. degree in data sciences and big data from the National School of Computer Science and Systems Analysis ENSIAS, Morocco, in 2018. She is currently pursuing the joint Ph.D. degree with the PRISME Laboratory, National Institute of Applied Sciences Centre Val de Loire (INSA CVL), France, and the IRF-SIC Laboratory, Ibn Zohr University, Morocco. Her research interests include machine learning algorithms and data fusion applied to agriculture.

ADEL HAFIANE (Member, IEEE) received the M.S. degree in embedded systems and information processing and the Ph.D. degree from the University of Paris-Sud, in 2002 and 2005, respectively. He worked one year in teaching and research at the University of Paris-Sud and later for one year at the INSA Centre Val de Loire (INSA CVL) (Former ENSI de Bourges). He was a Postdoctoral Fellow at the Computer Science Department, University of Missouri, from 2007 to 2008. He has been with INSA CVL, since September 2008. Since 2020, he has been the Head of the PRISME Laboratory, Images and Vision Group, University of Orléans-INSA CVL. He was also a Visiting Researcher at the University of Missouri for several periods, from 2009 to 2013. He is currently an Associate Professor with the Electrical Engineering and Computer Science Department, INSA CVL. His research interests include theory and methods of image processing, computer vision, and machine learning for several applications: medical, agriculture, and robotics.