Species Composition Prediction with High Spatial Resolution at Continental Scale Using Remote Sensing.

Industrial Internship report presentation..

Developed and presented by Yash Raj 202134

Department of Computer Science and Engineering,

Central University of Haryana , Mahendragarh

Species Composition Prediction

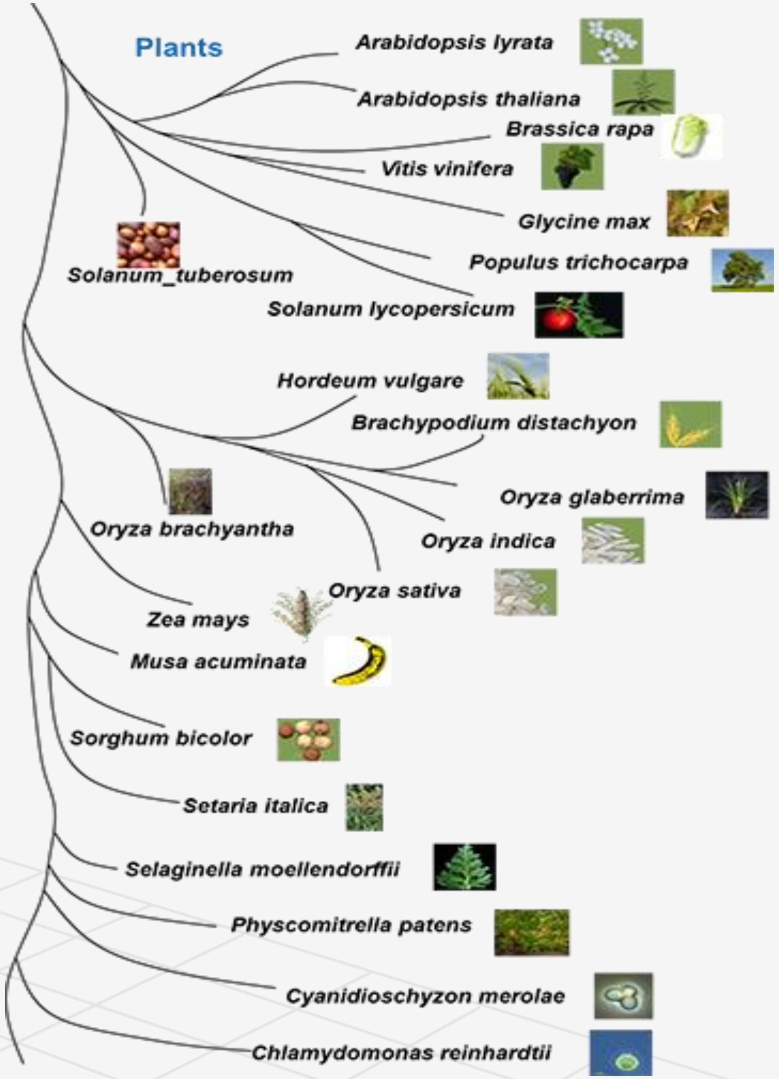
- It refers to the process of forecasting the variety and abundance of species within a particular ecosystem or habitat.
- This involves using data and models to estimate which species will be present and in what proportions.
- The prediction can be short-term or long-term and is crucial for various applications in ecology, conservation biology, and environmental management.

Plant species : Why to use Machine Learning and Deep learning?

- Millions of plant species.

- ML and DL are increasingly utilized for species composition prediction due to their ability to handle complex, high-dimensional data and to uncover patterns that might not be apparent through traditional statistical methods.

- Handling Complex Interactions.
- Improved Prediction Accuracy.
- Automation and Efficiency.
- Integration of Diverse Data Sources.



Contents

Abstract

Introduction

Objective

Methodology : Siamese neural network

Dataset

Modelling and evaluation

Transitioning to Modular Programming for Implementation

Challenges

Future Scope

Conclusion

Abstract

The "Location-based Species Presence Prediction" project aims to enhance species composition prediction using deep learning models and remote sensing data. By integrating 5 million plant species observations across Europe with various environmental datasets, the project developed models to predict species presence in 22,000 small plots. It uses a large-scale training set and a test set to improve biodiversity management and conservation efforts. A novel learning strategy was introduced to address biases in ecological modeling, leading to significant accuracy improvements. The project's outcomes aid in scientific understanding, conservation planning, policy-making, and education, supporting proactive biodiversity management and mitigating environmental impacts.

Introduction

The primary objective of the ″Location-based Species Presence Prediction″ project is to develop and fine-tune advanced machine learning models capable of accurately predicting plant species presence at specific locations and times, utilizing a diverse array of predictors such as satellite imagery, climatic time series, land cover, human footprint, bioclimatic, and soil variables.

Objective

To study large-scale plant species through various data modalities:

Landsat Cubes: Utilize satellite imagery data to capture the spectral information relevant to plant species.

Bioclimatic Cubes: Incorporate climate-related data, such as temperature and precipitation, to understand the environmental conditions affecting species distribution.

Sentinel Image Patches: Leverage high-resolution images to obtain detailed information about the habitat and local vegetation.

Environmental Rasters: Elevations , Human footprints , Soil grids, Climate rasters etc.

Objective

To develop a robust multimodal machine learning model that accurately predicts species distribution using diverse data sources:

Multimodal Integration: Combine data from Landsat, bioclimatic, and Sentinel sources to enhance the model's predictive capabilities.

Siamese Network Architecture: Employ a Siamese neural network to process each data modality with specialized encoders, and then integrate their outputs for final classification.

Performance Optimization: Experiment with various techniques such as data augmentation, mixup, and hyperparameter tuning to maximize model performance.

Objective

To convert the complete approach to modular programming for practical implementation and deployment:

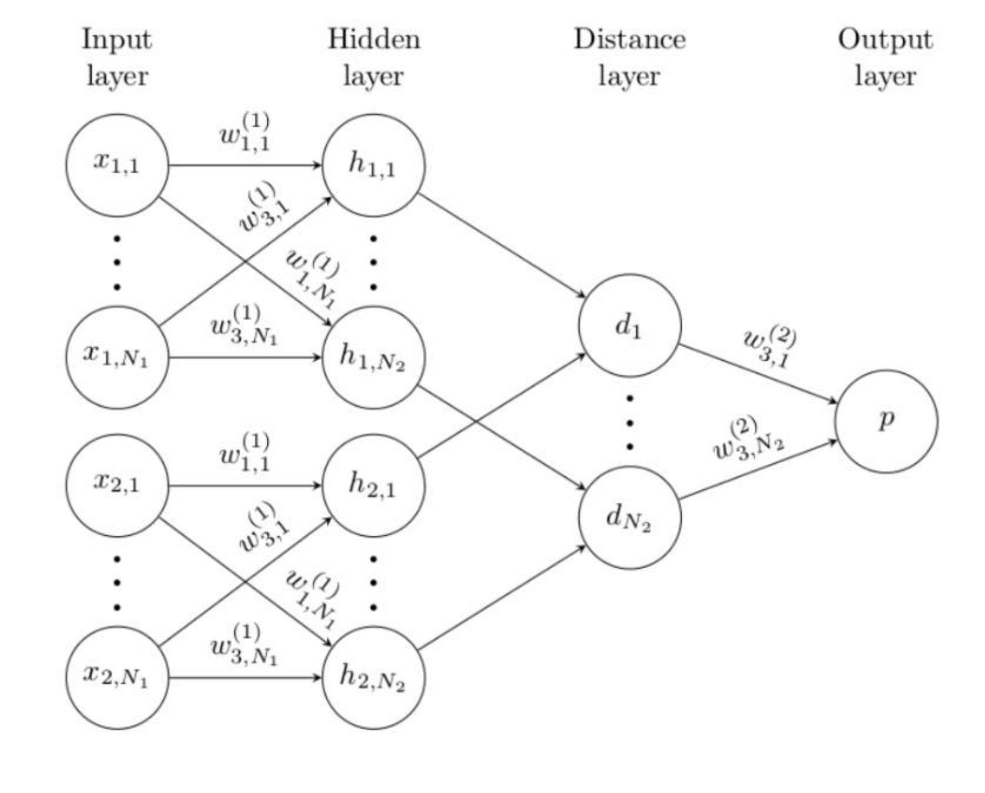
Modular Design: Break down the entire model into independent modules, including data preprocessing, model training, and evaluation, to facilitate easier maintenance and updates.

Scalability: Ensure the design can handle large-scale data efficiently, allowing for seamless integration of additional data sources or model enhancements.

Deployment Readiness: Prepare the model for real-world deployment by creating robust pipelines for data ingestion, model inference, and result visualization.

Methodology

The Siamese neural network serves as the cornerstone of our approach, facilitating the integration of multiple data modalities and enabling accurate predictions of plant species distribution.



The Siamese neural network methodology forms the backbone of our project, enabling us to effectively harness the power of multimodal data for accurate species distribution prediction.

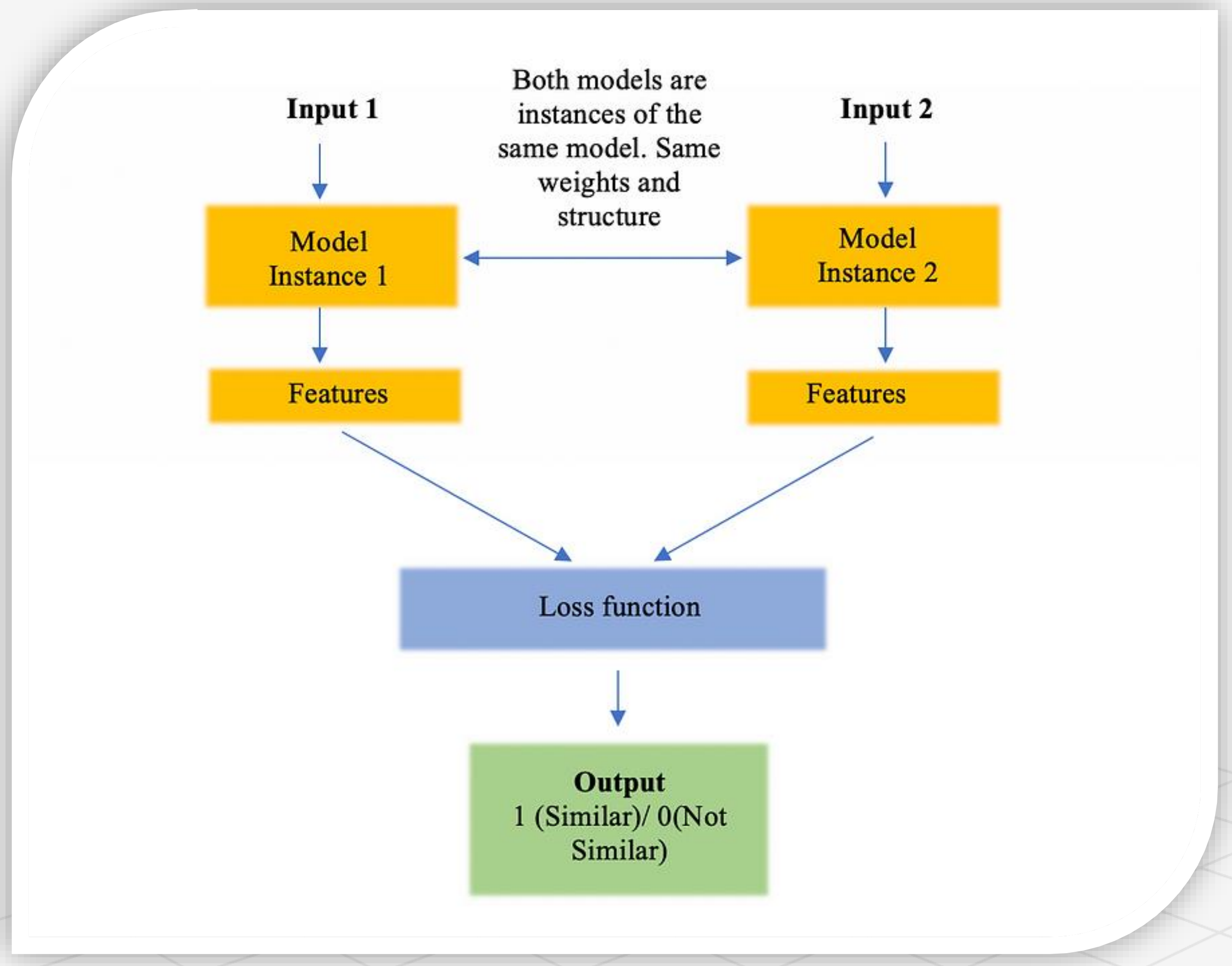
Methodology

What is a Siamese Network?

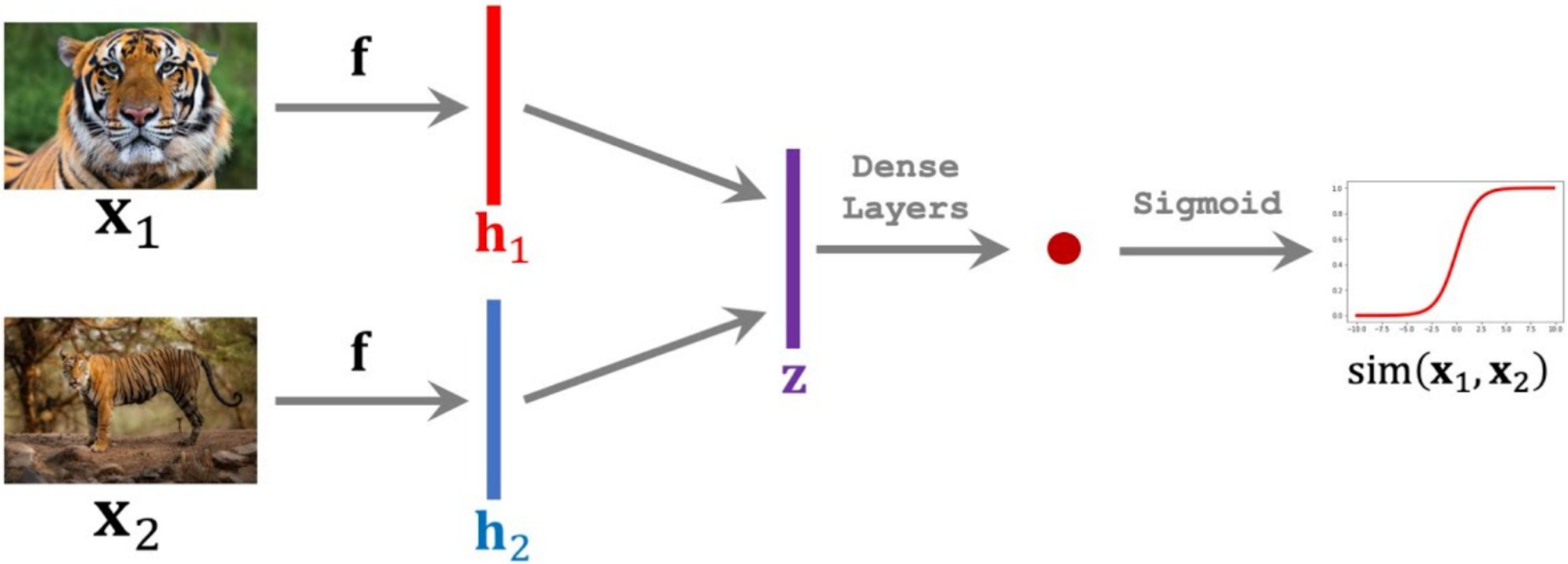
A Siamese network is a class of neural networks that contains one or more identical networks.

We feed a pair of inputs to these networks. Each network computes the features of one input.

And, then the similarity of features is computed using their difference or the dot product.

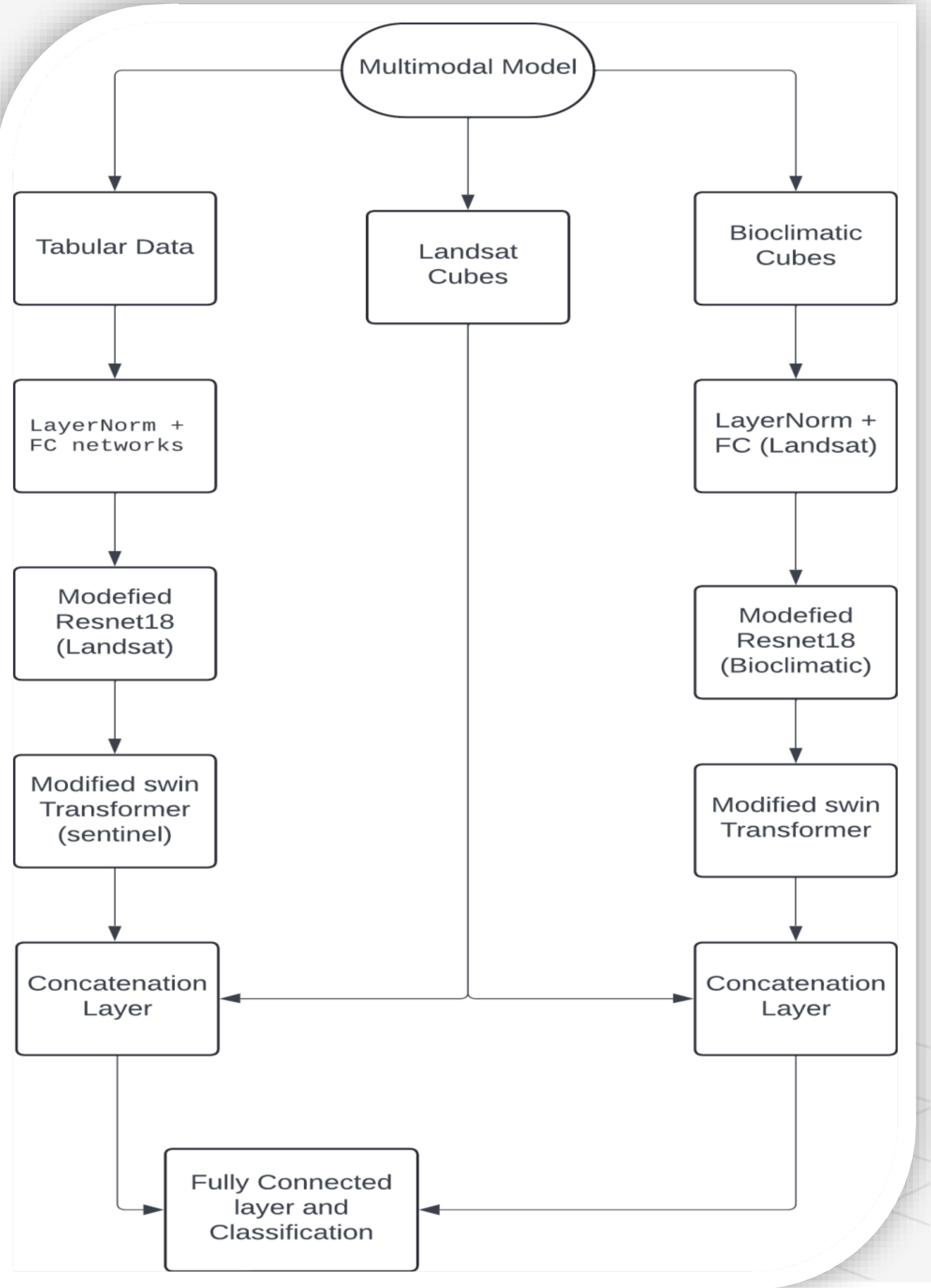

Input 1
Both models are instances of the same model. Same weights and structure
Input 2
Model Instance 1
Model Instance 2
Features
Features
Loss function
Output
1 (Similar)/ 0(Not Similar)

Siamese Network



Methodology

- The Multimodal Model utilizes the Siamese approach to handle inputs from various data modalities.
- Each modality, including Landsat cubes, Bioclimatic cubes, and Sentinel Image Patches, is processed separately by a distinct backbone or encoder.
- Encoders transform data into 1D vectors, which are concatenated and classified using a fully connected neural network.
- The model integrates information from multiple modalities, enhancing understanding and prediction of target classes.
- Processing steps include normalization, feature extraction, and fusion of diverse information sources to improve model performance.



Multimodal Model
Tabular Data
Landsat Cubes
Bioclimatic Cubes
LayerNorm + FC networks
LayerNorm + FC (Landsat)
Modefied Resnet18 (Landsat)
Modefied Resnet18 (Bioclimatic)
Modified swin Transformer (sentinel)
Modified swin Transformer
Concatenation Layer
Concatenation Layer
Fully Connected layer and Classification

Dataset

Species Observation Data:
Presence-Absence Surveys (PA): Approximately 90 thousand surveys covering 10,000 species of European flora. Used to address false-absences in presence-only data and calibrate models.
Presence-Only Occurrences (PO): Around five million observations from various datasets, spanning all countries within the study area. Opportunistic sampling led to varied biases. Absence of a species in PO data doesn't indicate true absence due to detection challenges, misidentification, or lack of interest.

Environmental Data:
Spatialized Geographic and Environmental Predictors:
Satellite Images: Four-band 128x128 images at 10-meter resolution around occurrence locations.
Time Series Data: Quarterly time series spanning over 20 years for six spectral bands at each location.
Raster Datasets: Various environmental raster datasets at European scale, including climatic, soil, land cover, human footprint, and elevation variables.
Monthly Climatic Variables: Monthly rasters of four climatic variables enabling extraction of time series data for any observation.

Dataset

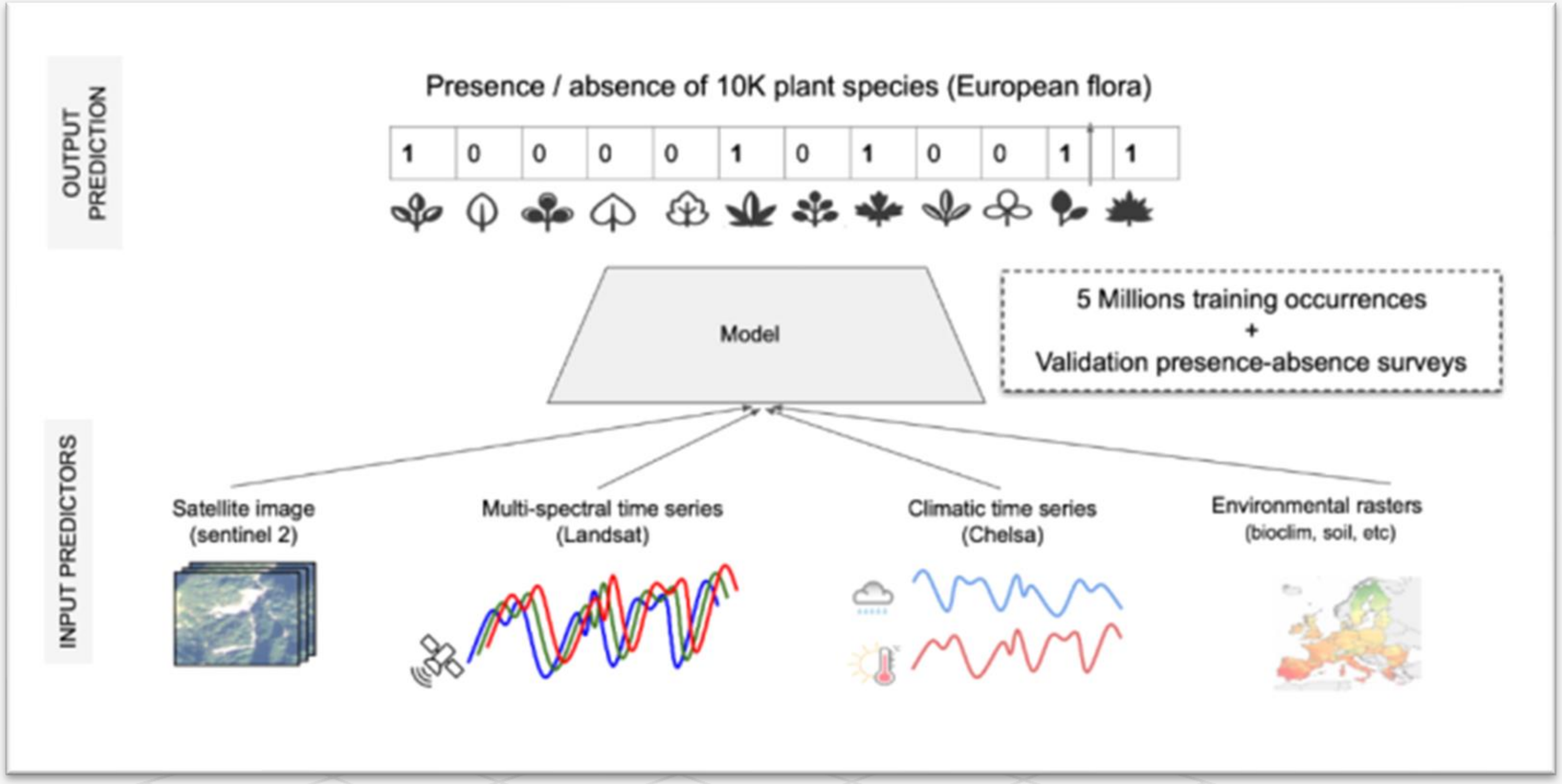
Standardized Biodiversity Observation Data: Presence-absence surveys conducted in small plots have limited spatial coverage and high renewal costs. This necessitates supplementing with crowdsourcing programs like Pl@ntNet and iNaturalist, which offer millions of precisely geolocated presence-only species records annually.

Challenges and Biases in Presence-Only Data: Presence-only records alone cannot indicate species absence, exhibit biases towards certain species, and only represent a fraction of species communities, introducing biases into species distribution models. Incorporating standardized presence-absence data can mitigate these biases but presents challenges in modeling due to strong class imbalance.

Integration of Environmental Data: Environmental data enriches the broader context but poses challenges in integration into traditional deep learning frameworks due to varying spatial resolutions. Carefully selected training data, including over 5 million presence-only records and approximately 5.9 thousand presence-absence surveys, were utilized for model training, calibration, and evaluation.

Dataset

Developing and assessing models for predicting the composition of species.


OUTPUT PREDICTION
Presence / absence of 10K plant species (European flora)
1 0 0 0 0 1 0 1 0 0 1 1
Model
5 Millions training occurrences
+
Validation presence-absence surveys
INPUT PREDICTORS
Satellite image (sentinel 2)
Multi-spectral time series (Landsat)
Climatic time series (Chelsa)
Environmental rasters (bioclim, soil, etc)

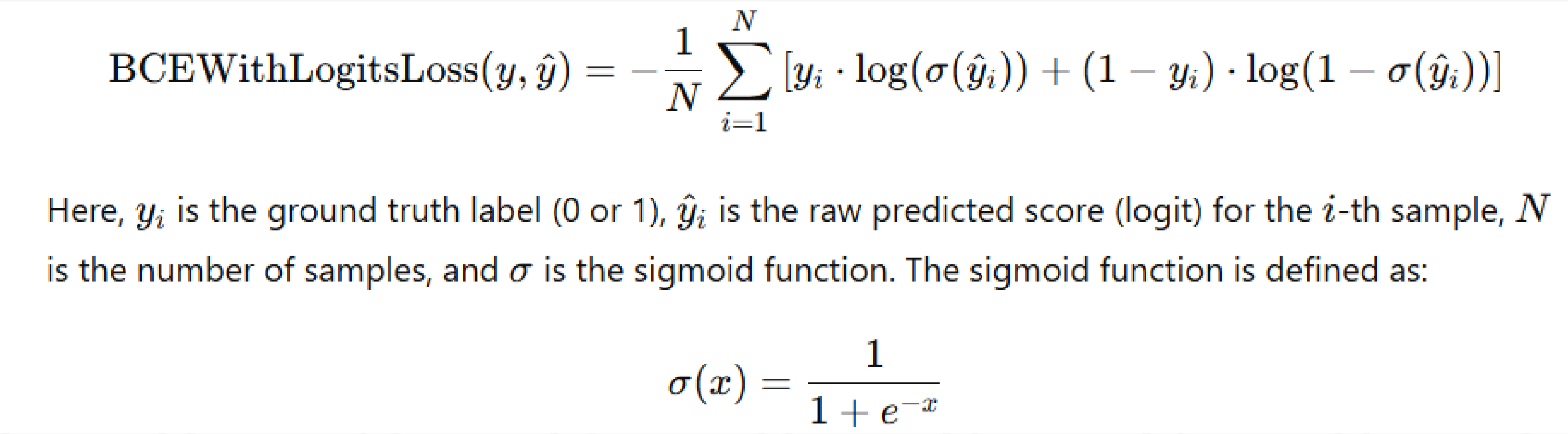
The main ambition is sought to create and assess models capable of forecasting ~10k plant species composition with high spatial resolution (approximately 10 meters) using various environmental predictors.

Modelling and Evaluation

Model Initialization

1. Define Optimizer: Utilize the AdamW optimizer for training the model, known for its effective handling of large-scale datasets and robustness against noisy gradients.

2. Define Loss Function: Employ the BCEWithLogitsLoss, a commonly used loss function for binary classification tasks, which efficiently combines a sigmoid activation function and binary cross-entropy loss.



$$\text{BCEWithLogitsLoss}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(\sigma(\hat{y}_i)) + (1 - y_i) \cdot \log(1 - \sigma(\hat{y}_i))\right]$$

Here, y_i is the ground truth label (0 or 1), $\hat{y}_i$ is the raw predicted score (logit) for the i-th sample, N is the number of samples, and σ is the sigmoid function. The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Modelling and Evaluation

Training Phase

1.For Each Epoch:

1. **Training Loop:**
   1. *For Each Batch of Training Data:*
      1. Apply mixup augmentation to enhance model generalization and robustness.
      2. Perform a forward pass through the model to compute predictions.
      3. Calculate the loss using the defined loss function.
      4. Conduct a backward pass to compute gradients and update model weights using the optimizer.
2. **Validation Loop (Every Few Epochs):**
   1. *For Each Batch of Validation Data:*
      1. Execute a forward pass through the model for validation.
      2. Compute the validation loss to assess model performance.
   2. *Track the best model based on validation loss for further analysis and evaluation.*

Modelling and Evaluation

Evaluation Phase

1. Load Best Model: Load the model with the lowest validation loss obtained during training to ensure optimal performance.

2. For Each Batch of Test Data:

1. Perform a forward pass through the model to generate predictions.
2. Compute predictions for each instance in the test dataset.

3. Post-process Predictions:

1. Sort predictions and select the top-k predictions for each instance based on confidence scores.
2. Compute the F1 score, a measure of model performance, considering both precision and recall.

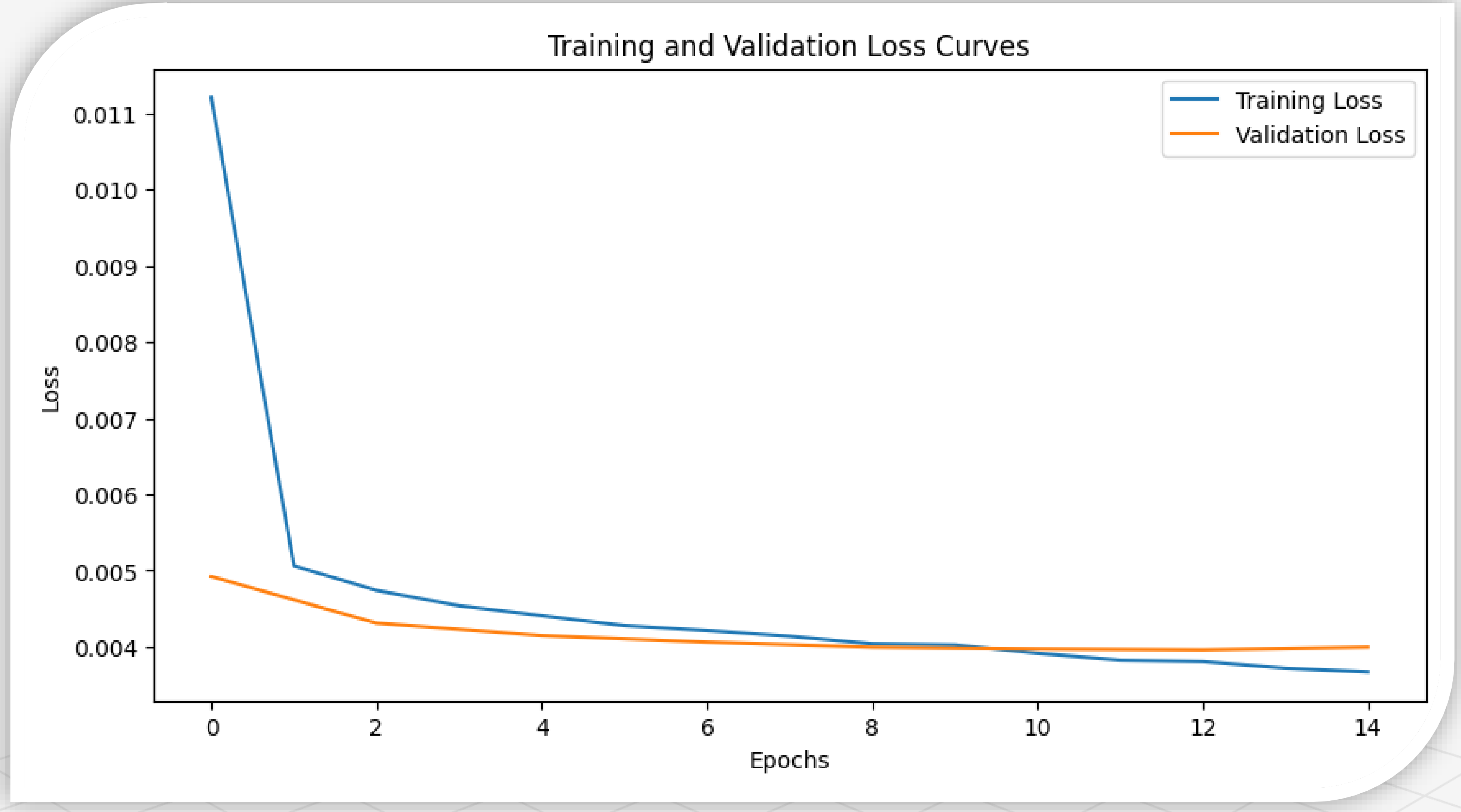
Modelling and Evaluation

Result Analysis

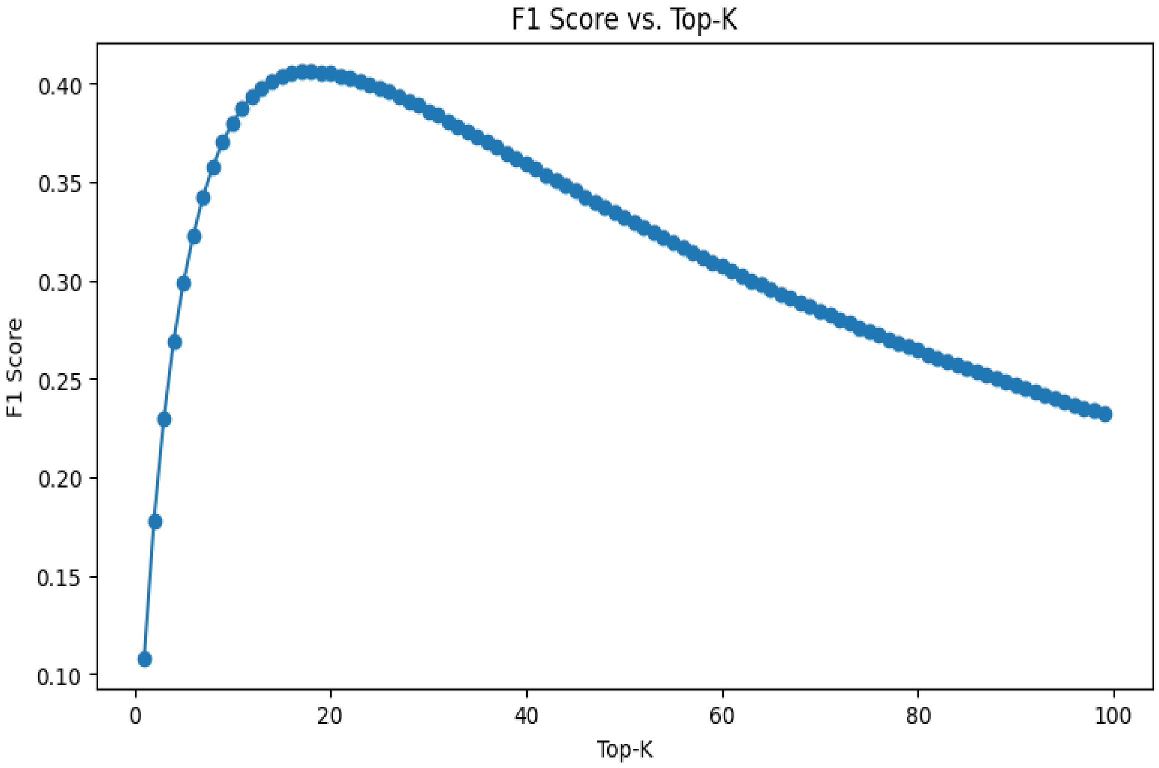
1.Plot Training and Validation Loss Curves: Visualize the training and validation loss curves over epochs to analyze the convergence and performance of the model during training.

1.Plot F1 Score vs. Top-k: Plot the F1 score against different top-k values to assess the model's ability to predict species distributions accurately across various thresholds.

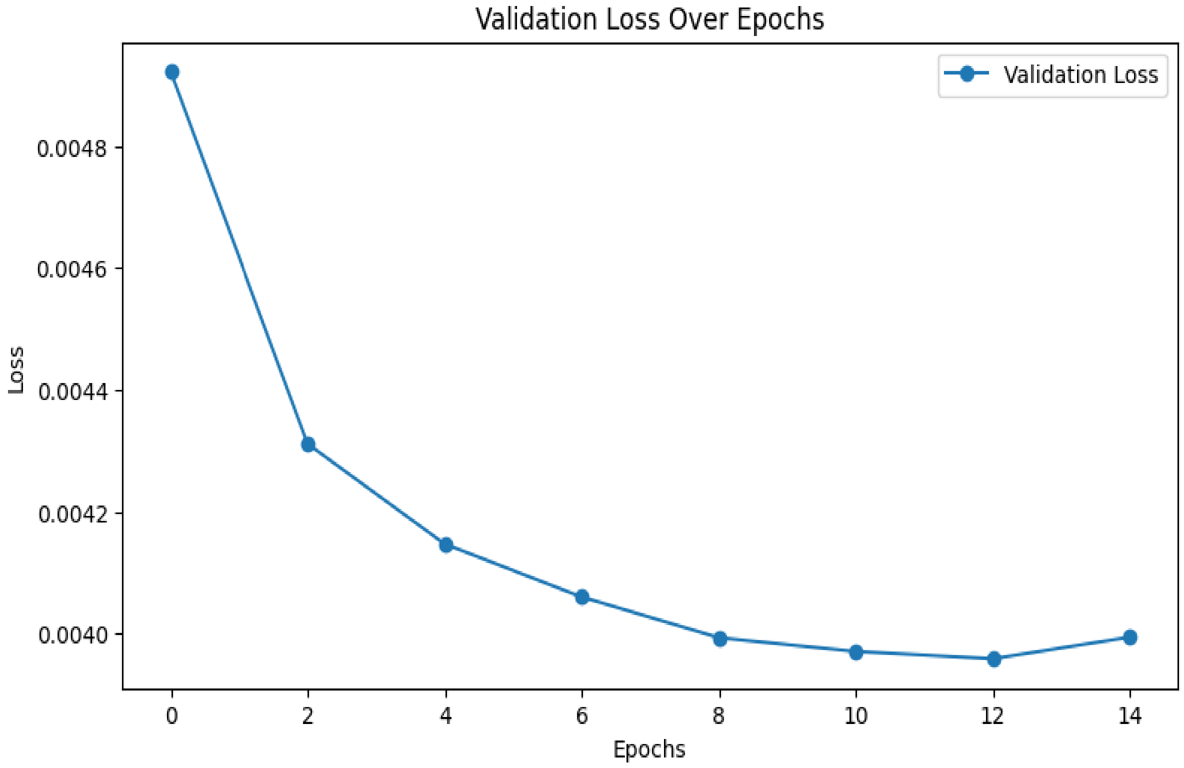
Modelling and Evaluation



Modelling and Evaluation



F1 Score vs. Top-K



Validation Loss Over Epochs

Modelling and Evaluation

Baseline Experiments Summary

Baseline with Bioclimatic Cubes:

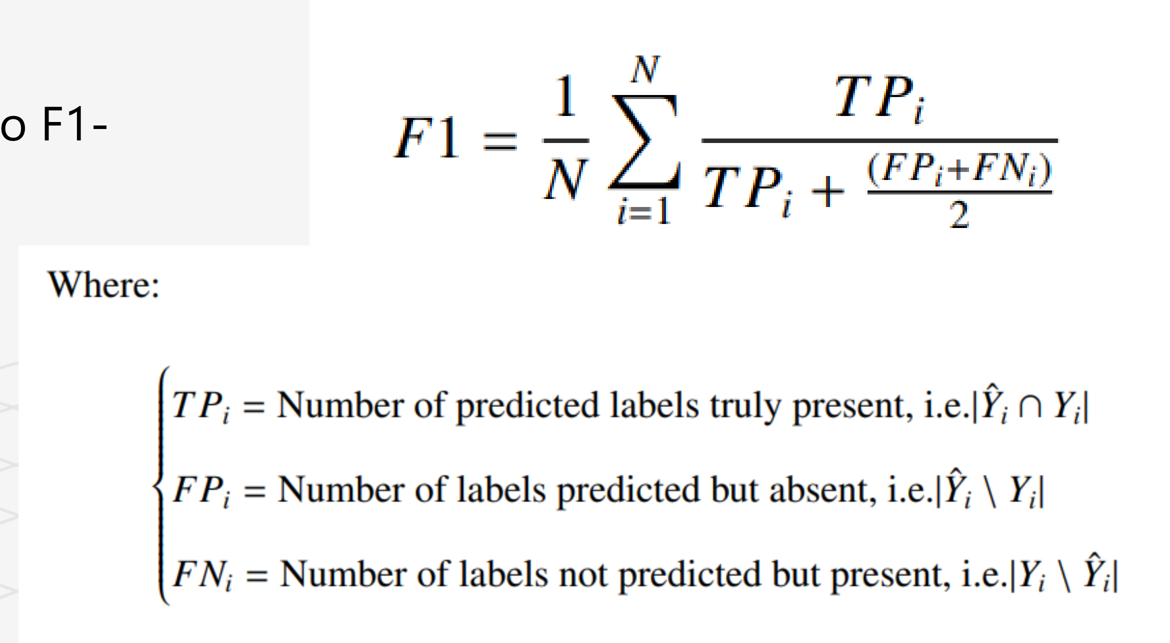
- **Methodology**: Utilized ResNet18 architecture with Binary Cross Entropy loss.
- **Performance Metric**: Achieved a score of [0.25784].
- **Insight**: Demonstrated the utilization of climatic history data to predict species composition.

Modelling and Evaluation

Evaluation metric

The technique used in this project was proposed as a multi-label classification task

The main evaluation metric for the project is the micro F1-score computed on the PA test set.



$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + \frac{(FP_i+FN_i)}{2}}$$

Where:

$$\begin{cases} TP_i = \text{Number of predicted labels truly present, i.e.}|\hat{Y}_i \cap Y_i| \\ FP_i = \text{Number of labels predicted but absent, i.e.}|\hat{Y}_i \setminus Y_i| \\ FN_i = \text{Number of labels not predicted but present, i.e.}|Y_i \setminus \hat{Y}_i| \end{cases}$$

Modelling and Evaluation

Baseline Experiments Summary

Baseline with Landsat Cubes :

- **Methodology**: Implemented ResNet18 architecture with Binary Cross Entropy loss.
- **Performance Metric**: Achieved a score of [0.26424].
- **Insight**: Explored the relationship between location values and species distribution using Landsat data..

Modelling and Evaluation

Baseline Experiments Summary

Baseline with Sentinel Image Patches :

- **Methodology**: Utilized Swin-v2-t architecture with Binary Cross Entropy loss.
- **Performance Metric**: Achieved a score of [0.23555].
- **Insight**: Demonstrated the potential of satellite imagery for capturing habitat characteristics relevant to species distribution.

Modelling and Evaluation

Baseline Experiments Summary

Baseline with Landsat + Bioclimatic Cubes + Sentinel images: (Combined)

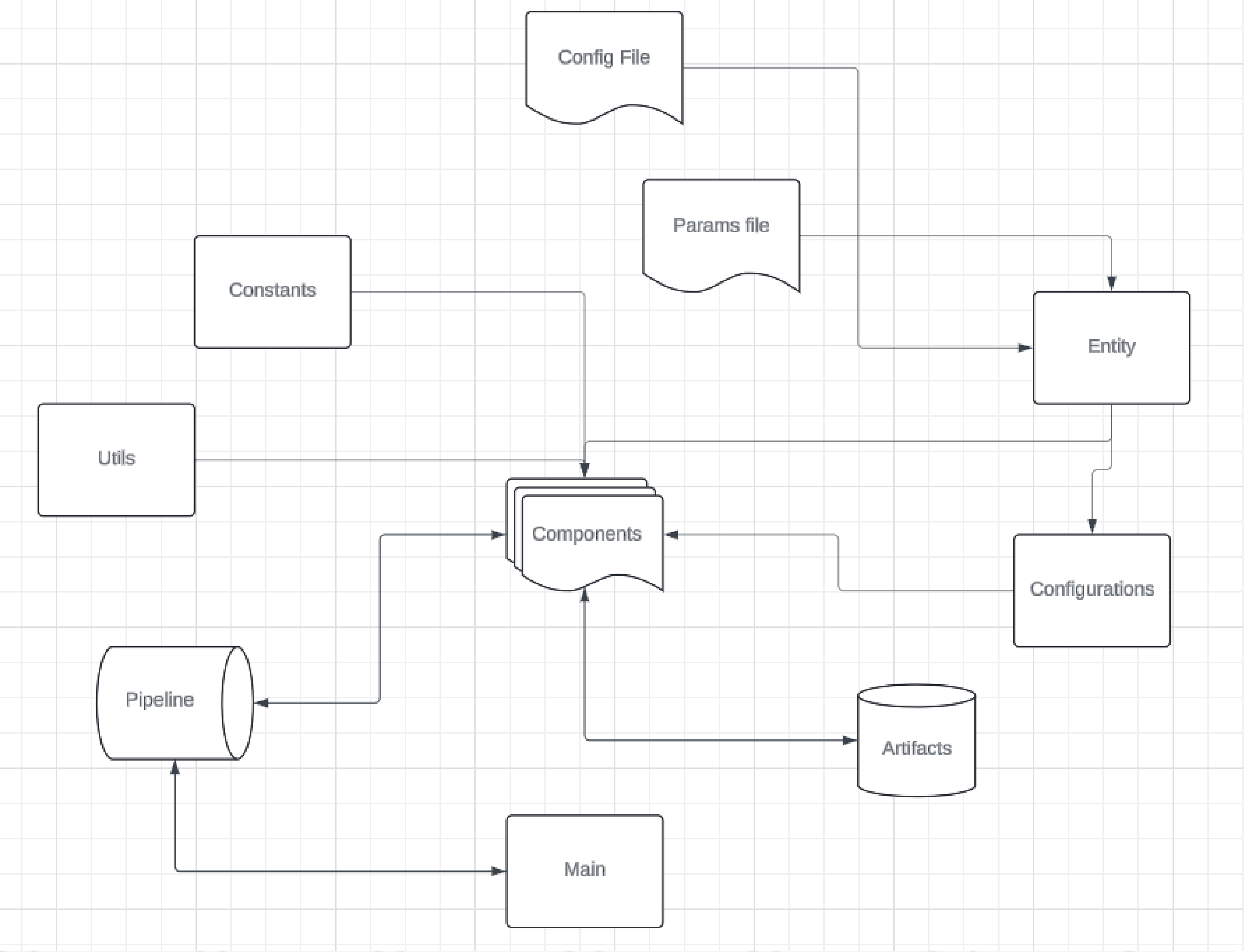
- **Methodology**: Implemented a Siamese Network approach.
- **Performance Metric**: Obtained a score of [0.31626].
- **Insight**: Integrated multiple data modalities, including Landsat and Bioclimatic cubes along with Sentinel images, showcasing improved performance through data fusion.

Transitioning to Modular Programming for Implementation

Code Restructuring: Modular programming involves breaking down the code into smaller units for better organization and scalability.

Modular Principles: Modules encapsulate specific tasks and adhere to principles like abstraction and single responsibility to ensure simplicity and clarity.

Interoperability and Collaboration: Standardized interfaces enable seamless communication between modules, fostering code reuse and facilitating collaboration among team members.



Challenges

The project faced several challenges in developing a robust model for predicting species distribution using multimodal data sources. These challenges included:

1. Multi-Label Learning from Single Positive Labels: Dealing with datasets containing single positive labels while requiring multi-label learning posed a significant challenge. Overcoming this required sophisticated techniques to ensure the model could predict multiple labels accurately.

2. Strong Class Imbalance: The dataset exhibited strong class imbalance, with some species being underrepresented. Addressing this issue was crucial to prevent bias towards more frequent classes and improve predictive accuracy for rare species.

3. Multi-Modal Learning: Integrating multiple data modalities introduced complexity due to their distinct characteristics and preprocessing requirements. Ensuring effective integration of diverse data types for coherent predictions presented a technical challenge.

4. Large-Scale Data Handling: Processing large-scale datasets, particularly high-dimensional satellite imagery and time-series data, required substantial computational resources and efficient data management strategies. This involved addressing challenges related to storage, memory management, and parallel processing capabilities.

Future Scope

- Refinement of Model Architecture.
- Integration of Additional Data Sources.
- Exploration of Advanced Techniques.
- Addressing Domain-Specific Challenges.
- Deployment and Integration.
- Collaboration and Knowledge Sharing.
- Continued Research and Development.

Conclusion

Objective Achievement: The project successfully realized its goals by developing a robust multimodal machine learning model and transitioning to a modular programming approach for implementation and deployment.

Data Integration: Diverse data sources, including Landsat cubes, bioclimatic cubes, and Sentinel image patches, were effectively integrated using a Siamese neural network architecture. This integration enhanced the model's understanding of species-environment relationships.

Overcoming Challenges: Despite facing technical challenges such as multi-label learning from single positive labels, strong class imbalance, and multi-modal learning, the project implemented innovative solutions to ensure the model's reliability and accuracy.

References

1. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

2. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

4. Brownlee, J. (2021). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning (Vol. 1). MIT press Cambridge.

6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

7. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Thank You !