A new perspective on using generative learning and multimodal data fusion for the classification of severe weather phenomena

Ştefan Alexandrescu*
*Faculty of Mathematics and Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania
stefan.alexandrescu@stud.ubblcuj.ro

Abstract*—Clouds play an essential role in Earth's climate, being the source of precipitation, controlling the amount of solar energy that reaches the surface of our planet, as well as affecting the overall Earth radiation budget. Thus, cloud classification is important as it helps in weather forecasting and supervising climate changes. Clouds are organized in many forms, their composition and density are variable, with different colours and heights, which makes their classification a challenging one. This paper proposes several Machine earning (ML) models and hybrid pipelines, including a Diffusion-based and a Vision Transformer-based classifier, for classifying various cloud images, and analyses their performance and ability of learning to extract and classify the features of distinct cloud types. Experiments are conducted on two ground-based cloud data sets of different sizes and characteristics, using data augmentation and by training each of the implemented approaches. The performances of the best proposed versions are then compared to that of other deep learning architectures used in the literature for clouds classification. The comparative analysis highlights metric scores around the same range, in terms of *Accuracy*, *Precision*, *Recall* and *F1-score*.

AMS - MSC Primary 68T07; Secondary 68T45;
ACM classifications:

- **Computing methodologies**
  - **Artificial intelligence**
  - **Computer vision**
  - **Computer vision problems**
  - **Object identification**
- **Computing methodologies**
  - **Machine learning**
  - **Learning paradigms**
  - **Supervised learning**
  - **Supervised learning by classification**
- **Applied computing**
  - **Physical sciences and engineering**
  - **Earth and atmospheric sciences**
  - **Environmental sciences**

1. Introduction

Due to the industrial revolution's progress, global warming has naturally triggered climate changes all over the world, leading to continuous increases of extreme weather events occurrences and hazards. Nowadays, the population is increasingly being affected by these phenomena, especially in the context of the latest extreme precipitation produced by cyclones all around the Eastern side of Europe and on a national level. As the number and intensity of severe meteorological phenomena increases, predicting them in due time to avoid disasters becomes highly demanding for meteorologists. Therefore, it is mandatory that we find better ways of identifying, examining and predicting such weather events, and to find solutions for overcoming such undesirable happenings with as minimal damages as possible.

Clouds represent means to study short-term weather conditions, as well as long-term climate changes [1], being important for weather analysis and forecasting. Cloud classification is important as it helps in short-term forecasting of severe weather events (e.g., severe rainfall, hailstorms detection). The various types of clouds are associated with certain weather conditions and, therefore, an accurate categorization of clouds types leads to better forecasts. Cloud identification is one of the main activities of the weather stations personnel. Previously, weather observers manually classified cloud types and the amount of sky covered by clouds, but modern automated weather stations networks, such as the New York State Mesoscale Network [2], capture high-resolution images of the sky at regular intervals, creating a need for automated cloud classification methods. However, factors such as similar shapes, reduced colour palette, overlapping of multiple cloud layers, the "whitening effect" given by the sun, make the clouds classification problem a highly complex one [3]. Clouds determine, in most cases, the weather appearance in the area above which they exist, making them some of the most important elements of meteorological observation [4]. Classifying clouds is thus employed in weather forecasting [5], meteorological reanalysis and climate studies.

Predicting the occurrence of meteorological phenomena is very hard to make on a larger time scale, due to the very dynamic nature of atmospheric movements and clouds. Even the slightest changes in initial conditions would change the outcomes tremendously, which also proves the base principles of chaos theory.

Predicting the occurrence of meteorological phenomena is very hard to make on a larger time scale, due to the very dynamic nature of atmospheric movements and clouds. Even the slightest changes in initial conditions would change the outcomes tremendously, which also proves the base principles of chaos theory. Within the Machine learning (ML) domain, generative learning rose as a cutting-edge paradigm, models being capable of creating new data of many forms such as text, images, audio content and videos. They essentially synthesise new data that closely resembles the distribution of original data provided during training stages. Generation of new content has many applications, and meteorology has lately shown up among them [6]. Diffusion models are also another strong example of generative algorithms that have revolutionised the Computer Vision scene. Their work essentially consists of two steps: forward and reverse diffusion processes that progressively alter an image and learn significant features in an efficient way. Additionally, Diffusion classifiers are an extension of regular Diffusion models which are adapted to classification algorithms. They represent one of the latest ideas in the field for such purposes as of now. Transformers are another great example of recent advancements in text processing [7], vision transformers (ViT) being proposed as a newer variant tailored to image processing [8]. This technique proves to be very efficient in extracting useful features from images, including potential weather-related ones, with the help of the so-called attention mechanism.

Additionally, multimodal data fusion can be beneficial for enhancing performance of deep learning models by integrating and processing information coming from multiple data sources. Moreover, by understanding and exploiting the complementary nature of different data sources, different information is brought together and the model is able to generalise better across tasks and domains. Despite its relevance in the meteorological domain, where data come from different sources (e.g., radar, satellite, weather stations observations), few approaches in the literature are addressing this topic.

1.1. Motivation

The current paper proposal aims to explore the effectiveness of using recent ML advancements (diffusion-based generative learning, ViT, multimodal learning) in severe weather analysis, a domain where the application of these models have the potential to advance the current state-of-the-art.

Due to its importance and relevance for weather analysis and forecasting, we decided to focus the current proposal on the problem of cloud images (obtained from ground weather stations, satellite imagery, radar data) classification. However, we note the generality of the targeted approaches proposed for cloud images classification, as they can easily be adapted and extended to other classification tasks (such as rainfall, hailstorm, snow, hoar frost, fog, etc) using images obtained from weather stations. Figure 1 also displays the ten standard types of clouds.

The issue of cloud classification from images is not trivial as many cloud characteristics, for example colour or shape, are not measurable parameters and might be subject to the meteorologist judgement. This problem becomes even more difficult in the case of automated weather stations, without human personnel. From a scientific standpoint, the diverse shapes of clouds constituting masses of air are important to be recognized and classified as they represent atmospheric manifestations of fluid motions that produce what we call clouds. From an empirical point of view, it was observed that clouds take several distinctive forms. These forms have been internationally given names so that weather observers could report the sky's local status in an understandable way, without the help of images [9]. This is due to the fact that many meteorological stations have to report the sky's status once every several hours, or even every hour.

The main objective of the paper is to enhance the performance of cloud classification from images (ground-based, satellite) and other types of weather-related data (e.g., radar data) using diffusion classifiers and vision transformers. In addition, fusing remote sensing (RS) data obtained from various sources (weather stations, radar data, satellite images, or outputs from Numerical Weather Prediction models) using multimodal data fusion techniques are envisaged for performance improvement. We also aim to address common data-related issues such as class imbalance by implementing a new form of data augmentation, with the use of automatically and artificially generated images through a diffusion model.

Besides the existing publicly available datasets we aim to benefit from the current collaboration of the faculty's Machine Learning research group with the Romanian Meteorological Administration (ANM) and use real world data provided by ANM (images and radar data provided by national weather stations).

1.2. Contributions

Our proposal currently aims to find innovative methods of identifying and classifying severe weather events, with a particular focus on cloud classification. Many studies have attempted various augmentation of training data techniques, but none of them have used generation with a diffusion model

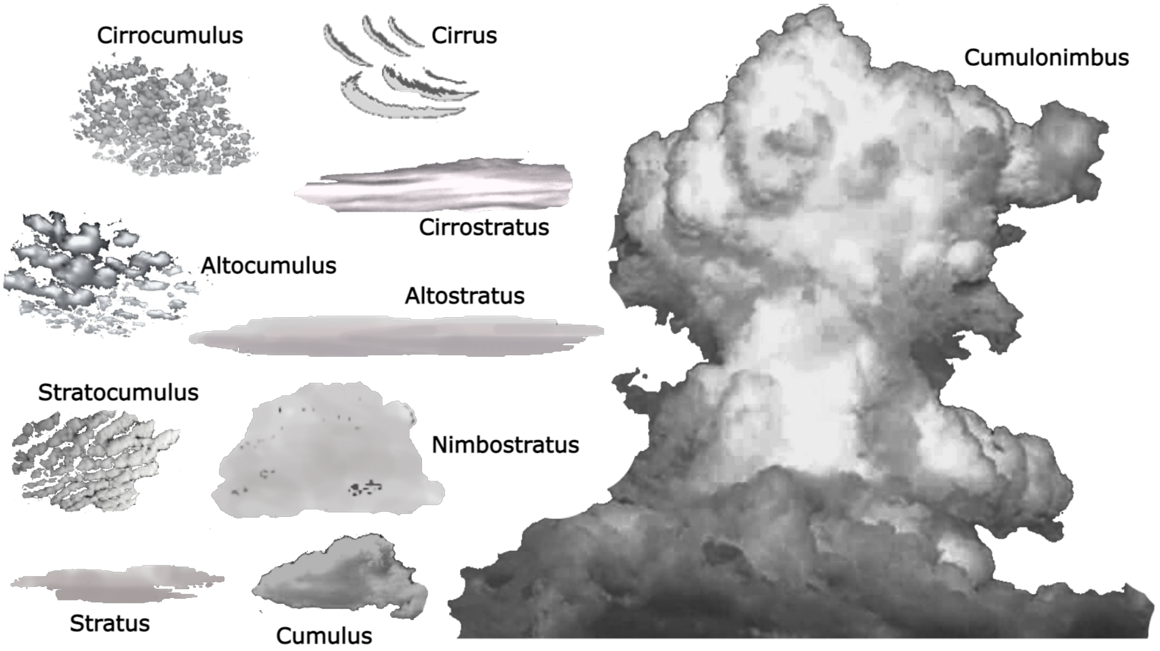

Figure 1. The ten types of clouds.

for the problem at hand. Most known augmentation techniques involve simple geometric or colour transformations, which might not bring a very diverse distribution of data as synthetic generation would. Additionally, many research papers approached various neural network-based pipelines with supervised, unsupervised or even self-supervised learning methodologies for cloud classification, but a hybridization of a Vision Transformer and a Diffusion Classifier was not yet tested either. Such a pipeline would be capable of extracting high-level features from cloud images by leveraging the unique strengths of both models, potentially leading to better performance and more robust classifications. Moreover, we plan to also integrate multimodal data fusion that was not yet addressed (existing approaches are mainly using satellite imagery).

To the best of our knowledge, no other approaches have currently tried severe weather identification and classification using either generative learning or multimodal data fusion. The topic addressed in the current proposal is of interest at the national level, as due to current climate variations there is an increased need to integrate ML-based automated tools for assisting Romanian operational meteorologists in weather-related tasks.

To summarize, the paper is aimed at answering the following research questions:

RQ1: Would a hybrid approach combining a ViT and a Diffusion Classifier be feasible for efficiently determining cloud morphology or meaningful cloud image features?

RQ2: Would the integration of RS data from different sources through a multimodal data fusion perspective be beneficial for a more accurate classification?

RQ3: To what extent does diffusion-based generative capability improve cloud classification performance through data augmentation and also provide other potential applications to the meteorological domain?

Firstly, for the RQ1, our proposed approach would include both a ViT and a Diffusion Classifier. Some of our ideas of implementing such a hybridization are detailed in the Methodology section. We believe that by combining these two models we could benefit from both of their advantages in building an efficient cloud classification scheme. More exactly, the ViT excels at extracting global contextual information in images through its self-attention mechanisms, while the Diffusion Classifier can refine representations and provide enhanced classification performance. This way, the final architecture could exhibit powerful classification performance, by providing improved representation features, while also proving its robustness to noise disturbances and especially including generative insights that enhance the decision-making process. Secondly, to answer RQ2, we would have to train the designed pipeline with instances from various data sources and conduct evaluations accordingly. More specifically, we aim to show that using remote sensing data collected in different contexts, formats or perspectives (e.g., ground and satellite-based imagery or radar-provided numerical products), could help a well-suited and configured hybrid model determine a better generalisation of cloud morphology and dynamics and provide even more precise class results. For addressing RQ3, we aim to devise a generative component within our proposal as well. One first use of this would be to synthetically create class instances due to the critical need of addressing the problem of class-imbalance of instances, which is present in some datasets. By performing such advanced data augmentation techniques to generate new data, we are confident that datasets would contain a better distribution of instances, which can then be used for training any ordinary classifier, including ours. Apart from that, the generator side of our designs can also have applications such as prediction of weather movements by performing slight modifications on latent space representations. From a practical perspective, the approaches targeted in the paper proposal could be helpful tools for assisting operational meteorologists in detecting and analysing severe weather phenomena.

The rest of the paper is structured as follows. Section 2 presents existing literature approaches on clouds classification from images. Section 3 describes the architectures and conceps we want to develop as a baseline approach and also as a further research into this subject. Afterwards, Section 4 explains the classification process a given test subset of images are going through, and presents the results obtained after evaluating the performance of both implemented models. Then, Section 5 details the importance and relevance of these results, especially when compared with the SOTA alternatives. Finally, Section 6 concludes by summing up all the important aspects we have discovered this far, and also present some other ideas of expanding and improving the current implementation.

2. Related work on clouds classification from images

Many research papers have approached the subject of weather prediction and classification of clouds using very diverse technical means, and in recent years there have been some novel ideas that bring very good results in certain contexts. Used learning mechanisms range from supervised to unsupervised, and the architectures consist of multiple structures with different purposes that rely on each other and provide the final results overall.

To begin with, the paper published by Luo et al. [10] described an architecture with various steps within. It receives ground-based images that suffer minor edge removals and class-adaptive “dark channel prior algorithm” to reduce haze and light disturbances caused by the sun, which helps preserve cloud edge details for thinner clouds and diminish overexposed areas for denser clouds. Then, after distortion filtering, re-sizing and standardisation, they are fed to a YOLOv8 based pre trained Deep Neural Network (DNN), which is made out of Backbone, Neck, Head and Loss components, and outputs a bounding box and the most probable class an image belongs to. Binary cross entropy and Distribution Focal Loss (DFL) with Complete Intersection over Union (CIoU) are used as Loss functions, and a slightly improved Stochastic Gradient Descent (SGD) as an optimizer. Additionally, image patches have finite element segmentation applied to them, which represents a technique that combines both the classical Normalised Red/Blue Ratio (NRBR) method for colour variations and k-means clustering for further separation of cloud and sky areas. This approach also represents the core novel idea of their study. Dataset is made out of 4000 images captured by an observation instrument at Yangbajing Observatory station in Tibet, that are then split in 4 categories, each one having roughly the same size: cirrus, clear sky, cumulus, and stratus. Results prove to be quite impressive, with the model approaching 100% accuracy overall during the training stage, but having lower results for the highly complex cumulus clouds. The researchers also observed local cloud fluctuations based on the seasons of the year or diurnal parts, for instance spring has more stratus clouds due to increased water evaporation, summer has more cumulus clouds due to convection processes and increased precipitation are more present then, while autumn and winter are rather dry and usually have clear sky.

Yousaf et al. [11] proposed the idea of developing an ML model that could have an accuracy for cloud classification as high as possible, while also finding solutions to interclass similarities and class imbalance problems found in the used datasets. Data used for the training stage is taken from Large-Scale Cloud Images Dataset for Meteorology Research (LSCIDMR), which it is claimed represents the first publicly available database and one of the most challenging ones, due to the aforementioned problems. It is made out of over 100,000 satellite images captured by the Himawari-8 satellite and divided into 11 classes [9]. A Convolutional Neural Network (CNN) and Residual Network (ResNet)-based model was proposed by He et al. [12]. It consists of 5 convolution blocks, each one having kernels of maximum 3x3 size, as the authors also took inspiration from the Visual Geometry Group (VGGNet) model [13]. Batch Normalisation (BN) inside each block and a separate Dropout layer were added, drastically diminishing the overfitting problem by providing some regularisation among data and randomly disabling neurons. Finally, after extracting visual characteristics, Global Average Pooling (GAP) is applied to reduce dimensions of resulting feature maps, which is then followed by a Fully Connected Layer (FCL) with softmax activation function for determining predicted classes. Data augmentation was also used for balancing the size of training classes, by applying simple colour and geometric transformations to some of the existing images, e.g. image flips and rotations and brightness and scale variations. The model was trained over the span of two snapshots of 100 epochs each, resulting in an accuracy of 97.25%.

Another study worthy to be mentioned comes from Togaçar et al. [14], who implemented a ShuffleNet-based model, which is a lightweight architecture with reduced number of trainable parameters, designed and optimised for mobile devices as well. Images used belong to Harvard University-provided CCSN database and also SWIMCAT-Ext, which is a new version of the SWIMCAT database. Instances belong to eleven and six classes respectively, and suffer from class imbalance. Therefore, Super-Resolution (SR) and Semantic Segmentation (SS) are performed, using a Residual Dense Network (RDN) and an Xception-based model. Additionally, Binary Sailfish Optimization (SFO) is used for feature selection. Classification is performed using multiple approaches, namely Linear Discriminant Analysis (LDA), Naïve Bayes, Softmax, and SVM methods. Many evaluations with different scenarios are carried out, the best accuracy being reached when using all aforementioned techniques and LDA as classifier, raising the value from 96.45% to 98.56%.

Dev et al. [15] proposed a texton-based texture classification model, which includes K-means clustering on computed feature filters, which results in a texton dictionary. S-filters (essentially rotation-invariant Gabor-like filters) are used in the convolution layers to combine colour and texture information. Singapore Whole-sky IMaging CATegories is used as a data source containing 784 patches belonging to 5 categories, captured around Singapore over the period of January 2013 to May 2014. After multiple evaluations, S-filters and hyperparameter tunings were tested, the model approached perfect classification accuracy, with a small exception for the veil-type clouds.

Vasylieva et al. [16] described a CNN to classify images captured by the NOAA-20 Visible Infrared Imaging Radiometer Suite satellite into 4 classes. Presented layers are convolution, max-pooling, followed by softmax activation. After a 20-epoch training session, accuracy resulted in 95% and 85% for training and validation data respectively, even though the model tends to make slight confusions between cirrus and stratocumulus cloud classes.

Some other interesting ideas include unsupervised models, which train on unlabeled datasets. The motives behind this are based on the fact that some very large image datasets lack labels, which would require a lot of time-consuming and possibly inefficient human experts work.

Geiss et al. [17] proposed a special type of self-supervised learning which makes use of a Siamese Neural Network (SNN). The two datasets which were used are Moderate Resolution Imaging Spectroradiometer (MODIS), a multispectral imager that orbits aboard NASA's EOS Terra satellite, and Advanced Baseline Imager (ABI), an imager aboard the GOES-17 satellite. After filtering out ill-fitting images, a total of 64.340 instances for MODIS and 91.077 for ABI are used for training. Testing data is used from Aqua and Terra MODIS data sources. The SNN is based on a ResNet-style encoder CNN and receives two batches of input data, for which simple data augmentations were also applied. Layer-Wise Adaptive Moments Optimizer for Batch Training (LAMB) is used for training, and Barlow Twins was employed as loss function, allowing a typically supervised SNN to handle an unsupervised task instead. The model was compared with other existing approaches for evaluations, outperforming them and proving that its accuracy well surpasses over 80%.

Kurihana et al. [18] came up with another idea of unsupervised learning. They used 789 GB worth of training multi-spectral images from the same dataset as the previously mentioned authors, namely MODIS. A deep Convolutional Autoencoder (CAE) is used to reduce dimensionality of images and learn relevant features, and the loss function represents the norm between original image versus the encoded-decoded variant. Loss function is a combination of 4 metrics, namely L1 and L2 loss, MSSIM (multi-scale structure similarity index) and a Sobel operator-based frequency norm error. Hierarchical agglomerative clustering (HAC) is used for grouping data points, and Adjusted Mutual Information score (AMI) is used as a measure of stability, leading to good performance results by observing displays of correlated distributions of physical variables for each detected label class.

3. The proposed approach

3.1. Vision Transformers (ViTs)

From a technical perspective, transformers are a complex type of Deep Neural Networks (DNN) that rely on the mechanism of attention [7]. They proved to perform better than other competitive alternatives such as Recurrent Neural Networks (RNNs) or CNNs in terms of computation time and efficiency. For an image processing task such as classification of clouds, this is guaranteed by its ability to capture both local and global relationships between patches, which constitute the original image. Figure 2 illustrates the architecture of a ViT model. The operation of a ViT can be divided into several crucial steps [8]. Firstly, the input image is divided into equal flattened patches, which are then papered into Patch Embeddings. Afterwards, a Positional Embedding is added, which is calculated using a specific scheme, e.g. sinusoidal encoding due to the smooth variation of sine and cosine functions and ability to distinguish position differences. Furthermore, the embeddings are fed into an encoder, where Multi-Head Self-Attention is employed. This performs multiple concatenated partial basic Self-Attention operations by involving three essential vectors, i.e., the patch-specific Queries (Q), attended patch-specific Keys

(K) and the actual information-holding Values (V). The operations can be formally represented as:

$$MSA = \bigcup(SA_1, SA_2, ..., SA_h);$$

$$SA_i(Q_i, K_i, V_i) = softmax(\frac{Q_i \cdot K_i^T}{\sqrt{d_k/h}})V_i,$$

where d_k is the embedding dimension, h represents the number of parts the matrices were divided into (each one having d_k/h dimensionality), and i is a specific part [8]. Finally, embeddings are processed through a Feed-Forward Network (FFN) and then a prepended globally-representative classification token (`[CLS]` token) is extracted as the final output.

3.2. Diffusion Classifiers

Diffusion Probabilistic Models [19] fundamentally differ from other existing approaches, e.g. Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). In essence, they effectively try to learn a bijection between original domain and pure noise space by incrementally performing noising and de-noising transformations to a given image, which are also called forward and reverse diffusion.

Firstly, the forward diffusion performs multiple slight noising operations to the original image, to altering each pixel to certain degree, which is dictated by the scheduler, which depends on timestep t and can be linear or cosine-based. The operation can be represented by the formula:

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

, where q is the forward process function, x_{t-1} and x_t are the input and the output at step t respectively, and N represents the normal distribution. Afterwards, the denoising process is performed through multiple timesteps instead of directly transforming pure noise into the real expected image, which is the case of GANs. For each step, the noise would be passed through an architecture that outputs the most likely noise which could be subtracted from the input to receive the best estimation of the real image. The reverse diffusion pipeline is typically a neural network, such as U-NET which makes use of self-attention mechanisms. The specific formula [20] of the reverse process function p is :

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}, \mu_\theta(x_t, t), \Sigma_\theta(x_t, t));$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t, t)),$$

where $\epsilon_\theta(x_t, t)$ represents the model's output, $\alpha_t = 1-\beta_t$, $\overline{\alpha_t} = \prod_{s=1}^{t}\alpha_s$, and μ_θ and Σ_θ represent the mean and variance respectively.

The Diffusion Classifier [20] relies on the same core principles, but instead of giving a generated image as output, it tries to predict the class y of the input image. It tries to minimise a certain loss function, such as cross-entropy, while also performing the same forward and reverse diffusions. This evaluation occurs at various noise levels by receiving a conditional class index within each downsampling and upsampling block while processing the noisy image xt at timestep t: p(y—xt). This repeated process makes the classifier robust to noisy images. Core principles of such a classifier are shown in Figure 3. In essence, the underlying formula for probability of having class ci from image x relies on the Bayes formula,

$$p_\theta(c_i|x) = \frac{p(c_i)p_\theta(x|c_i)}{\sum_j p(c_j)p_\theta(x|c_j)}$$

3.3. Multimodal remote sensing data fusion

Multimodal Data Fusion [21] combines different data streams with various characteristics to produce more helpful information. It integrates data of diverse dimensionality, resolution, and type (such as RS data) to achieve specific goals in multiple applications. Deep learning (DL) has been successfully applied to multimodal remote sensing data fusion, providing significant improvement compared with traditional methods. Li et al. [21] recently reviewed DL-based multimodal remote sensing data fusion methods. Remote sensing modalities include Panchromatic (Pan), Multispectral (MS), Hyperspectral (HS), Light detection and ranging (LiDAR), SAR, infrared, night time light, and satellite video data. The paper divided existing methods into two main groups: homogeneous fusion (e.g., pansharpening, spatiotemporal fusion) and heterogeneous fusion (e.g., remote sensing geospatial data). According to Li et al. [21], commonly used supervised models for DL-based homogeneous and heterogeneous fusion are Autoencoders (AEs), CNNs, GANs and ViT-based architectures.

3.4. Potential hybrid architectures

Using both a ViT and a Diffusion Classifier is a novel idea that is currently explored within the research community, without being employed for cloud image classification and related tasks yet. The hybridization involves designing multiple pipeline ideas for combining both mechanisms. Below we present three of our proposals for hybridisation, with an emphasis on the first two. One of our future aims is to implement two of the proposed architectures in order to comparatively assess their performance.

3.4.1. Conditional Diffusion via ViT Features. With this approach, the ViT would be used to extract high-level semantic features from the input images, which would then be used as conditioning information for the Diffusion model throughout the denoising process. Additionally, the Diffusion model can be used for both classification and generation of

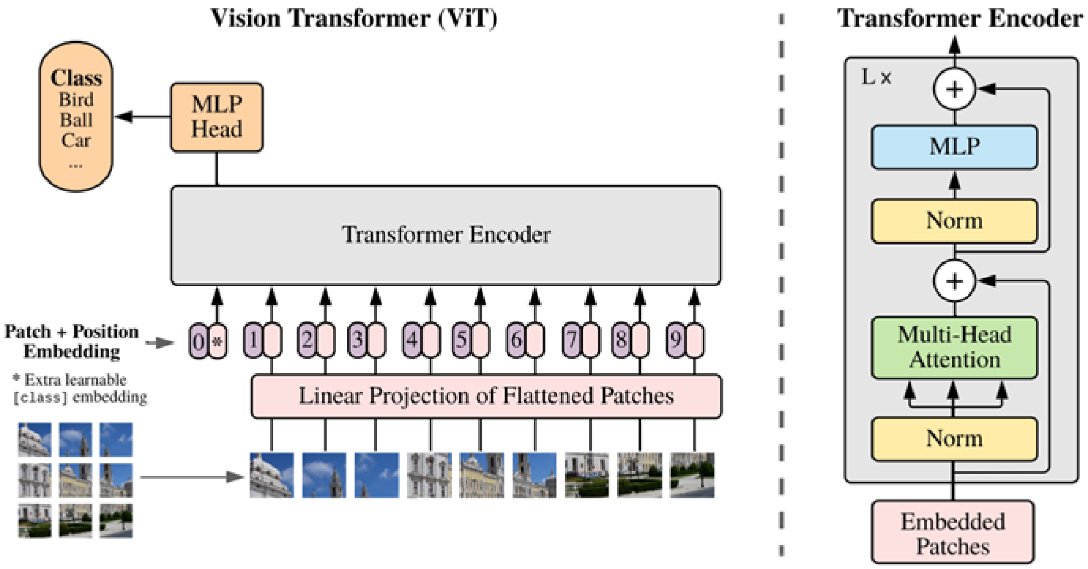

Figure 2. Architecture of the ViT model [8].

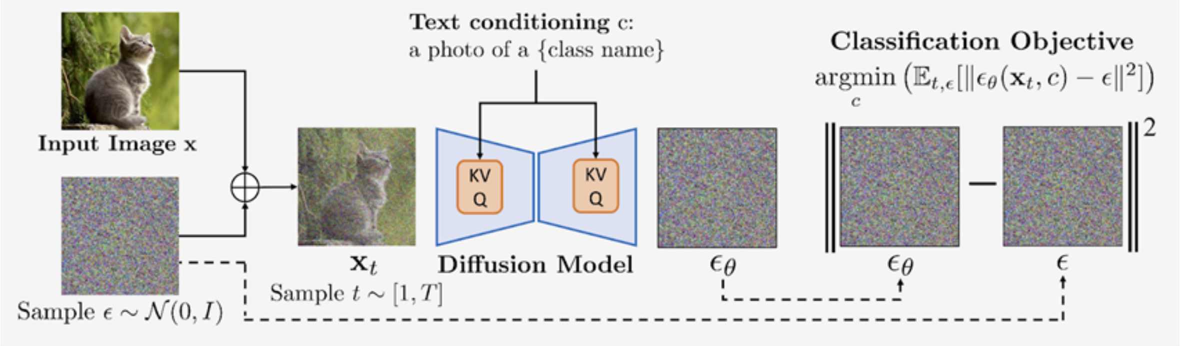
Input Image x
Sample $\epsilon \sim \mathcal{N}(0, I)$
$\mathbf{x}_t$
Sample $t \sim [1, T]$
Text conditioning c:
a photo of a {class name}
KV
Q
KV
Q
Diffusion Model
ϵ_θ
Classification Objective
$\underset{c}{\text{argmin}} \left(\mathbb{E}_{t,\epsilon}[\|\epsilon_\theta(\mathbf{x}_t, c) - \epsilon\|^2]\right)$
ϵ_θ
ϵ
2

Figure 3. A simplified overview of the Diffusion Classifier model [20].

cloud images. At each timestep, while progressively denoising the random noise back into an image within a neural network-based pipeline like U-Net, the ViT provided features are used as a conditioning value, by performing concatenation, feature modulation with the existing processed information, or by using them to control noise levels. Training of such an architecture can be split into two stages, first one only involving the pre-training of the ViT for image classification tasks. Afterwards, a Joint Training of both models could be performed, potentially giving better results by having a pre-trained ViT. This pipeline can ultimately be used for both classification and generation, and therefore the loss function could be computed to correct the models for both scenarios. Figure 4 illustrates a high-level design of this architecture.

3.4.2. Diffusion Transformer Hybrid. For this architecture design, the ViT model is essentially integrated within the Diffusion Classifier. It is responsible for providing the next step's prediction of noisy image representations at each timestep of the reverse diffusion process. The ViT tries to minimise the noise by applying self-attention. Consequently, it optimises its attention weights and feature extraction, ensuring that global context and semantic structure of the image are always preserved. Nevertheless, the overall noise schedule is still managed by the Diffusion model. This hybrid model uses loss functions for both denoising and classification, such as Denoising Loss (for both models), Self-Attention Loss (for the ViT), and Cross-Entropy Loss for classification. The total loss function would be a combination of all these losses, where the lambdas represent the importance of the corresponding loss:

$$L_{total} = L_{denoise} + \lambda_{classify} \cdot L_{classify} + \lambda_{attention} \cdot L_{attention}$$

3.4.3. Parallel Processing with Cross-Attention. Another possible design is that of the two models performing operations in parallel but interacting through a cross-attention mechanism. For instance,

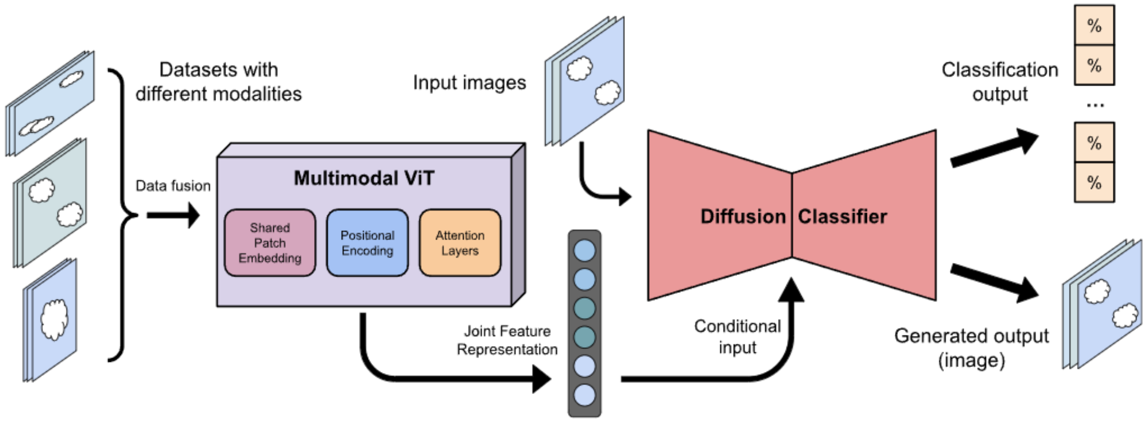

Figure 4. An overview of the proposed Conditional DiT Classifier architecture.

the ViT receives an image and captures global structure from it, while the Diffusion Classifier separately performs de-noising steps progressively. Afterwards, cross-attention is used to help one model attend to the features of the other model and vice-versa. This idea would match either for classification of images, where the final result can be computed by fusing the features from both models, or for generation, such that the diffusion model's generated image can adhere to the extracted features of the ViT, refining and improving image details.

4. Experimental validation

4.1. Datasets

Experiments will be performed on both public and real datasets. First, two publicly available datasets are targeted, CCSN and Ground-based Cloud Dataset (GCD), which differ in size, number of cloud classes and complexity. The CCSN dataset, previously used in the clouds classification literature, consists of 2543 images having a resolution of 256×256 pixels, divided into 11 classes [9]. The GCD dataset includes 19000 images having a resolution of 512×512 pixels captured in China with a total of 7 classes and the split of training and testing instances is 10000 and 9000 images respectively. Besides the publicly available datasets, we aim to use RS data such as satellite images provided by European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and radar data provided by the Romanian ANM. We will use satellite imagery from the Meteosat-11 satellite of EUMETSAT, a geostationary satellite orbiting Earth at 35.786 km above the Equator, providing a continuous stream of images (one image every 15 minutes) of the entire Earth disc, to monitor the weather over Europe and Africa. In addition to the satellite images, radar data collected by a Doppler single-polarisation radar located in central Romania is targeted. In a full volume scan, completed every 6 minutes, the radar outputs many different products related to the location, intensity and the movement of precipitating clouds and their associated meteorological phenomena. The most important radar products are the base Reflectivity (R) product and the base Velocity (V) product. The radars collect these base products at 9 elevation angles, gathering 9 sets of velocity or reflectivity data at each time step.

4.2. Case study on small data

The purpose of this study is to illustrate the methodology and potential of our approach to classifying cloud images, using a small selective set of instances. To present some preliminary results of model functionality, we fine-tuned a pre-trained U-Net model on a smaller subset of GCD dataset, and also run some evaluations with a significant ratio of the GCD for a ViT-based classifier. We will further describe the data set structure, along with pipeline elements and the steps each image instance went through.

4.2.1. Data Collection. For this study, we utilised a smaller and more simplified version of the GCD dataset (also called `mini-GCD`), which includes real cloud images into three categories: ”`4_clearsky`, ”`1_cumulus`” and ”`6_cumulonimbus`”. The dataset is divided into three subdirectories: *train*, *validation* and *test*; each one with its specific role.

4.2.2. Code implementation. Both the U-Net and the ViT based models were implemented in Python language, leveraging popular libraries such as `pytorch`, `torchvision`, `transformers` or `sklearn`. We also configured CUDA framework for enabling acceleration of the Graphics Processing Unit (GPU) at our disposal for achieving exponentially faster computations. As a baseline model implementation, we employed the principal and most commonly used component of a diffusion model, namely a pre-trained U-Net model on ImageNet.

	Precision	Recall	F1-score	Support
1_cumulus	0.68	1.00	0.81	15
4_clearsky	0.76	1.00	0.86	37
6_cumulonimbus	1.00	0.67	0.80	58
Accuracy			0.83	110

Table 1. TABLE OF CLASSIFICATION RESULT METRICS FOR THE TEST DATASET, FOR THE U-NET MODEL.

Additionally, we appended a classifier head to perform image classification, which was composed of a multilayer perceptron (MLP) with both linear and nonlinear activated fully connected layers, and regularisation techniques such as Dropout layers and Batch Normalisation.

4.2.3. Preliminary results. Train dataset was used for feeding the U-Net classifier during the training stage, and a smaller validation dataset was also used for analysing current loss of potentially new data given to the model and to address early stopping caused by the stagnation of this specific loss if necessary. Figure 5 displays the detailed losses of both training and validation data.

4.2.4. Performance Evaluation. To evaluate the model's performance, we selected a subset of images as the test dataset and calculated various performance metrics. A confusion matrix displaying the evaluation results is presented in Figure 6. The results indicate that the model achieves a decent performance in classifying cloud images, with an accuracy of 83%. However, there is still room for improvement, particularly in handling underrepresented classes. The precision, recall, and F1 score are also shown in the Table 1.

4.2.5. Discussion on sample data. We have randomly selected a sample of 1 batch containing 8 images to observe how does the model affect the instances. This subset of images was chosen to provide a comprehensive understanding of how the model processes and classifies cloud images. Each image in the sample underwent a series of transformations and computations as it passed through the various layers of the U-Net architecture.

Initially, the images were preprocessed using standard augmentation techniques such as resizing, normalization, and data augmentation to enhance the model's generalization capabilities. These preprocessed images were then fed into the U-Net model, which is composed of an encoder and a decoder path. The first path, consisting of a series of convolutional layers followed by max-pooling operations, progressively downsampled the images to extract high-level features. Simultaneously, the decoder path, comprising upsampling layers and skip connections, reconstructed the spatial dimensions of the images while incorporating the high-level features from the encoder. Throughout this process, the model applied mathematical formulas such as convolutions, activations, and pooling operations to transform the input data. For instance, the convolutional operation can be mathematically represented as:

$$Conv(I, K) = \Sigma_m \Sigma_n I(x - m, y - n) \cdot K(m, n)$$

where I is the input image, K is the convolutional kernel, and (x, y) are the pixel coordinates. The activation functions, such as ReLU, introduced non-linearity, enabling the model to learn complex patterns. The ReLU activation function is defined as:

$$ReLU(x) = max(0, x)$$

The pooling operations reduced the spatial dimensions, focusing on the most salient features. Max-pooling, for instance, can be described as:

$$MaxPool(I, s) = max_{i,j\epsilon s}(i, j)$$

where s is the pooling window size.

The intermediate feature maps generated by the U-Net model were visualized to provide insights into the feature extraction process. These feature maps highlighted the regions of the images that the model deemed significant for classification. Figure 7 displays such details, along with real labels and the ones predicted by the model with a certain probability.

4.3. Case study on real data

A subsequent evaluation of the proposed models was performed using the same GCD dataset, with the difference that all the 7 classes have been included in the training and validation process. Nevertheless, due to the sheer size of the dataset, we imposed a ratio limit to the number of instances, resulting in a couple of thousand instances overall.

For this scenario, we implemented the Vision Transformer classifier architecture to examine the performance of the another one of our presented ideas.

Despite the results produced by the previously presented and case studied variant (i.e. the U-Net model), the ViT-based model presented significantly better results, even approaching optimal results. More specifically, the accuracy which is obtained by determining whether an instance with a predicted cloud class label actually belongs to that class, reached 95%, while all the other metrics show similar good outcomes. The confusion matrix showing all the accuracy details for each class is presented in Figure 8. Moreover, Table 2 shows all the evaluation statistics for each class as well.

5. Discussion

5.1. Comparison to related work

Our proposed approaches—a diffusion/U-Net based cloud classification model and a Vision Transformer (ViT) classifier—offer a novel perspective

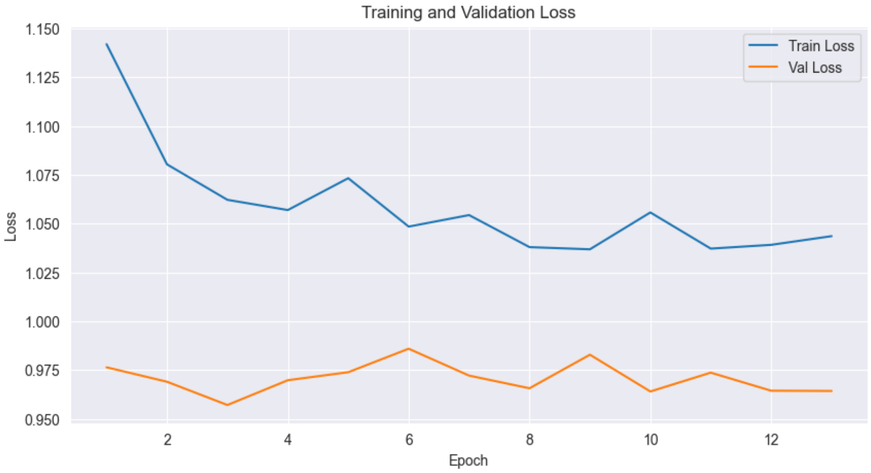

Figure 5. Evolution of training and validation losses during the fine-tuning period of the initial U-Net based classifier.

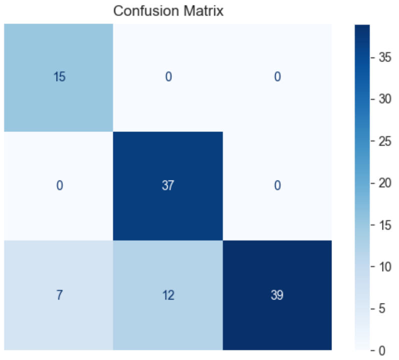

Figure 6. Confusion matrix of test data with 3 classes, as performed by the U-Net model.

	Precision	Recall	F1-score	Support
1_cumulus	0.94	1.00	0.97	15
2_altocumulus	1.00	1.00	1.00	15
3_cirrus	1.00	0.95	0.97	19
4_clearsky	1.00	1.00	1.00	37
5_stratocumulus	0.94	0.86	0.90	36
6_cumulonimbus	0.90	0.97	0.93	58
7_mixed	1.00	0.89	0.94	9
Accuracy			0.95	189

Table 2. Table of classification result metrics for the test dataset, for the ViT model.

on this challenging task. These architectures have been meticulously designed to address the intricacies of cloud classification, leveraging advanced deep learning techniques to achieve superior performance.

Our diffusion/U-Net based model, while not achieving the highest accuracy, demonstrates significant potential. The U-Net architecture, known for its effectiveness in segmentation tasks, provides a robust framework for feature extraction and classification. The U-Net model consists of an encoder-decoder structure with skip connections, which allows it to capture both high-level and low-level features effectively. This architecture is particularly well-suited for tasks that require detailed spatial information, such as cloud classification. Despite its current performance, the U-Net model has the potential to surpass state-of-the-art (S.O.T.A.) methods with further refinement and optimization. The integration of diffusion techniques enhances the model's ability to capture complex patterns and nuances in cloud images, making it a promising candidate for future advancements in this field.

The Vision Transformer (ViT) classifier, on the other hand, has shown remarkable performance, achieving near-perfect accuracy and excellent results across various metrics. The ViT architecture, with its transformer-based approach, excels in capturing global context and long-range dependencies in the data. This makes it particularly well-suited for cloud classification tasks, where understanding the broader spatial relationships within images is crucial. The ViT classifier's outstanding performance underscores its effectiveness and positions it as a leading contender among existing methods. The ViT model leverages self-attention mechanisms to weigh the importance of different parts of the image, allowing it to focus on the most relevant features for classification. This architectural advantage sets the ViT classifier apart from traditional convolutional neural networks (CNNs) and other models used in related work. Overall, our proposed architectures, especially the ViT classifier, demonstrate the potential to set new benchmarks in the field of cloud classification.

5.2. Addressing research questions

We are confident that the presented results prove that using innovative Machine Learning methods in the field of meteorology, such as Vision Transformers or generative-based approaches like diffusion architectures tailored to image classification, indeed represent promising perspectives for the future. We have shown have generative models can indeed make

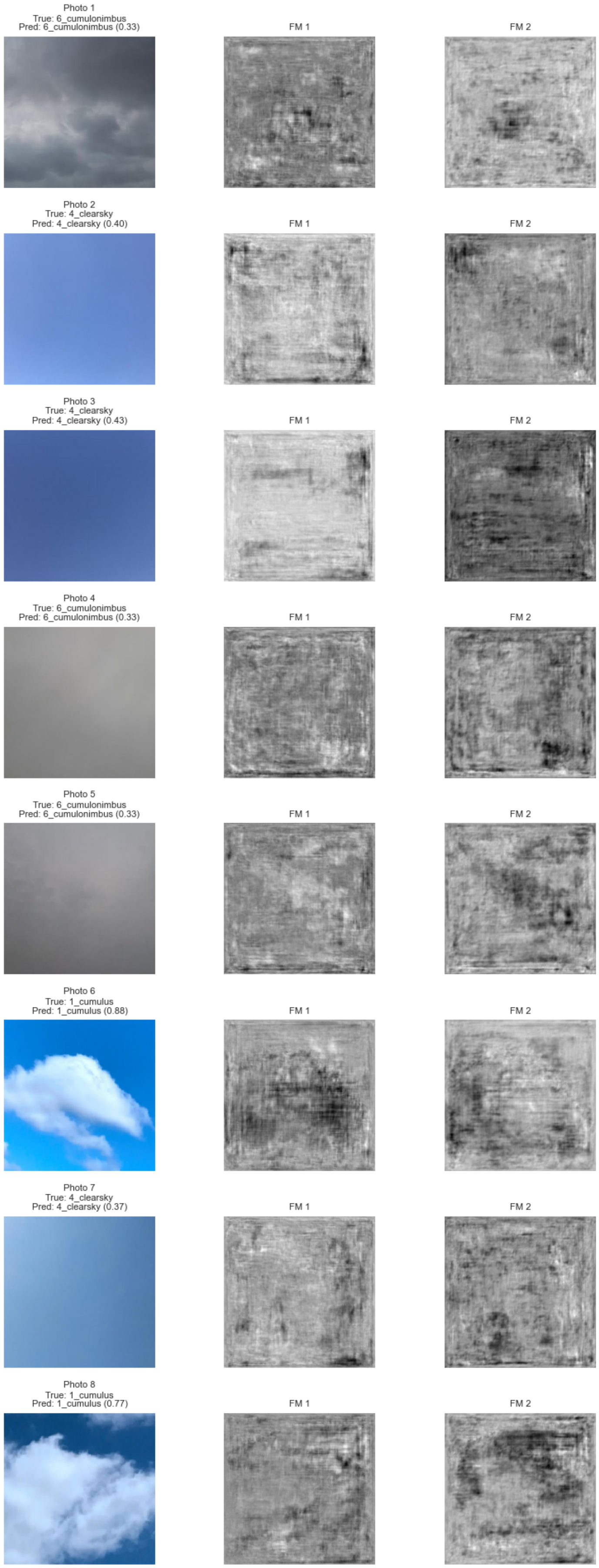

Figure 7. Corresponding feature maps of the sampled images, along with predicted and actual labels.

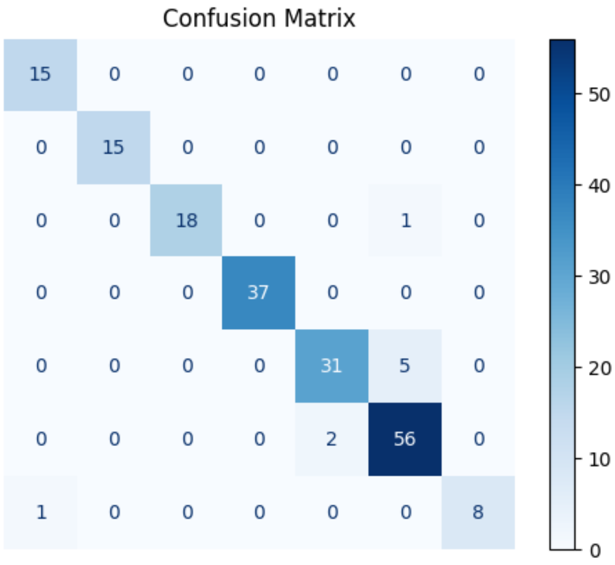

Figure 8. Confusion matrix of test data with all 7 classes, as performed by the ViT model.

some progress on learning morphological features and characteristics of different clouds, and especially the attention-based transformers seem to excel of such tasks, also being able to adapt to different types of datasets, if trained properly.

6. Conclusions and future work

Summing up, our illustrated concepts, representative for cutting-edge and large-scale solutions, could undoubtedly improve cloud classification, and therefore severe weather identification and classification, helping the society adapt in an ever-so-changing world in which modern ideas are necessary. The need for developing automatic ways of properly labelling weather phenomena without extensive and laborious human efforts is substantial.

As mentioned in the evaluation section, we ought to reiterate that the datasets were limited to a certain ratio (more exactly to 10% of their size) due to the very high time complexity which could have severely impacted the training and the research process. In the eventuality of having more advanced hardware at disposal, one could use the entire dataset for training the ViT model and most probably achieve even better results.

Another aspected that should be mentioned and which is highly related to the previous one is the problem of class imbalance. GCD suffers from having classes with significant differences in size (one class may have less than five hundred instances, while another one such as `6_cumulonimbus` has well over two thousand representative images). If not handled properly, this can lead to significant biases of the model, or even over-fitting on training data. In our study, we have already used some data augmentation techniques for trying to overcome this situation, but there are also other concepts worth studying such as generative augmentation using a diffusion model.

Presented models behave well on different types of datasets, especially for ground-based imagery. An interesting idea that we aim to study and develop is multimodality through the use of data fusion by incorporating multiple datasets made from different angles, points-of-view (such satellite images), or even non-image data (such as numerical data from radar stations).

Last but not least, generative learning has the potential to unlock new alternatives for meteorological-related problems. For instance, by using diffusion models we could generate images that resemble and share the characteristics of other images, or even use generative AI to modify, expand, or enrich images or datasets with similar content.

The Diffusion-based classifier and the Vision Transformer concepts we introduced and implemented in this paper represent just the baseline for our long-term research. We plan to further experiment on these models and develop multiple hybrid architectures that combine and make use of both of these models, thus using the best elements out of every one. We intent to employ the first two approaches we have detailed in a previous chapter, namely *Conditional Diffusion via ViT Features* and *Diffusion Transformer Hybrid*.

References

[1] G. L. Stephens, "Cloud feedbacks in the climate system: A critical review," *Journal of climate*, vol. 18, no. 2, pp. 237–273, 2005.

[2] J. A. Brotzge, J. Wang, C. Thorncroft, E. Joseph, N. Bain, N. Bassill, N. Farruggio, J. Freedman, K. Hemker Jr, D. Johnston *et al.*, "A technical overview of the new york state mesonet standard network," *Journal of Atmospheric and Oceanic Technology*, vol. 37, no. 10, pp. 1827–1845, 2020.

[3] V. H. Phung and E. J. Rhee, "A deep learning approach for classification of cloud image patches on small datasets," *Journal of information and communication convergence engineering*, vol. 16, no. 3, pp. 173–178, 2018.

[4] W. Ye, Y. Li, and D.-L. Zhang, "Generation of extreme precipitation over the southeastern tibetan plateau associated with tc rashmi (2008)," *Weather and Forecasting*, vol. 37, no. 12, pp. 2223–2238, 2022.

[5] J. L. Bytheway, M. R. Abel, R. Cifelli, K. Mahoney, and J. M. English, "Demonstrating a probabilistic quantitative precipitation estimate for evaluating precipitation forecasts in complex terrain," *Weather and Forecasting*, vol. 37, no. 1, pp. 45–64, 2022.

[6] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need study notes," in *NIPS 2017*, pp. 5998–6008.

[8] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR 2021*, 2021.

[9] W. M. Organization, "International Cloud Atlas," https://cloudatlas.wmo.int/en/home.html, 2022, [Online; accessed 13-December-2024].

[10] J. Luo, Y. Pan, D. Su, J. Zhong, L. Wu, W. Zhao, X. Hu, Z. Qi, D. Lu, and Y. Wang, "Innovative cloud quantification: deep learning classification and finite-sector clustering for ground-based all-sky imaging," *Atmospheric Measurement Techniques*, vol. 17, pp. 3765–3781, 06 2024.

[11] R. Yousaf, H. Z. U. Rehman, K. Khan, Z. H. Khan, A. Fazil, Z. Mahmood, S. M. Qaisar, and A. J. Siddiqui, "Satellite imagery-based cloud classification using deep learning," *Remote Sensing*, vol. 15, no. 23, p. 5597, 2023.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[14] M. Toğaçar and B. Ergen, "Classification of cloud images by using super resolution, semantic segmentation approaches and binary sailfish optimization method with deep learning model," *Computers and Electronics in Agriculture*, vol. 193, p. 106724, 2022.

[15] S. Dev, Y. H. Lee, and S. Winkler, "Categorization of cloud image patches using an improved texton-based approach," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 422–426.

[16] A. Vasylieva and R. Melnyk, "Classification of cloud types on satellite images using deep learning," *Manazerska Informatica*, vol. 3, pp. 1–13, 2022.

[17] A. Geiss, M. W. Christensen, A. C. Varble, T. Yuan, and H. Song, "Self-supervised cloud classification," *Artificial Intelligence for the Earth Systems*, vol. 3, no. 1, p. e230036, 2024.

[18] T. Kurihana, I. Foster, R. Willett, S. Jenkins, K. Koenig, R. Werman, R. B. Lourenco, C. Neo, and E. Moyer, "Cloud classification with unsupervised deep learning," *arXiv preprint arXiv:2209.15585*, 2022.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[20] A. C. Li, M. Prabhudesai, S. Duggal, E. Brown, and D. Pathak, "Your diffusion model is secretly a zero-shot classifier," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2206–2217.

[21] R. Bokade, A. Navato, R. Ouyang, X. Jin, C.-A. Chou, S. Ostadabbas, and A. V. Mueller, "A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing," *Expert Systems with Applications*, vol. 165, p. 113885, 2021.