

MSc Embedded Systems
Final Project

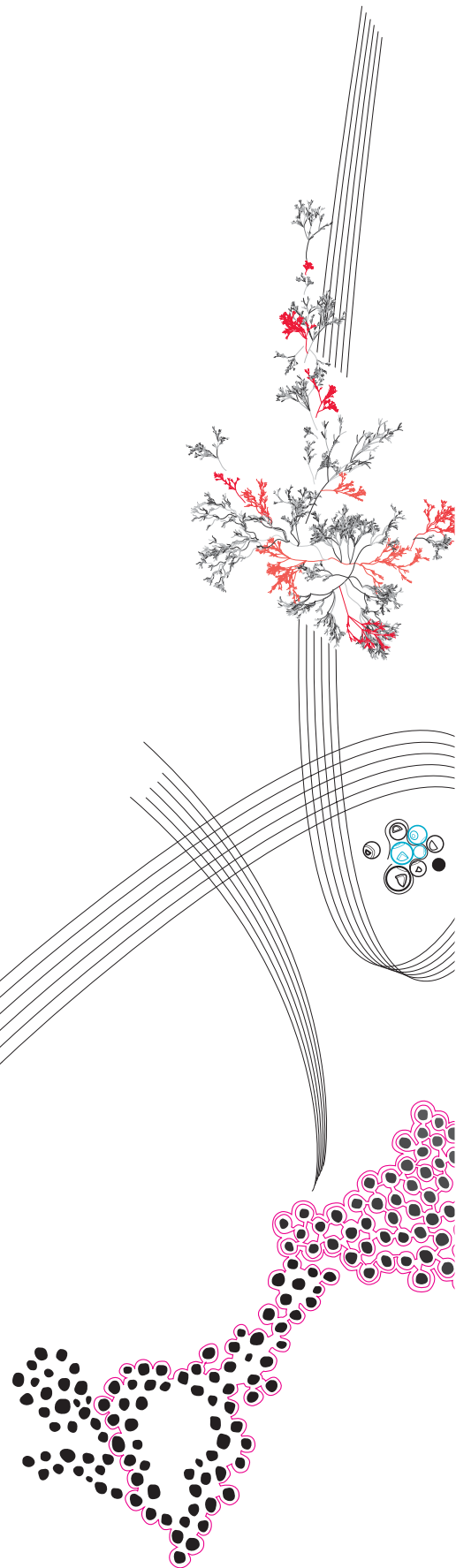
Species Distribution Modelling: A Multimodal Learning Approach

Pranesh Velmurugan
s2578050

Graduation Committee:
Prof.Dr.P.J.M.Havinga
Dr.A.Kamilaris
Dr.E.Talavera Martínez

October, 2023

Pervasive Systems,
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente.



Acknowledgements

I am filled with great pleasure as I reflect upon the completion of my thesis, focused on the development of a Species Distribution Model. This endeavor allowed me to pursue my passion for deep learning and computer vision with a profound commitment to environmental preservation.

First of all I would like to thank Pervasive Systems group for their support and assistance in helping me complete this thesis.

I extend my deepest appreciation to my supervisor, Andreas Kamilaris for his valuable feedback, guidance and suggestion throughout the thesis.

I also wish to convey my gratitude to my committee chair, Paul Havinga and external examiner, Estefanía Talavera Martínez for dedicating their time and evaluating my work.

A special thanks goes to Chirag Padubidri (CYENS Centre of Excellence) for his guidance and helping me from the inception of this thesis.

Lastly, I am thankful to my family and friends for providing me the strength and being there with me throughout my masters programme.

Have fun reading my thesis!

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation | 1 |
| 1.2 | Problem Statement | 2 |
| 1.3 | Research Objective | 2 |
| 1.4 | Thesis Outline | 3 |
| 2 | Scientific Background | 4 |
| 2.1 | Multimodal Learning | 4 |
| 2.2 | Evaluation Metrics | 4 |
| 2.2.1 | Area Under Curve (AUC) or Area Under Receiver Operating Characteristic Curve (AUROC) | 4 |
| 2.2.2 | True Skill Statistic (TSS) | 5 |
| 2.2.3 | Percentage Correctly Classified (PCC) | 5 |
| 2.2.4 | Top-k Accuracy | 5 |
| 2.3 | Activation Function | 6 |
| 3 | Related Works | 7 |
| 3.1 | Current state-of-the-art methods | 7 |
| 3.1.1 | Conventional Statistical Methods | 7 |
| 3.1.2 | Machine Learning Methods | 8 |
| 3.1.3 | Deep Learning Methods | 9 |
| 3.1.4 | Summary of Methods for SDMs | 10 |
| 3.2 | Multimodal Learning based SDMs | 12 |
| 3.3 | Graph Neural Networks | 15 |
| 3.4 | Research Gaps | 15 |
| 3.4.1 | Only limited works have been performed that uses pseudo-absence data in building a SDM | 15 |
| 3.4.2 | Feature importance assessment | 15 |
| 3.4.3 | Raw-dataset Balancing | 16 |
| 3.5 | Contributions of this thesis | 16 |
| 3.6 | Research Questions | 16 |
| 4 | Methodology | 18 |
| 4.1 | Dataset Description | 18 |
| 4.1.1 | Environmental Covariates (Predictor variables) | 18 |
| 4.1.2 | Frog Occurrence Dataset (Target Variable) | 19 |
| 4.2 | Pre-Processing Techniques Used on the Datasets | 20 |
| 4.2.1 | Grid Approach | 20 |
| 4.2.2 | Feature engineering | 20 |

| | | |
|----------|---|-----------|
| 4.2.3 | Dataset Balancing | 21 |
| 4.2.4 | Log Transformation | 25 |
| 4.2.5 | Image Augmentation | 25 |
| 4.2.6 | Other Basic Pre-processing Techniques | 26 |
| 4.3 | Pseudo-Absence dataset | 26 |
| 4.4 | Dataset for GNN Approach | 28 |
| 4.5 | Model Architecture | 30 |
| 4.5.1 | Model for Frog Counting Task - Regression | 31 |
| 4.5.2 | Model for Frog Presence/Absence Classification | 32 |
| 4.6 | Experimental Setup | 32 |
| 4.6.1 | Validation Metrics | 32 |
| 4.6.2 | Implementation details - Experiments Conducted | 34 |
| 5 | Results | 39 |
| 5.1 | Evaluation Metric | 39 |
| 5.1.1 | Mean Absolute Error (MAE) for Frog Counting Task | 39 |
| 5.1.2 | Accuracy and AU-ROC score for Presence/Absence Classification | 40 |
| 5.2 | Experiment Results | 40 |
| 5.2.1 | Balancing Dataset | 40 |
| 5.2.2 | Performance Comparison | 40 |
| 5.2.3 | Feature Importance Assessment | 43 |
| 5.2.4 | Generalization | 45 |
| 5.2.5 | Performance Evaluation of the pre-processed terraclimate data using XGBoost | 45 |
| 5.2.6 | Presence / Absence Classification | 46 |
| 5.2.7 | Comparison of different pseudo-absence data generation method | 47 |
| 6 | Discussion | 48 |
| 6.1 | Results - Analyses and Discussion | 48 |
| 6.1.1 | Dataset Balancing | 48 |
| 6.1.2 | Performance Comparison | 48 |
| 6.1.3 | Feature Importance Assessment | 50 |
| 6.1.4 | Generalization | 52 |
| 6.1.5 | Performance Evaluation using XGBoost | 53 |
| 6.1.6 | Presence / Absence Classification | 54 |
| 6.1.7 | Comparison of different pseudo-absence generation methods | 55 |
| 6.2 | Addressing the Research Questions | 55 |
| 6.3 | Limitations | 57 |
| 7 | Conclusion and Future Work | 58 |
| 7.1 | Conclusion | 58 |
| 7.2 | Future Work | 58 |
| A | Appendix | 68 |
| A.1 | Cross Validation Results | 68 |
| A.2 | Balancing Approach - K means Clustering | 68 |
| A.3 | Terraclimate Variables - Correlation matrix | 69 |

List of Figures

| | | |
|------|--|----|
| 2.1 | ReLU Activation [61] | 6 |
| 2.2 | Sigmoid Activation [61] | 6 |
| 3.1 | Multimodal Analysis - Generic Architecture [60] | 13 |
| 3.2 | Comparison [77] | 13 |
| 4.1 | RGB Patch- Costa Rica | 19 |
| 4.2 | Grid Formation for Australia | 21 |
| 4.3 | Presence points comparison | 22 |
| 4.4 | Histogram- Balanced Frequency of Frog Counts | 24 |
| 4.5 | Histogram- Frequency of Frog Counts | 24 |
| 4.6 | Histogram of Frog count before and after log transformation | 26 |
| 4.7 | Missing Values - Terraclimate | 26 |
| 4.8 | Frog Count - Outliers | 27 |
| 4.9 | Pseudo-Absence Data Generation | 28 |
| 4.10 | Australia Pseudo Absence point. Red - presence point, Grey - Pseudo Absence point | 29 |
| 4.11 | South Africa Pseudo Absence point. Red - presence point, Grey - Pseudo Absence point | 29 |
| 4.12 | Costa Rica Pseudo Absence point. Red - presence point, Violet - Pseudo Absence point | 29 |
| 4.13 | Pseudo Absence Points Comparison | 29 |
| 4.14 | Edge List - Sample Points | 30 |
| 4.15 | Pipeline Architecture - Frog Counting Challenge | 31 |
| 4.16 | Pipeline Architecture - Frog Presence/Absence Classification | 32 |
| 4.17 | Sliding Window | 34 |
| 4.18 | Weighted Average Ensemble | 35 |
| 5.1 | Train and Test MAE curves for Imbalanced data | 41 |
| 5.2 | Train and Test MAE curves for balanced data | 41 |
| 5.3 | Sample patch that shows the distinction between RGB, LC and NDVI data of the same location | 42 |
| 5.4 | Results - Comparison of various methods | 43 |
| 5.5 | Feature Importance - Histogram | 44 |
| 5.6 | AU-ROC Curve for LC & Numeric data | 46 |
| 5.7 | AU-ROC Curve for NDVI & Numeric data | 47 |
| 6.1 | Results - Model-W | 48 |
| 6.2 | Scatter Plot on submission data | 49 |
| 6.3 | Frog Counting Challenge - Leader Board Scores | 50 |
| 6.4 | Results - MAE obtained for Lower value of Frog count | 51 |

| | | |
|------|--|----|
| 6.5 | Scatter Plot on submission data for lower value of frog count | 52 |
| 6.6 | Tmax and Tmin of Australia and Costarica | 53 |
| 6.7 | Mean Precipitation of Australia and Costarica | 53 |
| 6.8 | Sample Land Cover patches of Australia and Costarica | 53 |
| 6.9 | Sample Presence and Absence Points of South Africa | 54 |
| 6.10 | Tmax and Tmin of presence and absence point - South Africa | 55 |
| 6.11 | Accumulated Precipitation of presence and absence point - South Africa . . | 55 |
| A.1 | K-means Clustering - Feature Space | 69 |
| A.2 | Heatmap - Correlation matrix of Terraclimate variables | 70 |

List of Tables

| | | |
|------|--|----|
| 3.1 | Performance Comparison of Existing State-of-the-art Models. AUC- Area Under Curve, PCC- Percentage Correctly Classified, TSS- True Skill Statistics, RMSE- Root Mean Squared Error, A10%DQ- Accuracy on 10% Densest Quadrats | 11 |
| 3.2 | Performance Comparison of Existing Multimodal Models. *- Not reliable due to a bug. | 14 |
| 4.1 | Overview of datasets used. | 20 |
| 4.2 | Overview of Training Parameters. | 33 |
| 5.1 | Results - MAE Comparison between Balanced and Imbalanced Dataset . . | 40 |
| 5.2 | Train and Test MAE obtained for the frog counting task on three sets of input data and the MAE obtained using sliding window and resizing approach on the submission data | 41 |
| 5.3 | Results - After Log Transformation | 42 |
| 5.4 | Results - Weighted Average Ensemble | 43 |
| 5.5 | Terraclimate Feature Importance | 43 |
| 5.6 | Results - MAE for top 6 features | 44 |
| 5.7 | Results on Costarica using the model trained on Australian data | 45 |
| 5.8 | Results of Ensemble method on Costarica using the model trained on Australian data | 45 |
| 5.9 | Results of XGBoost model trained on terraclimate dataset | 46 |
| 5.10 | Classification Accuracy for Frog Presence / Absence detection | 46 |
| 5.11 | Pseudo-absence data performance comparison | 47 |
| 6.1 | F1-Score of Ensemble model | 50 |
| 6.2 | Comparison of Model-W and Proposed model on Test data | 50 |
| 6.3 | MAE obtained on Submission data for lower values of frog count | 51 |
| 6.4 | Two Comparison Approaches - Overview | 51 |
| 6.5 | Comparison of XGBoost and Fusion model | 54 |
| 6.6 | Training Time for models using different modalities of data | 57 |
| A.1 | Cross Validation Results. * Not Completed | 68 |

Abstract

A species distribution model (SDM) makes use of environmental factors at a location to predict whether one species, or potentially several, will be present there. Some SDMs could even predict the count of the species present there. This work aims to predict the count of *Anura* (frogs) present in a location by building a SDM based on multiple modalities of data. Eventually, these will provide us valuable information about the condition of the environment. While there has been many methods of building SDMs, this work specifically aims to build a SDM based on Multimodal Learning which takes as input environmental features not just from one data type but from multiple modalities to predict the presence of the species. This work describes the proposed architecture and evaluates the results obtained from the model. Moreover, this paper compares the results obtained from the proposed model to the existing State-of-the-art methods.

The fusion architecture proposed in this model makes use of both tabular data and satellite image data. The results evaluated are compared with the winner of the frog counting challenge [6]. According to the leader board of the challenge, the proposed work achieved a F1-score of 0.36, which is placed second, and the winner of the challenge achieved a score of 0.42.

Apart from the task of counting frogs, this work also performs the classification of a location as presence / absence. The best performing model achieved an accuracy of 89.19% and an AUC score of 0.96. Though there are no direct comparison available for this task, still the results are on par with the existing classification SDMs.

For the task of classifying the location as presence/absence, a novel method of generating pseudo-absence dataset has been presented and is compared with some of the existing methods. The proposed method performed better than the distance criteria method by almost 4% better accuracy and by 19% better accuracy than the random selection method.

Overall, this work provides ways to use multiple modalities of data in building a SDM and suggests ways to improve the performance further.

Keywords: Species distribution model, multimodal learning, covariates, pseudo-absence data.

Chapter 1

Introduction

This section has been divided into motivation for this thesis, problem statement and Research objective of this work.

1.1 Motivation

Considering the importance of ecological research in preserving our environment and also to understand climate change, it is vital for the researchers to have a tool to monitor the distribution of flora and fauna. This provides them with valuable information regarding the consequences faced by animals and plants due to the changes in climate, pollution caused due to human activities. These tools also play an important role in preserving and protecting endangered species. There has been a number of citizen science projects where researchers and volunteers participate together to gather data on various species and their occurrences. Though these projects already provide the researchers with valuable information regarding the Geo-location of the species, it cannot be used to make predictions on the species occurrence in regions that have not been evaluated or recorded by citizen scientists.

To address this issue, the concept of Species Distribution Modelling has been introduced. Species Distribution Models can be defined as "*a quantitative , empirical models of species-environment relationships developed using geo-location of species data and the environmental features that affect those species distributions. The methodology or techniques used to develop such Species Distribution Models are called Species Distribution Modelling*" [22]. Species Distribution Models (SDMs) are an important tool that contributes greatly to the research of biodiversity which in turn helps us in the conservation of ecology [48]. SDMs provides us with a measurable entity which explains the relationship between the input variables or covariates (which could consist of variety of entities ranging from environmental features, climatic factors to remote sensing images) and the distribution of a species. SDM's function is to gather the spatial distribution of a species, given the details of the occurrence of the species obtained through multiple sources of species observation data. By employing SDMs to gather the spatial distribution of species, such as frogs (*Anura*), researchers can better understand environmental issues.

In order to take actions towards environmental conservation and address those issues, researchers need to know the factors that affect the ecosystem. In order to be aware of such factors, there should be some sort of indications that can help in the analysis of environmental problems. Bio-indicators [52] help in such analysis. By definition "*bio-indicators are living organisms of any kind such as animals, micro-organisms, plants etc, that assists*

in the screening of the natural ecosystem in the environment." [52]. One such bio-indicator are frogs (*Anura*), that helps in the assessment of the quality of the environment and the changes observed in the environment . Frogs serve as an important bio-indicator due to their high sensitivity towards even minor changes in environmental conditions [27].

The technological aspect of motivation for this work comes from the enormous potential shown by one of the major areas of deep learning called multimodal learning [7]. It involves processing and integrating information from multiple sources. This work aims to leverage multimodal learning in the context of building a SDM. The motivation behind employing this method stems from the promising results exhibited by this approach in a wide range of applications as mentioned in [7],[75],[49],[35] including in SDMs [60]. This further underscores the relevance of exploring this option further in the context of building a SDM. More on multimodal learning is explained in the following chapter.

This work describes the process of building a Species Distribution Model for predicting the presence of frog population and also to predict the count of frogs present at a location using multimodal learning under the assumptions that using more than one modality of data will improve the predictions of SDMs.

1.2 Problem Statement

Based on the above motivation, this work aims to address the following problem statement. "To fulfill the critical need for monitoring the distribution and population count of frog species as a bio-indicator for analyzing environmental changes, there is a need to develop a SDM that can effectively predict the presence and abundance of frog populations".

Additionally, the SDM built, will also be used to predict the presence / absence of frogs in a location. This task requires absence data, which is not available with the dataset used in this work. For that purpose, pseudo-absence data will be generated.

1.3 Research Objective

The concept of SDM has been in the picture for many decades now. There has been numerous methods available for building a SDM and it has evolved since. SDMs based on statistical methods [69][51][25] are a popular model which produced good results. On the other hand due to the evolution of deep learning methods and the effectiveness of Convolutional Neural Networks (CNNs), SDMs based on the above technologies has gained traction.

So far, SDMs based on models that uses covariates like temperature, moisture content of soil, elevation data etc. as input to classify the species using deep learning and SDMs that takes high resolution remote sensing images as input to classify via deep learning (CNNs) exists. Eventhough these model provide better predictions, important details are missed by those models mainly due to the nature of the input data being single modality. Research on making use of multimodal learning in building SDMs has not gained much attention. Refer scientific background section for more information on multimodal learning. So, the main objective of this thesis is to build a SDM based on MultiModal Learning [49][75] and compare the results obtained with existing state-of-the-art methods.

1.4 Thesis Outline

The thesis is organised as follows. Chapter 2 describes the relevant scientific background information. Chapter 3 outlines the literature survey conducted. Chapter 4 discusses the methodology followed, which includes the dataset preparation ,the model architecture and the experimental setup. In chapter 5, the results obtained are presented. Chapter 6 explains and analyses the results obtained and answers the research questions framed. Finally, chapter 7 gives a overall conclusion and also discusses the future research areas.

Chapter 2

Scientific Background

2.1 Multimodal Learning

In multimodal learning, the deep neural network learns various features over multiple modalities. In traditional deep neural networks, the input data involves a single modality. But as discussed in the research objective section, SDMs often take inputs from multiple modalities, so it is only logical that we employ a multimodal learning based model. By processing and learning from diverse data sources together, the models can identify complex correlation and dependencies that might not be captured when considering the different modalities separately. Having said this, multimodal learning can be defined as *the process of developing algorithms and architectures, that enable the models to handle and process information from multiple modalities or sources of data.* [49], [75].

2.2 Evaluation Metrics

In this part, some of the evaluation metrics that have been used in literature in measuring the model performance is explained.

2.2.1 Area Under Curve (AUC) or Area Under Receiver Operating Characteristic Curve (AUROC)

The performance of a binary classification model can be visualized graphically using the Receiver Operating Characteristic (ROC) curve. The curve is obtained by plotting True Positive Rate (TPR) against False Positive Rate (FPR). By measuring the Area Under the ROC curve (AUC), we obtain a degree of separability. A higher AUC value means, the model is better at predicting true positives and true negatives. In [11], the author proposes that AUC can be calculated by using trapezoidal integration, mathematically AUC can be calculated using equation 2.1 [11]

$$AUC = \sum_i ((1 - \beta_i) \cdot \Delta\alpha) + \frac{1}{2}(\Delta(1 - \beta) \cdot \Delta\alpha) \quad (2.1)$$

Where,

$$\Delta(1 - \beta) = (1 - \beta) - (1 - \beta_{i-1})$$

$$\text{and } \Delta\alpha = \alpha_i - \alpha_{i-1}.$$

$$\alpha = P(F_p) \text{ [False Positive Rate]}$$

and $\beta = 1 - P(T_p)$ [True Positive Rate].

2.2.2 True Skill Statistic (TSS)

TSS is a frequently used evaluation metric in assessing the performances of SDMs, especially when it comes to binary classification (distinguish between the presence and absence of a species). TSS is often credited for being independent of the proportion of the presence or absence of a species in the sampled locations (i.e. prevalence). Mathematically TSS is defined in equation 2.2 [65]

$$TSS = sensitivity + specificity - 1 \quad (2.2)$$

Where,

sensitivity = true positive rate (TPR),
and *specificity* = true negative rate (TNR).

$$TPR = \frac{TP}{TP + FN} \quad (2.3)$$

Where,

TP(True positives) = positive instances correctly classified,
and *FN*(False negatives) = negative instances incorrectly classified.

$$FPR = \frac{FP}{FP + TN} \quad (2.4)$$

Where,

FP(False positives) = positive instances incorrectly classified,
and *TN*(true negatives) = negative instances correctly classified.

2.2.3 Percentage Correctly Classified (PCC)

PCC is one of the common and easiest way to evaluate a model's classification ability. PCC corresponds to the proportion of observations that has been correctly classified. Mathematically, it is given by equation 2.5

$$PCC = \frac{\text{No.of correct classifications}}{\text{Total no of samples}} * 100 \quad (2.5)$$

2.2.4 Top-k Accuracy

For problems involving multi-class classification, where the model has to predict from N classes, this evaluation metric is the most used one. In top-k predictions, where k can be

any positive integer. In top-1 prediction, the prediction is correct only if the most probable prediction is correct, whereas, in top-k prediction, the prediction will be considered correct if one of the top k predicted values has the correct prediction. Top-k prediction is defined mathematically by equation 2.6 [58]

$$C_k(q) = \{y_{[i]} \in Y | 1 \leq i \leq k\} \tag{2.6}$$

Where,
 $Y = \{1, 2, \dots, N\}$ is the label space , N is Number of classes,
 $q = (q_1, q_2, \dots, q_N)$ is the normalized values
 obtained as the output after softmax activation,
 and $k = (1, 2, \dots, N - 1)$.

2.3 Activation Function

Activation functions that are used in this work is explained here.

Rectified Linear Unit:Rectified Linear Unit or ReLU is an activation function that outputs the input value directly if the value is positive or else zero.

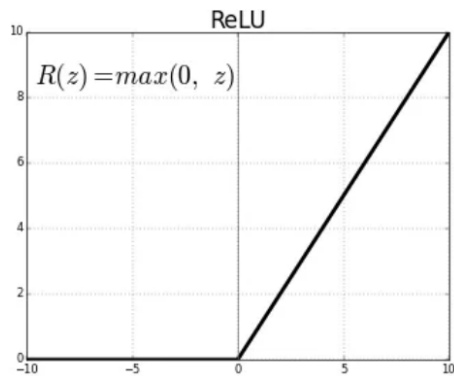


Figure 2.1: ReLU Activation [61]

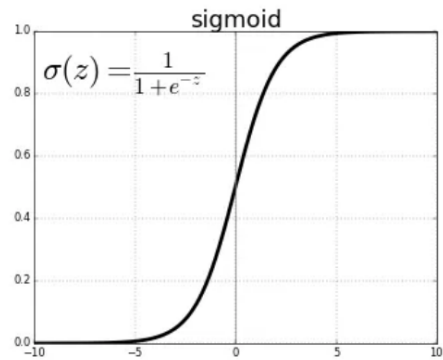


Figure 2.2: Sigmoid Activation [61]

Sigmoid:Sigmoid takes any input value and outputs a value in the range of 0 to 1. When the prediction is a probability, sigmoid is used. Especially when it comes to binary class classification.

Chapter 3

Related Works

This section is subdivided into two parts, the first part is about the research done on the existing methods for building a SDM and the second part is about Multimodal learning which will be the approach used in this thesis to build a SDM. This chapter also includes some preliminary research work done on using Graph Neural Networks (GNN) for building a SDM.

3.1 Current state-of-the-art methods

This section is categorized based on the types of techniques employed, progressing from elementary approaches to more intricate methods.

3.1.1 Conventional Statistical Methods

Statistical methods based SDMs are a trusted and proven approach. In literature one can find many SDMs based on statistical approaches. Statistical approaches based on presence-only methods are a popular approach in building SDMs. Before going into the method it is important to know the two major types of data available. Presence-only data contains only the presence details. This means that if the species is observed at a location, it is recorded as a presence point, however it fails to provide the location where the species is absent. This gap is overcome by presence-absence data which consists of both locations where the species is present and absent. More on the dataset that will be used in this work is discussed later. Several works [2][12][13] have been proposed based on presence-only method. But, the most popular one is MaxEnt [54]. In the context of Species Distribution Modelling, Phillips et.al., in [55] suggests that given the presence details of the species, the region of interest and the covariates, the distribution of the target species over the geographical region of interest is predicted by calculating the distribution of maximum entropy provided that the expected value of each feature is equivalent to its empirical average. Though this method is consistent and produces good results, which is evident from the experiments conducted by the authors in [54], there are some serious drawbacks with this approach. One such issue is that this method is based on presence-only dataset and this may lead to bias in the occurrence localities.

Several works such as [24] [25] [51] have used statistical methods, generalized linear models for instance. Manel et.al., in [47] compared different methods of building a SDM to predict the presence of *Rhyacornis fuliginosus*. One of the methods used is based on

Logistic Regression. In this context, logistic regression is treated as more of a statistical method than a machine learning based method because of the linear dependence between the variables. The authors employed a generalized linear model based on multiple logistic regression to model the presence/absence of the species. The equation 3.1 represents the linear function of 32 predictor variables.

$$(p) = \log \frac{p}{1-p} = b_0 + \sum_{i=1}^{32} b_{1i}x_i \quad (3.1)$$

Where,

(p) represents the logit transformation of presence/absence (p) ,
 b_0 and b_{1i} are the regression constants.

The authors of [47] when compared logistic regression with other methods like artificial neural networks and multiple discriminant analysis, found logistic regression to have better performance in than others, but the difference is minimal. Though logistic regression have an edge over artificial neural networks in terms of efficient usage of resources, it comes with a serious issue. In the case of logistic regression, a situation where predictor variables could have a significant impact on the presence/absence of a species by coincidence or random chance rather than due to the representation of actual relationship between covariates and occurrence can arise. This is due to the building of model based on only correlative data. This makes it one of the reasons to look beyond statistical model in building SDMs.

3.1.2 Machine Learning Methods

Even though statistical methods have yielded good results in building SDMs, due to the impact of machine learning it is hard to ignore it. And also, a linear function may not be sufficient to explain the relationship between the environment and the species. As a result several works based on machine learning methods such as random forests classification [14], support vector machines [19] and boosted regression trees [21] are proposed for SDMs.

One such machine learning based model was presented by Cutler et.al., in [14]. The authors used random forest classifier to build Species Distribution Modelling. RF method is compared with 4 other different classification methods, such as Linear discriminant analysis (LDA), logistic regression, additive logistic regression and classification trees. Accuracy of each method is calculated using overall percentage correctly classified (PCC), sensitivity (the percentage of presences correctly classified), specificity (the percentage of absences correctly classified), kappa and AUC.

This method is applied to 3 examples for classifying 3 groups of organisms, vascular plants, non-vascular plants, and vertebrates. In all three examples it is found that RF outperforms other classifiers. It is concluded that RF should outperform linear methods like logistic regression and LDA which has high interactions among variables. In addition to it RF has the ability to measure the variable importance better than other ML based methods such as SVMs. But the downside with this approach is that the relationship between predicted values and covariates is complex and this makes interpreting ecological information difficult. Lek et.al., in [41] compared two techniques for modelling SDMs namely multiple regression and neural networks. The authors concluded that neural net-

works performed better when there exists a non-linear relationship between variables when compared to multiple regression.

3.1.3 Deep Learning Methods

The environmental variables can be of many types including elevation data, land cover, soil type in addition to climate data. As the input variables increase it is important to make sure that the complex relationship between the input variables is addressed properly. This is where deep learning comes into the picture. But, before going into deep learning, neural networks with a single hidden layer has been utilized in the field of SDM [42] [40] many years back. Though these networks produced better predictions, deep neural networks with more than one hidden layer perform better when the complexity is high [9]. Botella et.al., in [10] put forward a deep learning approach for building a SDM. According to the authors in [10] the main goal of SDMs is to obtain a function that outputs the density of the species at a location given the environmental features as the input. This means that the species is constrained to one specific ecological niche that is distinguished by the distribution over the environment. But, it has already been established that the function is more complex than expected. The paper [10] provides a classical equation which helps us understand the correlation between species abundance and covariates.

$$g([y|x]) = \sum_j f_j(x_j) + \sum_{j,j'} h_{j,j'}(x_j, x_{j'}) \quad (3.2)$$

Where,

y is the target variable whose presence is to be predicted,

x is the input variable,

f and *h* are the monovariate and bivariate functions that describe the relation between the inputs,

g is to make sure that the expected value is within the space of *y*.

From equation 3.2 [10] we can understand that, it is the case in most of the time that the pairwise interaction effect between the covariates is expressed by the product of their values, which simplifies the model and makes it easier to interpret. This approach may not work always as it assumes a simple correlation between the covariates and the species response. In reality, this does not reflect the complexity of the environmental patterns that influence the species occurrence. So, neural networks with several layers can negate this issue as their architecture can accommodate complex interactions between input variables. The deep neural network proposed in [10] consists of a feedforward network with six hidden layers and uses ReLU activation function. From the experiments conducted in [10], the authors found deep neural networks to outperform the classical MaxEnt approach. Similarly several works [79] [67] have used deep learning to build SDMs and have achieved better results. However, when the dataset on which the deep neural network is trained on is small, the model will overfit leading to a degraded performance on the validation dataset [1].

Another drawback with using deep learning approach in building SDMs is that extra work should be put in deciding the hyper-parameters and architecture before training the neural network. Shiferaw et.al., in [62] compared various machine learning algorithms and deep neural networks in predicting a invasive plant species. From the experiments

conducted, the authors obtained a 92% accuracy when using a random forest approach. However deep learning approach only yielded an accuracy of 73%. The authors could not definitively explain the reasons for the under performance of DNN, but suspect that various factors such as proper weight initialization, selection of right optimizer and further hyper-parameter tuning could result in a better performance.

A special case of deep neural networks are the one that has a convolutional layer in it. Convolutional neural networks (CNNs) [39] are a specific class of neural networks that specializes in image data due to the convolution operation between the matrices. CNNs have seen a lot of success in the field of image classification [37], pattern recognition [73] etc. due to their ability to extract spatial features from the image. So, the spatial patterns in the covariates contains information that are missed by usual machine learning methods.

Deneu et.al., in [15] proposed a SDM based on CNN. The authors suggest that Convolutional Neural Networks, have a distinct characteristic in that they rely on spatial environmental tensors, which are representations of the spatial distribution of environmental factors surrounding each point, rather than simply using local values. Given their ability to capture rich information through the use of spatial environmental tensors, CNN-SDMs are well-suited to model how complex ecological niches and spatial dynamics influence the distribution of numerous species within a given region. CNN-SDM presented in [15] consists of the input data subjected to a non-linear transformation initially. This results in a feature vector of lower dimension. For this transformation the authors make use of the Inception v3 model [66]. The feature vector obtained captures information about the environmental characteristics. This feature vector acts as a input in predicting the species at the end using a generalized linear model. Based on the experiments conducted in [15], the authors suggest that CNN-SDMs performed better when the occurrence data are limited.

3.1.4 Summary of Methods for SDMs

The table 3.1 compares and summarizes the main state-of-the-art methods that were discussed so far. Since each of the methods discussed in the literature have used different target variables (species) and different metrics of evaluation, it is difficult to directly compare them and come to a conclusion that one method is better than other. However, the following summary made a diligent attempt to provide a comprehensive comparison of all the methods by utilizing the available data and highlighting their respective drawbacks.

From the discussion about the random forests classifier approach earlier, various metrics were used to evaluate the different methods that were utilized to predict various species. From those only PCC & AUC metrics are chosen to compare as they are the most common method of evaluation in the literature. Also only one species (*Verbascum thapsus*) is chosen because of the highest number of observation among others. From the experiments conducted in [14] Random forests classifier outperformed all other classifier.

Similar to Random forests classifier, in MaxEnt [55] approach AUC is chosen as the evaluation metric for comparison purpose. This method obtained a high AUC value compared to the other commonly used presence-only method called Genetic Algorithm for Rule-Set Prediction (GARP).

In [47] logistic regression based model fared better than ANN and MDA methods. This is due to a straightforward linear method of predicting the distribution.

When it comes to deep learning approach carried out by Botella et.al., in [10], in order to not be partial towards the selection of the category of species, 1000 species were chosen randomly from 7626 species and from 1000, 200 were chosen and finally 50 species

| S.No | Method | Species/Dataset | Evaluation Metric | Value |
|------|--------------------------------|-------------------------------|-----------------------------------|--------|
| 1 | Random Forests Classifier [14] | <i>Verbascum thapsus</i> | PCC | 92.6% |
| | | | AUC | 0.940 |
| 2 | MaxEnt [55] | <i>Microryzomys minutus</i> | AUC | 0.982 |
| 3 | Logistic Regression [47] | <i>Rhyacornis fuliginosus</i> | Prediction performance (Accuracy) | 75% |
| 4 | Deep Learning [10] | Randomly selected 50 species | Mean Loss | -0.927 |
| | | | RMSE | 2.61 |
| | | | A10%DQ | 0.519 |
| 5 | CNN-SDM [15] | GBIF dataset | Mean top-k accuracy | 0.34 |
| | | | AUC | 0.818 |
| | | | TSS | 0.450 |

Table 3.1: Performance Comparison of Existing State-of-the-art Models. AUC- Area Under Curve, PCC- Percentage Correctly Classified, TSS- True Skill Statistics, RMSE- Root Mean Squared Error, A10%DQ- Accuracy on 10% Densest Quadrats

were chosen. And the evaluation metric chosen are mean loss, RMSE and accuracy on 10% densest quadrats, which are different from the usual AUC. From the experiments conducted, deep neural network outperformed MaxEnt method when there are multiple species at the output. However, for mono response version MaxEnt performed better.

In the case of CNN-SDM described in [15], the evaluation metric chosen are mean top-k accuracy, AUC and TSS. And the dataset is GBIF consisting of 4520 plant species. When compared with the other methods such as deep neural network, random forests classifier and boosted regression trees, CNN-SDM had the highest mean top-k accuracy. However, when AUC & TSS metric are taken into account, there was not much difference when compared with random forests. But, it was observed that CNN-SDM performed well when it comes to rare species with less observation data while random forests and boosted trees performed better for frequently occurring species.

The table comprises of the results obtained from the best performing method which were presented in the literature explained so far. Eventhough it is difficult to compare these methods in a straightforward manner, it can be infered that each method has its own pros and cons.

3.2 Multimodal Learning based SDMs

In multimodal learning [49] [75], deep neural networks learns features by taking in inputs from multiple modalities. Several applications that uses multimodal learning has been developed so far. For example, video and image captioning [68] [71], generation of images from texts [72] [74]. These models uses images and text as input data. There are also speech recognition[20] which combines audio and visual data. Likewise there are several other applications which are developed where the deep neural network learns across multiple modalities. However, there hasn't been a lot of work done using multimodal learning in the field of SDMs so far.

Though deep learning based SDMs have produced good performance, the input to those models are only of one modality. Mostly the input will be either of environmental variables or high resolution satellite images. When the model uses only one type of data, for example environmental variables like temperature, soil type, humidity etc., it misses out on important information about spatial patterns of the area. If only satellite images are used, the model misses out on important climatic data. Either way the model performance takes a hit due to missing out on important information.

So, in order to tackle this issue, multimodal learning is utilized to handle the heterogeneous nature of input data used in building a SDM.

When it comes to multimodal learning, one of the task to keep in mind is the fusion of representation of input data of different modalities. The stages at which the fusion occurs plays a major role in the performance of the multimodal model. There are only a few works that has focused on building an SDM using multimodal learning, and the following discussion will be about the related works that focuses on the different fusion methods.

Deneu et.al., in [16] used input data of multiple modalities and fused them at the input stage itself. The combined data is then fed to a CNN model and the features are extracted from it. Though this is a simple method, the drawback of this approach is that, since the input data are combined at an early stage itself, it might be difficult to capture and leverage unique information from different modalities especially when the modalities of input data are significantly different from other.

Seneviratne in [60] tried incorporating multimodal images in building an SDM for habitat prediction of 30,000 species. The author trained a ResNet50 model with the base

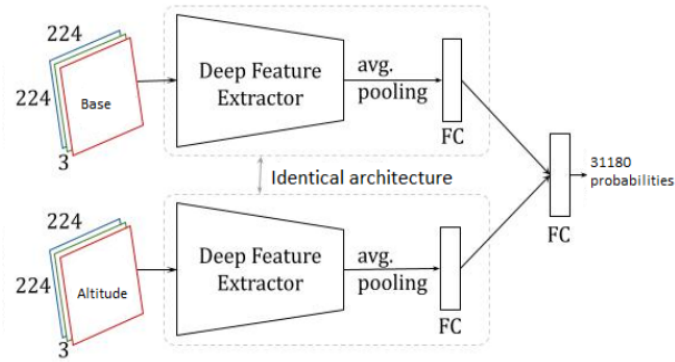


Figure 3.1: Multimodal Analysis - Generic Architecture [60]

| Model | Input Data | | | Top-30 Accuracy |
|-----------------------------|------------|------------------------|------------------------|-----------------|
| | RGB Images | Environmental Variable | Longitude and Latitude | |
| Top-30 most present species | ✗ | ✗ | ✗ | 4.36 |
| Random forest | ✗ | ✓ | ✗ | 21.68 |
| ss model | ✓ | ✗ | ✗ | 26.32 |
| ResNet50 | ✓ | ✗ | ✗ | 25.72 |
| SDMM-Net v1 | ✓ | ✓ | ✗ | 25.8 |
| SDMM-Net v2 | ✓ | ✓ | ✓ | 26.4 |
| Swin-T | ✓ | ✗ | ✗ | 25.13 |
| PreFuzeMM-Swin | ✓ | ✓ | ✓ | 29.56 |
| PostFuzeMM-Swin | ✓ | ✓ | ✓ | 26.07 |
| MidFuzeMM-Swin | ✓ | ✓ | ✓ | 29.37 |

Figure 3.2: Comparison [77]

RGB imagery initially and then included altitude images and compared the results. The architecture used for multimodal analysis consisted of an additional branch for extracting features of the altitude imagery other than the one used for RGB imagery. Instead of fusing the lower level features of the two modalities at the beginning itself, this architecture fuses higher level features at the end to facilitate the learning of more fine grained information about the different image domains. This leads to the network being adapted easily to different image modalities. The architecture used in [60] for multimodal analysis is shown in figure 3.1. From the experiments conducted in [60], multimodal structure achieved lower error rate compared to the unimodal structure. However, it was discussed that higher GPU memory footprint is seen as a drawback with this approach.

Another simpler approach is to train the different models with different modalities of input variables independently and then evaluate the prediction by taking the average of all the model’s predictions. Eventhough it is a lot more easier approach, there is no single layer where the fusion of different modalities occurs in a significant manner and this could lead to a degraded prediction.

Zhang et.al., in [77] presented two approaches of multimodal learning in building an SDM. The first approach constitutes a two branch approach, one each for two different modalities of predictor variables. The first branch consists of a ResNet architecture, used for extracting spatial features from remote sensing RGB images. The second branch is made up of a fully connected layer for processing 27 environmental variables passed as 27-dimensional vector. The two branch’s vector are then concatenated together as single vector which is then passed through a fully connected layer and a softmax layer to get the final prediction.

Attention mechanism is a concept where the weighted combination of feature vectors are generated using scalar weights. These weights are generated by taking into account the

| S.No | Method | Species/Dataset | Evaluation Metric | Value |
|------|--|------------------|-------------------|--------|
| 1 | Fusion of CNN and RF Prediction [16] | GeoLifeCLEF 2020 | Top-30 Accuracy | 0.24* |
| 2 | ResNet50 - Fusion of RGB and altitude Imagery [60] | GeoLifeCLEF 2021 | Top-30 Error rate | 0.748 |
| 3 | ResNet50 - SDMM-Net [77] | GeoLifeCLEF 2020 | Top-30 Accuracy | 26.4% |
| 4 | Swin-T Transformer (PreFuze) [77] | GeoLifeCLEF 2020 | Top-30 Accuracy | 29.56% |
| 5 | Ensemble of 3 Models (leaderboard) [18] | GeoLifeCLEF 2022 | Top-30 Error rate | 0.684 |

Table 3.2: Performance Comparison of Existing Multimodal Models. *- Not reliable due to a bug.

importance of vectors in an image. For example setting higher scalar weights to more important vectors and lower weights to less important ones. This way an attention mechanism generates vectors of an image that highlights the significant features of an image.

The second approach presented in [77] is based on Swin Transformer which is a model based on attention mechanism. [46]. For comparison purpose, out of different versions of swin transformer, Swin-Tiny is used as it is similar to ResNet50 in terms of computational purpose. 4 different fusion methods are implemented namely pre-fuze, post-fuze, mid-fuze and feature addition and concatenation method. In pre-fuze method the Swin-T structure is used to extract the features after fusion, where in post-fuze method Swin-T structure is used to extract only features of remote sensing images and the fusion with the features of environmental data occurs at the end. In mid-fuze the fusion occurs at the end of every stage of the Swin-T structure. These methods so far are called data fusion technique, where feature fusion techniques such as addition and concatenation are also used to build an SDM. From the figure 3.2 we can observe that pre-fuze method yielded highest accuracy among other methods proposed in [77].

In addition to the methods mentioned above, the winning model of GeoLifeCLEF 2022 [18] is also discussed. It is a combination of 3 models of which 2 are deep learning based and 1 is a Random forest model. The first model is a bi-modal network which uses a pre-trained ResNet34 for remote sensing images and a FCN layer for environmental vector. The second model is similar to the first one but it uses MobileNetV3 instead of ResNet34. The third model is a Random forest with 32 estimators. Finally the predictions from all 3 models are merged using mean probabilities approach. However there is not much information about other parameters and the architecture used in this approach. This method yielded an error rate of 0.684.

From table 3.2 the results suggest that the Swin-T Transformer (PreFuze) method outperforms the other methods in terms of accuracy in predicting species or evaluating SDMs on the respective datasets. Though the datasets are same in most of the cases, it is hard to pick one method which we can say is the best just from these four cases.

Also due to the limited number of published studies in the field of multimodal learning for constructing SDMs, it is challenging to draw significant conclusions or derive extensive

data from the performance comparisons presented so far.

3.3 Graph Neural Networks

This section gives a short background on using Graph Neural Networks (GNNs) in building a SDM (**Under the GNN approach only the dataset has been generated. The experiments conducted were not conclusive so it will not be explained in this thesis**).

Graphs in general are a type of data structure that consists of nodes and edges. Edges represents the relationship between the nodes. Using this type of data structure in deep learning have gained attention. GNNs are a class of machine learning / deep learning models that works on data represented by graphs [78]. GNNs have found usage in a variety of application areas like bioinformatics [76], wireless networks [32], computer vision [56], weather forecasting [36] etc. When it comes to species distribution modelling, GNNs have been seldom used. However, Han et.al., in [30] used message passing in GNNs for predicting species. In [38], GraphCast is introduced where the GNNs are employed in a encode-process-decode structure for forecasting the weather. The input data is first encoded in a graph structure which then uses GNNs to learn the complex interactions between the data through message passing and then finally decoded to output the predicted weather. Unlike the latitude-longitude based grid division, GraphCast introduces multi-mesh which is obtained by successively dividing a icosahedron into many levels.

3.4 Research Gaps

From the extensive literature survey conducted, the following research gaps are identified, which will be looked to address in this thesis.

3.4.1 Only limited works have been performed that uses pseudo-absence data in building a SDM

Since, collecting absence data is harder than collecting presence data, researchers often use pseudo-absence data in predicting the species occurrence. There are only minimal works that has actually used pseudo-absence data, especially when it comes to using multimodal learning. Only few notable works like [45] has conducted research on how the prediction accuracy varies when using pseudo-absence data. Moreover method for generating pseudo-absence data is little explored. The pseudo-absence data generated should be reliable in predicting the species occurrence. The method chosen for generating the pseudo-absence data should be accurate and the generated data should be comparable to that of an actual absence data. Only few notable works [59] are published for generating a pseudo-absence data. This work compares the proposed method of generating the pseudo-absence data with few of the already existing methods in the literature.

3.4.2 Feature importance assessment

From the literature reviewed so far, the impact of using relevant features has been relatively explored less in the context of deep-learning based SDMs. Assessing the importance of different features can provide valuable insights into identifying the most influential factors that drives species occurrence.

3.4.3 Raw-dataset Balancing

The problem faced with dataset used in building SDMs is that they are raw, similar to all tasks in deep learning. But with SDMs there are some specially associated difficulties.

1. **Spatial Bias:** Data collected from some location could be in large number compared to other location. This leads to over representation in some parts and could lead to bias in model prediction.
2. **Generalization:** The purpose of SDMs is to use it on unseen locations. So merely oversampling the data will not be enough. This leads to a more complex balancing technique.

These challenges are taken into consideration while balancing the dataset. Proper balancing of dataset should be performed before building a SDM. This area is little explored so far in existing literature especially when it comes to regression task (counting frogs).

3.5 Contributions of this thesis

- A novel method for generating pseudo-absence data points.
- Dataset balancing method for SDMs (regression challenge).
- How different climatic features affect the distribution of frogs is addressed here.
- How multimodal learning is leveraged in building a SDM for frogs is presented in this work.

3.6 Research Questions

Based on the research objective and the literature survey conducted, the following research questions are framed and will be addressed in this project.

RQ1: *What are the major limitations of the existing methods for building an SDM in predicting the presence of a species in a particular location?*

It is crucial to identify the limitations that are present in the existing methods. Because by doing so, those limitations will be looked to address in the proposed method.

RQ2: *How multimodal learning can be made use of in building a SDM?*

This question is framed to explain how the technology of multimodal learning is leveraged to build a SDM. The techniques and how input data of multiple modalities are made use of is answered through this question.

RQ3: *How does the performance of the proposed SDM based on multimodal learning compare to the existing state-of-the-art methods?*

In order to gain insights into the performance of the proposed model, it is necessary to compare with a baseline model. This will eventually be helpful in determining where the

proposed model stands and also identify the areas of improvement.

RQ4: *What is the impact of the different covariates used(temperature, land cover etc) in contributing towards the prediction of target variable?*

It is important to identify the factors that influence the most in prediction of the target variable. By doing so, important questions like how a particular factor impact the habitat selection of the species.

RQ5: *How is the performance of the model using pseudo-absence points generated by the proposed method compared to the existing ones?*

Since one of the contribution of this work is to provide a better way of generating pseudo-absence points, the generated points are compared with some of the existing methods.

Chapter 4

Methodology

In this chapter, section 4.1 describes about the datasets used, followed by the pre-processing techniques used in section 4.2. Then the proposed method for generating the pseudo-absence data is explained in section 4.3. Section 4.4 describes the dataset generated for the purpose of using GNN. Sections 4.5 and 4.6 explains the model description and the experimental setup respectively.

4.1 Dataset Description

This section explains the datasets that are used as predictor variables for building the SDM and the frog occurrence dataset that contains the frog count which will be used as the target variable.

4.1.1 Environmental Covariates (Predictor variables)

The Microsoft Planetary Computer Portal [5] is an open source platform that provides access to a number of geospatial data which includes high-resolution satellite images. The dataset consisting of patches for the required grids are downloaded using *pystac* API client. Out of all the available data catalogs, the following covariates are chosen for the task in hand.

Sentinel-2 Level-2A

This dataset provides high resolution (10m to 60m) satellite imagery in 13 spectral bands. Out of all the spectral bands, this work utilizes RGB and NIR bands as predictor variables. A sample RGB patch in Costa Rica is shown in figure 4.1.

JRC Global Surface Water

JRC - GSW are comprehensive datasets that provide information about the occurrence, seasonality and transition of surface water on a global scale.

Esri 10-meter land cover (10 class)

This dataset provides information regarding the type of Land Cover (LC) present at various locations on the earth's surface. The dataset includes land types such as water bodies, trees, grass, crops etc. The dataset is characterized by a spatial resolution of 10 meters.



Figure 4.1: RGB Patch- Costa Rica

Copernicus DEM GLO-90

This dataset represents the surface of the earth that includes, buildings, infrastructure and the elevation data. As researchers suggest that the distribution of several species, especially amphibians could change and move towards higher altitudes due to rising earth's temperature [44], the elevation data of the species occurrence is taken into consideration.

TerraClimate

This dataset provides climate and water balance information of the earth in a monthly basis. The dataset includes a wide range of climate variables, from which temperature, precipitation, Palmer Drought Severity Index, vapour pressure, soil moisture etc., are chosen. These specific variables are chosen due to their high influence on frog habitat and their distribution. Temperature (tmin and tmax) affect frog's distribution as pointed out by Gerick et.al., in [28] as the species' distribution will experience a temperature beyond their thermal optimum capability. But their thermal safety margin is between 3.2 to 3.8°C. So, it is important to take temperature into account when building a SDM. Similarly, experiments conducted in [44] suggests that severe drought will affect a species' distribution and their existence. [70] points out the importance of soil moisture in the distribution of amphibians. And precipitation is considered because, wet-skinned organisms like frogs require a moist skin for respiration and for maintaining their body temperature as pointed out by Lertzman-Lepofsky et.al., in [43].

An overview of all the dataset used is outlined in table 4.1

4.1.2 Frog Occurrence Dataset (Target Variable)

The frog presence dataset for the frog counting tool has been provided by [6] for three countries namely, Australia, South Africa and Costa Rica in CSV format. For Australia the dataset is part of FrogID project [3] and for South Africa and Costa Rica the dataset is part of iNaturalist Research-grade observations [4]. Both these projects involves dataset collected as a part of citizen science project where people share bio-diversity information and create a database of different species observed at different locations.

The initial dataset provided pertaining to all three countries encompassed a variety of information, including details such as the species name, coordinates where the species was observed, the corresponding date and time of observation etc. The initial grid size for

| S.No | Dataset | Variable/Image Type | Resolution |
|------|-------------------------------------|---|------------|
| 1 | Sentinel-2 Level 2A | RGB and NIR band | 10m |
| 2 | JRC Global Surface Water | Transition, seasonality, recurrence, occurrence, maximum water extent | 10m |
| 3 | Esri 10-meter Land Cover (10 class) | Global land cover (Water, trees, grass, flooded vegetation, crops, shrub, built area, bare ground, snow, clouds) | 10m |
| 4 | Copernicus DEM GLO-90 | Elevation data | 90m |
| 5 | Terraclimate | Temperature (Tmax, Tmin), vapour pressure (Vap), vapour pressure deficit (Vpd), precipitation (Ppt), soil moisture (Soil), palmer drought index (Pdsi), wind speed (Ws), runoff (Q), radiation flux (Srad), evapotranspiration (Aet). | – |

Table 4.1: Overview of datasets used.

calculating the frog density given was 225sq km, since the submission data has the same resolution. But, the frog density was calculated for 30 sq km for the purpose of having more quantity of data. However, this raw data as such cannot be utilized for predicting the density or count of frogs at specific areas.

4.2 Pre-Processing Techniques Used on the Datasets

To make the data usable for such predictions, a number of transformations including data preprocessing, removal of outliers etc, were undertaken. The pre-processing techniques that are performed in this work is explained here.

4.2.1 Grid Approach

In order to calculate the frog density of all three countries, we create grids of 30 sq.kms area such that the entire country is partitioned into several grids. The grids are enclosed by bounding box coordinates (min_latitude, min_longitude, max_latitude, max_longitude). Once the grids are created, we obtain the frog count of each grid by iterating through each grid and subsetting the frog presence points available from the original dataset provided. An illustration of the grid creation for Australia is shown in figure 4.2. The presence points of Costa Rica, South Africa and Australia are shown in figure 4.3 . The points are visualized using QGIS software.

4.2.2 Feature engineering

Apart from the raw data obtained from the datasets mentioned above, it is possible to derive meaningful features from the available datasets. This process of deriving new features from the existing features is called feature engineering [57] [33]. For the task of frog counting, two such features are identified which could help in predicting the count better.

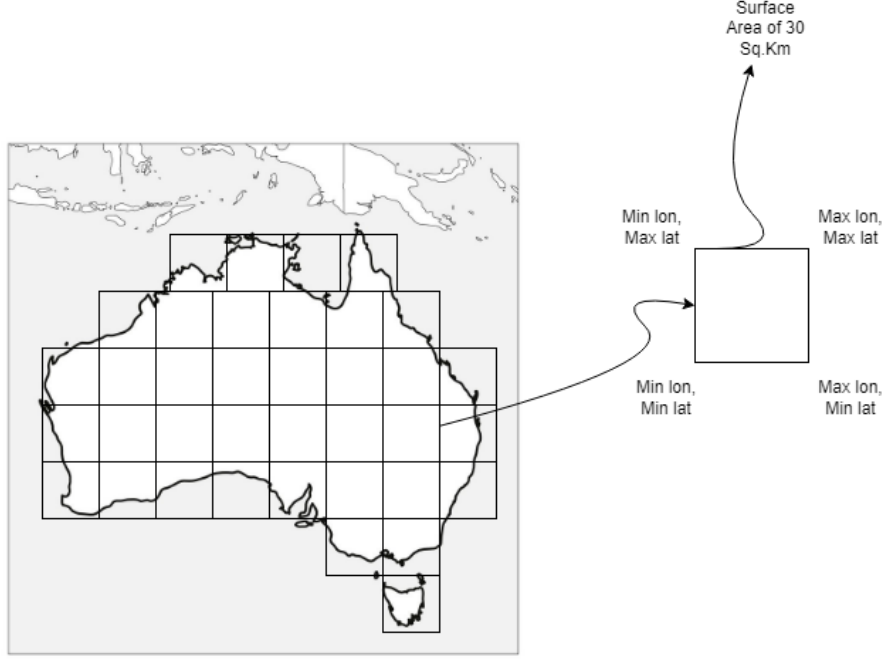


Figure 4.2: Grid Formation for Australia

1. **Normalized Difference Vegetation Index (NDVI) [34]:** This gives us a quantifiable entity that helps us in measuring the vegetation of a particular location. This could be one of the major factor in the habitat suitability of frogs. NDVI is calculated by equation 4.1 and its value lies between -1 and +1.

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (4.1)$$

Where,

NIR and *Red* are the spectral reflectances of *NIR* (Near infrared) and *Red* Channels respectively.

2. **Normalized Difference Water Index (NDWI) [26]:** NDWI provides us with an index which tells us the surface water content, which also highly influences frog habitat. NDWI is calculated by equation 4.2

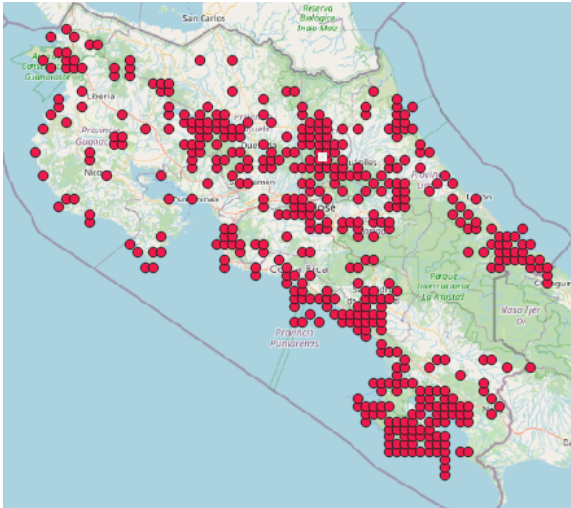
$$NDWI = \frac{Green - NIR}{Green + NIR} \quad (4.2)$$

Where,

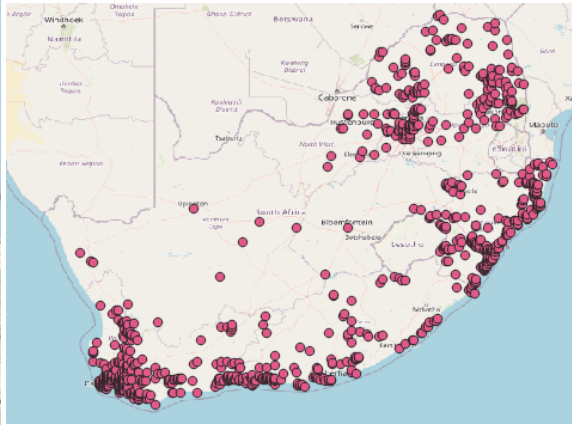
NIR and *Green* are the spectral reflectances of *NIR* (Near infrared) and *Green* Channels respectively.

4.2.3 Dataset Balancing

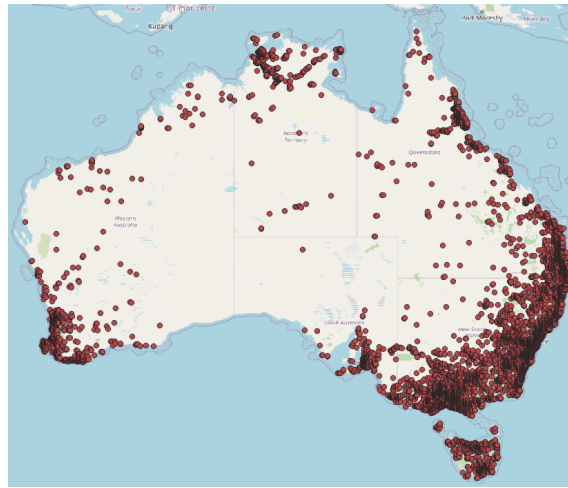
Imbalanced dataset is a major problem in the field of machine learning. An imbalanced dataset is a situation where the distribution of the data is heavily skewed, meaning one



Presence points - Costa Rica



Presence points - South Africa



Presence points - Australia

Figure 4.3: Presence points comparison

class has significantly more instances than others. In our case, since the frog count is our target variable, some values of counts are represented more compared to others. From figure 4.5 we can see that, the count of frogs are heavily concentrated between 1 and 10. This is a heavily imbalanced dataset. This imbalance leads to several problems such as

- Bias in model performance: Models trained on imbalanced dataset shows bias towards majority class.
- Overfitting: This is caused when the model performs well on training data but fails to give similar performances on unseen data.
- Under represented class: The model may fail to learn patterns from the minority class.

In order to tackle this issue, one of the methods to balance the dataset called oversampling is performed. Oversampling is where the representation of minority class is increased by duplicating it n - number of times to balance the dataset. But a mere duplication of the data can lead to the dataset losing diversity and variability. For this purpose K-means

clustering algorithm is used to group similar data points together based on the defined terraclimate features. Once clusters are formed, one instance from each cluster is extracted. This way only unique data points are present in the dataset without losing the diversity of the original dataset.

So, the procedure described below is followed to oversample the original dataset, without losing the variability.

1. Obtain the frequency of frog counts and identify the minority range of frog counts.
2. Apply oversampling of the data points that have less frequency of frog counts.
3. Retrieve the resultant data frame with somewhat balanced dataset. This balanced dataset is obtained by merely oversampling the available datapoint. To restore the diversity of the original dataset K-means clustering is employed.
4. **Feature Selection:** Define the features for performing k-means clustering. The features are composed of the various parameters from terraclimate dataset. These features define the dimensions along which the similarity is measured. As a result a feature matrix is formed.
5. Define the number of clusters (n).
6. **Perform k-means clustering:** K-means clustering works by randomly instantiating n number of centroids. Centroids are nothing but the centre of clusters. Then in the first iteration, all the datapoints are assigned to one of the cluster based on how closer they are to the centroids. This way all similar points are present in the same cluster. The closer together two data points are the more similar, the farther apart the less similar. During each iteration of the algorithm, centroid of each cluster is updated. The center of each cluster is found by the mean feature values of datapoints within the cluster. This corresponds the centroid.
7. **Convergence:** The K-means clustering algorithm continues by updating the centroid until convergence. This occurs when there is no significant changes in the center of clusters between iterations.
8. **Cluster Label:** Each data point is assigned a cluster label indicating the cluster it belongs.
9. **Unique Indices:** For each cluster, only one instance is kept. This way only unique values are kept maintaining the diversity of the dataset.
10. Get the resultant dataframe with balanced dataset without losing variability.

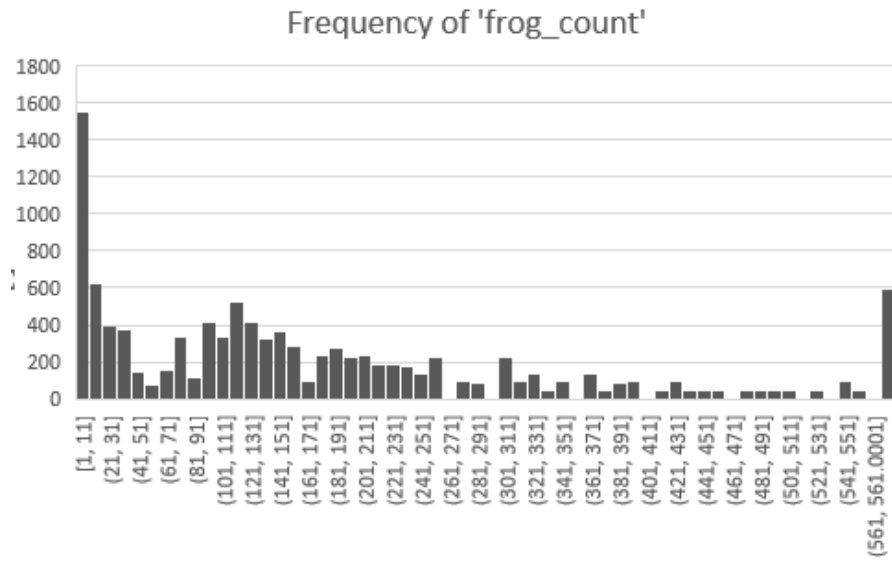


Figure 4.4: Histogram- Balanced Frequency of Frog Counts

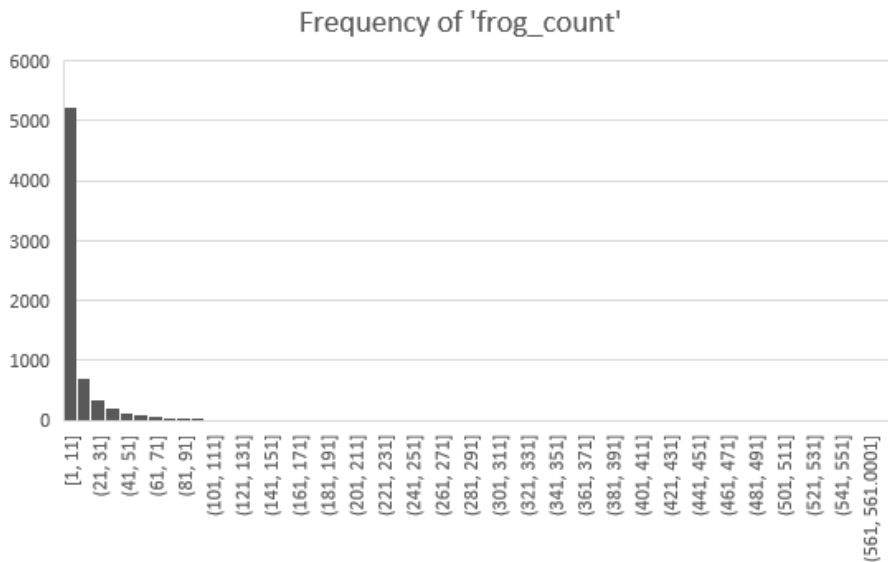


Figure 4.5: Histogram- Frequency of Frog Counts

After performing the above steps we obtain a better balanced dataset shown in figure 4.4 compared to the original one.

Apart from the data imbalance which was present based on the frequency of frog counts, the dataset also had imbalanced data countrywise. Out of the three countries for which the frog occurrence data is present, Australia covers about 82% of the data while South Africa taking up 11% and Costarica filling up the remaining 7%. This type of imbalance can lead to biased models with lower performance on the minority classes.

So, to negate this issue, weighted loss function [23] [8] is implemented in this thesis. The idea is to assign higher weight to minority class (Costarica and South Africa), and lower weight to majority class (Australia). The weights assigned to each class is calculated by equation 4.3

$$\text{Weight of class } x = \frac{\text{Total.no.of.samples}}{\text{No of samples in class } x * \text{No of Classes}} \quad (4.3)$$

These weights assigned to each class is multiplied with the original loss value and then added with the regularized loss with L2 regularization. This is given in the equation 4.4

$$\begin{aligned} \text{Total_loss} = & (\text{Weight.of.class}(x) \times \text{loss_value}) \\ & + \text{regularization_loss} \end{aligned} \quad (4.4)$$

4.2.4 Log Transformation

Skewness is a common problem faced with respect to the distribution of the data. Skewness is a measure of the asymmetry of data distribution. When the data is not normally distributed then the performance could take a hit. This is the case for the data used here even after the balancing step. So, log transformation of the dependant variable is performed to reduce the effect of skewness. Our particular data is positively skewed, i.e. the data distribution has a long tail in the positive direction of the number line. Refer figure 4.6 for the histogram of frog count before and after log transformation. It is important to note that the predictions obtained by the model trained using log transformed data will be in log scale. So, to get back the predictions for interpreting, the inverse of log transformation has to be taken.

4.2.5 Image Augmentation

One of the techniques used to balance the dataset is introducing duplicate data. By filling the dataset with duplicate data the uniqueness and variability of the images are compromised. In order to preserve the variability and uniqueness of the images various image augmentation techniques are utilized. Performing image augmentation before training also makes the model more generalized and prevent overfitting. The image augmentation techniques used are explained below.

Horizontal Flip : This technique flips the image along its vertical axis. Horizontal flip is performed with a probability of 0.5, meaning there is a 50% chance that the image will be flipped before training.

Rotation: Here a random rotation is applied to the image. The angle of rotation is randomly chosen between -10 and 10 degrees. So, instead of uniform angle applied every time, the image will be rotated randomly within the specified range.

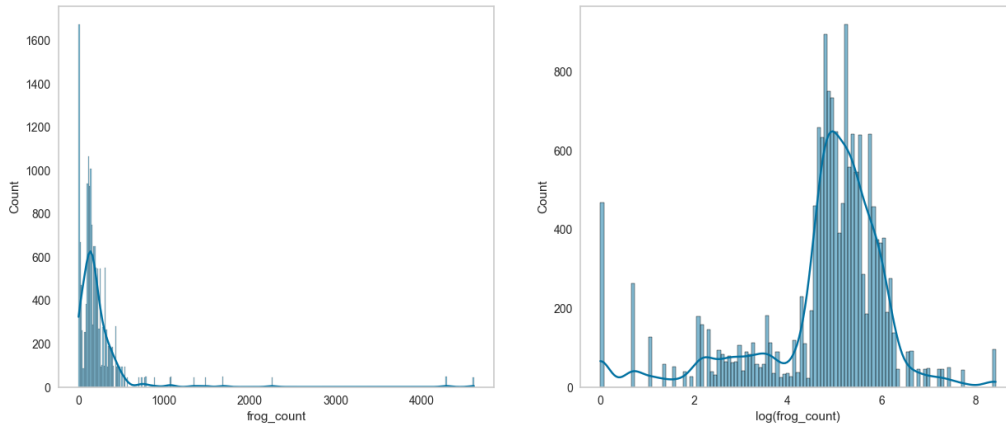


Figure 4.6: Histogram of Frog count before and after log transformation

```

missing value in aet : is 96 ---> 1.3316687473990845%
missing value in def : is 96 ---> 1.3316687473990845%
missing value in pdsi : is 96 ---> 1.3316687473990845%
missing value in pet : is 96 ---> 1.3316687473990845%
missing value in ppt : is 96 ---> 1.3316687473990845%
missing value in q : is 96 ---> 1.3316687473990845%
missing value in soil : is 96 ---> 1.3316687473990845%
missing value in srad : is 96 ---> 1.3316687473990845%
missing value in tmax : is 96 ---> 1.3316687473990845%
missing value in tmin : is 96 ---> 1.3316687473990845%
missing value in vap : is 96 ---> 1.3316687473990845%
missing value in vpd : is 96 ---> 1.3316687473990845%
missing value in ws : is 96 ---> 1.3316687473990845%

```

Figure 4.7: Missing Values - Terraclimate

Scaling: Depending upon the scaling factor, the image is scaled up or down. The scaling factors are randomly sampled between the interval 0.6 and 1.4. This applies scaling in both height and width of the image.

Resizing: Random resizing of images can make the model robust to various image sizes. So, before training the images are resized randomly.

4.2.6 Other Basic Pre-processing Techniques

Missing Values: In the terraclimate dataset, most of the variables have missing values. These should be removed before training. Figure 4.7 shows the number of missing values present in each of the terraclimate variables

Outlier removal for Target Variable: Since frog count is the target variable, the outliers present should be removed. Such outliers can influence the prediction accuracy and can cause biased outcomes. Figure 4.8 shows that most of the frog counts are concentrated below 500. So, the data points that have count above 500 are removed.

4.3 Pseudo-Absence dataset

One of the research objective is to predict the presence/ absence of frogs at a location. For that we need absence data to know the characteristics of absence location. It is comparatively difficult to collect absence data of any species compared to collecting presence data.

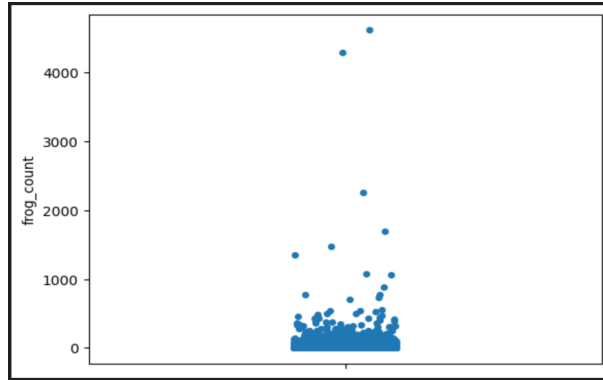


Figure 4.8: Frog Count - Outliers

This is because, if a species is not observed at a particular location, it does not necessarily be considered as true absence point. There are multiple reasons for the species to not be present at a location [59].

- The species could have been present but was not recorded by the observer due to human error.
- The species could have been extinct at that particular location due to human activities or due migration in spite of being environmentally suitable.

Considering these difficulties researchers sort to other methods of obtaining the absence data. The generated data is called pseudo-absence data. This section describes a novel procedure created to obtain the pseudo-absence data. The factors that are taken into account for generating the pseudo-absence data that could potentially have an impact on the accuracy of the SDMs are:

1. Number of pseudo-absence points (i.e. the ratio of presence points to pseudo-absence points). Since having a large number of pseudo-absence points compared to the presence points could make the dataset imbalanced and result in a biased performance.
2. Covariates chosen to filter the data.
 - Geographical extent (Distance criteria)
 - Land cover patches

The motivation for selecting the above covariates is that, there is a high probability that the person who observed the presence of frog at a location would have also been present at a location close to the presence point (distance criteria). If the land cover of those points are similar to that of the presence point, then there is a high chance that the observer have not observed frog presence at those points. This is the reason for selecting the mentioned pseudo-absence technique. The steps followed to obtain the pseudo-absence points are:

1. Extracting potential pseudo-absence points: First of all, the whole study area is divided into 30sqkm grids and from this, the existing grids for which the presence data available are separated. The remaining grids constitute the potential pseudo-absence points.

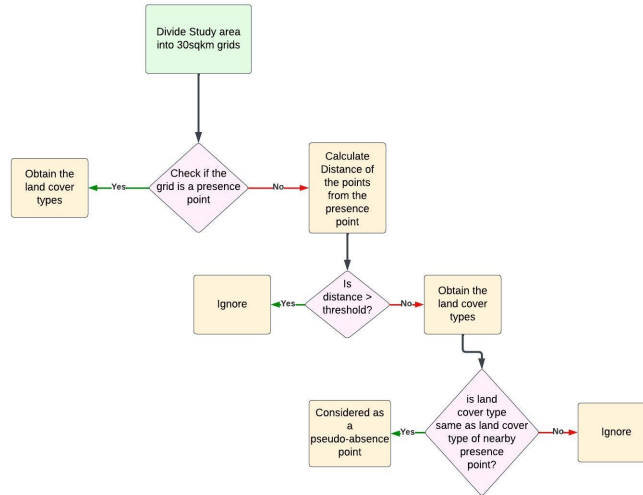


Figure 4.9: Pseudo-Absence Data Generation

2. From the known presence points, select only those points that are located at some X distance. The distance between two points are calculated using the haversine formula by utilizing the coordinates of these points[50]. Haversine formula is shown in equation 4.5. The threshold set for distance X is different for three countries. Considering the first factor explained earlier, since Australia has more presence points compared to other two countries, fixing the same threshold will result in large number of pseudo-absence points. So, the threshold is set as 10 Kms, 20 Kms and 28 Kms for Australia, South Africa and Costarica respectively. The points obtained from this step for all three countries are shown in figure 4.10,4.11,4.12 (example patch).
3. Identify the land cover types present at the presence and the points obtained from the previous step. For this, the Esri 10-meter land cover dataset can be utilized.
4. If the points which are present at x distance (obtained from step 2) contain the same land cover as the presence points, then those points are a possible pseudo absent point. Because these points have a high possibility of being visited by a citizen scientist and probably those points don't have the presence of frogs.

$$d = 2r \sin^{-1} \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\psi_2 - \psi_1}{2} \right)} \right) \quad (4.5)$$

Where,

d is distance between 2 points , r is radius of Earth
and ϕ and ψ are latitude and longitude

4.4 Dataset for GNN Approach

The task of classifying a location as presence / absence is tried as a node classification problem in GNN. The nodes present in a graph are made up of data points which are

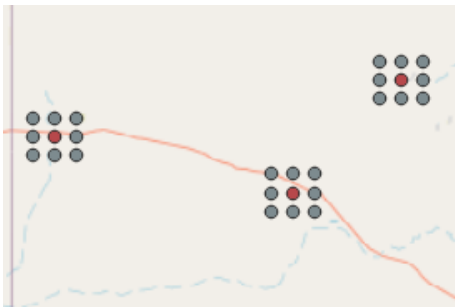


Figure 4.10: Australia Pseudo Absence point. Red - presence point, Grey - Pseudo Absence point

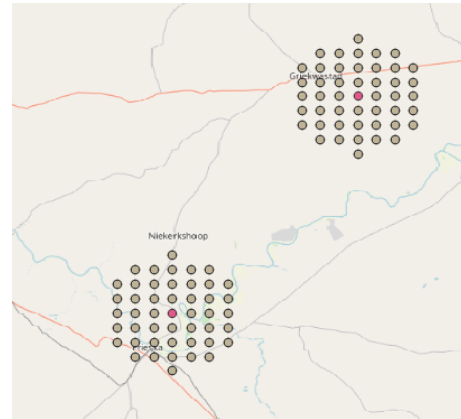


Figure 4.11: South Africa Pseudo Absence point. Red - presence point, Grey - Pseudo Absence point

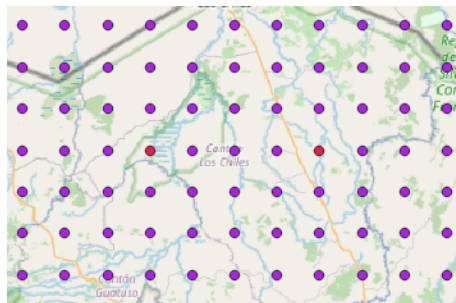


Figure 4.12: Costa Rica Pseudo Absence point. Red - presence point, Violet - Pseudo Absence point

Figure 4.13: Pseudo Absence Points Comparison

one of presence or absence location. In a graph, a node is represented not just by its own features but also by the features of the neighboring nodes. So, the idea is to make use of the information of the neighboring nodes also in order to classify a particular node as presence / absence. In order to obtain information about the neighboring nodes, message passing is incorporated. Message passing is a technique which each node uses to obtain information from its neighbours and update its embedding.

Since it is a relatively new approach used in building a SDM and also considering time constraint, only one modality of data is used (Tabular) and on only one country data (Costarica). The graph dataset is created by making a edge list. And the feature vectors are made up of terraclimate data. The edge list represents the connectivity of nodes in the graph. The edge list is formed based on the distance criteria. The threshold for the distance is chosen as 10 kilometers. So, for a specified node, 2 nodes will be a neighbour in both horizontal and vertical direction, while 1 node will be a neighbour in diagonal direction. Refer figure 4.14 for the representation of a edge list.

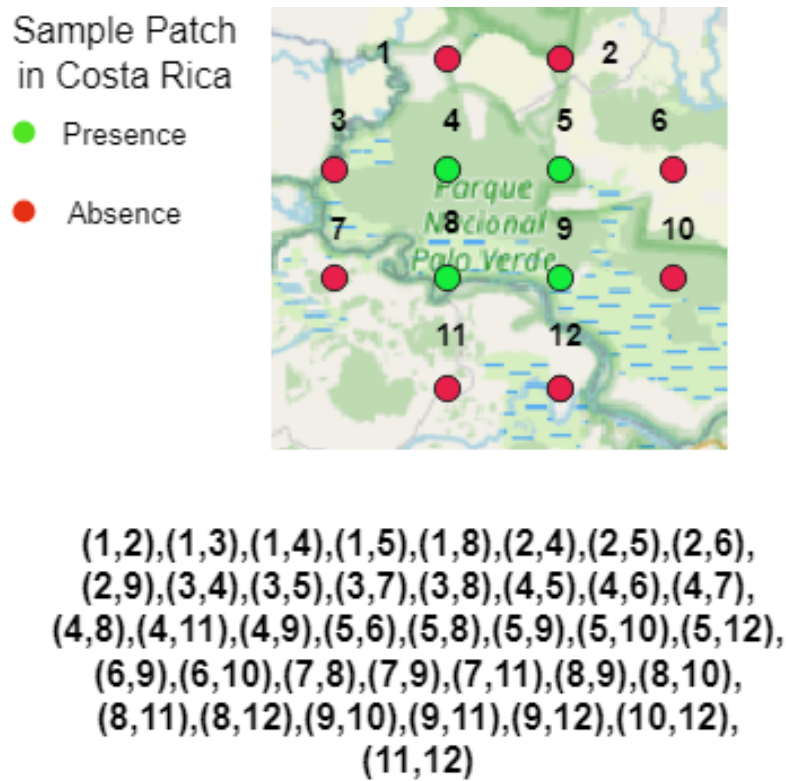


Figure 4.14: Edge List - Sample Points

4.5 Model Architecture

This section describes the model architecture used in building the SDM. This section is subdivided into two according to the nature of the problem in hand. The first part describes the pipeline used for the purpose of frog counting tool which is a regression task and the second part is for classification of the location into presence or absence of frogs conducted.

4.5.1 Model for Frog Counting Task - Regression

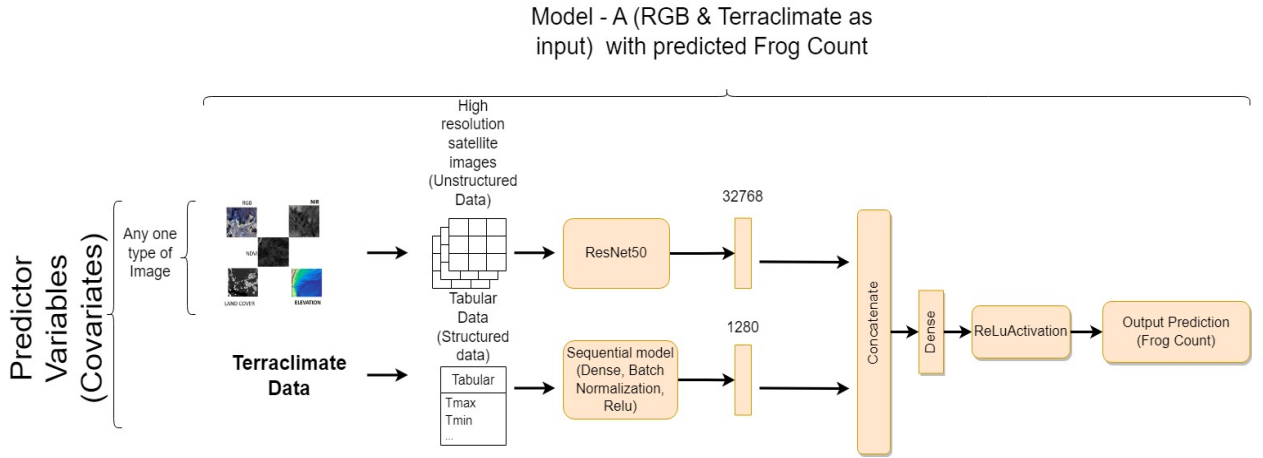


Figure 4.15: Pipeline Architecture - Frog Counting Challenge

In general multimodal learning involves architecture that is able to combine input data from multiple modalities. The fusion module is where the combination happens where a single representation of different modalities is obtained that will be used for regression or classification tasks. As outlined in section 3.2, several fusion methods are present like early fusion, late fusion, weighted fusion and attention mechanism based fusion etc.

In this work late fusion is applied. This is because each modality will have its own neural network for feature extraction. This flexibility of having data specific neural network will be influential in capturing the input data's feature more effectively. Since, each modality has its own path for processing, unique characteristics and features of each data source can be captured with high accuracy.

The architecture used for the purpose of counting frogs in a particular location is somewhat similar to the one used in [77]. The exact architecture used is shown in figure 4.15. The model is a fusion model where it combines information from two different input sources. The model takes input from two modalities of data, one is images which constitutes high resolution satellite images and the other is of numerical or tabular data which consists of terraclimate data. These inputs make up the predictor variables or covariates. The model is made up of two branches. Input images are fed to a ResNet50 [31] model pre-trained on imagenet dataset [17]. Output of the ResNet50 model consists of the features extracted from the image inputs and it is flattened to get 32768 features. The output from the sequential model consists of a flattened array of 1280 features of numeric data. Both the obtained features are concatenated to get the combined features of both inputs of length 34048. It is then passed through a dense layer and finally through Relu Activation to get the predicted frog count.

The model shown in figure 4.15 takes one set of input data which consists of RGB image and terraclimate image. Similarly, two other different models are used which takes a different image as input. For example instead of RGB, Model-B takes land cover patches and terraclimate data as input and Model-C constitutes inputs from NDVI and terraclimate data.

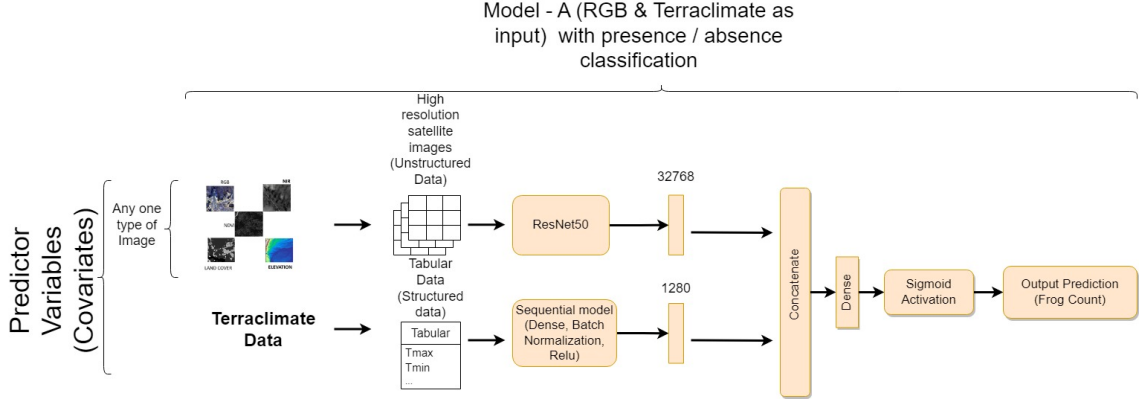


Figure 4.16: Pipeline Architecture - Frog Presence/Absence Classification

4.5.2 Model for Frog Presence/Absence Classification

The architecture used for the classification of area based on the presence or absence of frogs is similar to the one used for counting problem, but with the use of softmax activation at the end. The output predicts either presence or absence of frogs. Similar to counting problem, this task was also implemented for three sets of data (RGB & terraclimate, Land cover & terraclimate, NDVI & terraclimate). The architecture is shown in figure 4.16

4.6 Experimental Setup

4.6.1 Validation Metrics

Model Architecture

As mentioned earlier, the choice of architecture chosen for both the tasks is ResNet50. The model is pre-trained on ImageNet dataset. ResNet is chosen because of its effectiveness against vanishing gradient problem [53]. Kaiming et.al., in [31] introduced deep residual learning framework that consists of *skip connections* to negate the vanishing gradient problem. Due to its effectiveness in extracting features of images, Resnet50 is chosen.

Loss Function

For the frog counting task, the loss function used is Mean Squared Logarithmic Error (MSLE). Since the target variable is a wide range of continuous numbers, MSLE treats small differences between actual and predicted value the same as big differences between actual and predicted values. MSLE is calculated using the formula given in equation 4.6

$$MSLE = \frac{1}{N} \sum_{i=1}^N (\log(y_i + 1) - \log(\hat{y}_i + 1))^2 \quad (4.6)$$

Where,

N is Number of data points, y represents true value, and \hat{y}_i represents predicted value.

For the purpose of frog presence/absence classification task, Binary Cross-Entropy Loss is used. Since, it is a binary classification problem, the Binary Cross-Entropy Loss

| Task | Input Data Type | Optimizer | Loss Function | Learning Rate | No of Epochs | Batch Size |
|------------------|-----------------|-----------|---------------|---------------|--------------|------------|
| Frog Counting | RGB & Numeric | Adam | MSLE | 0.005 | 1500 | 16 |
| | LC & Numeric | Adam | MSLE | 0.001 | 2000 | 16 |
| | NDVI & Numeric | Adam | MSLE | 0.005 | 1500 | 16 |
| Presence/Absence | RGB & Numeric | Adam | BCE | 0.01 | 1000 | 16 |
| | LC & Numeric | Adam | BCE | 0.02 | 1000 | 16 |
| | NDVI & Numeric | Adam | BCE | 0.01 | 1000 | 16 |

Table 4.2: Overview of Training Parameters.

function measures the difference between predicted probabilities and actual binary labels. The mathematical formula is given in equation 4.7

$$\text{Binary Cross-Entropy Loss} = -\frac{1}{N} \sum_{i=1}^N (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)) \quad (4.7)$$

Where,

N is Number of data points, y_i represents true value, and p_i represents predicted probability.

In addition to the loss function, in order to prevent overfitting, regularization is employed while training. Specifically, L2 regularization also called as ridge regression is used where the square of the magnitude of coefficient is added as a penalty to the loss function.

Optimizer and Learning Rate

The optimizer used in training is Adam. Eventhough Adam handles learning rate optimization on its own for each parameter in the model, a learning rate scheduler is used as a warm-up phase. This warm-up phase results in a smooth transition from the initial set learning rate to the target value.

Software and Hardware

For training the model and to conduct various experiments the GPUs and CPUs with the following properties are used.

- GPU: NVIDIA A16 and a CPU with 72 cores and 256 GB of memory.
- NVIDIA Quadro RTX-6000 with 24GB of memory.
- NVIDIA GeForce RTX-2080ti with 11GB of memory.

The following software and the version are used:

- CUDA version 11.3
- Python version 3.10.11
- TensorFlow version 2.11.0

4.6.2 Implementation details - Experiments Conducted

Balancing the Frog count dataset

As said earlier in section 4.2.3, balancing the dataset appropriately is one of the areas which is rarely explored in building SDMs. This experiment is conducted to analyse the impact of the balanced dataset by comparing the performances of the model trained on both the original imbalanced dataset and balanced dataset. The trained model is used to make prediction on the submission data to know the differences in their performances.

Performance comparison between proposed method and winner of frog counting challenge (model-w)

RQ3: Performance of proposed multimodal based SDM. In order to answer the research question about the performance of the developed method, this experiment is conducted. The training data consists of grids of size 30sqkm split into train and test data. As said earlier in section 4.5, the model is trained on three sets of data separately. In order to obtain the model performance and make a fair comparison with model-w , the trained model is tested on the submission data provided by [6]. The submission data consists of grids of size 225 sqkm. So, the model prediction is done using two approaches.

Sliding Window approach Since the model is trained on images of size 512*512 and

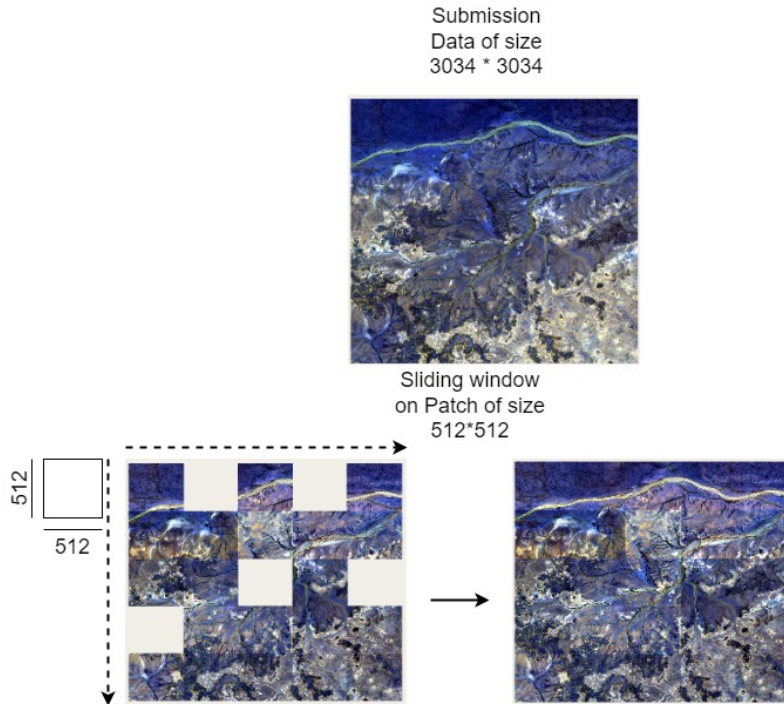


Figure 4.17: Sliding Window

the image on which the prediction done on is of size 3034*3034, sliding window method is employed while predicting. The window size is set to 512 and the stride is set equal to the window size to avoid overlapping. The model is then used to make prediction on the current window, starting from top left and slides through the entire image. The predictions from each window are accumulated and then finally the sum of all the predictions constitutes the prediction of the current location. An illustration of the approach is shown in figure 4.17

Resizing approach This approach simply resizes the image on which predictions are done to the same size as the training images.

Predictions are done using both the approaches and the results obtained using all three models are discussed in the next chapter. The results obtained by model-w is used as a baseline model and compared with the result obtained here.

Weighted Average Ensemble

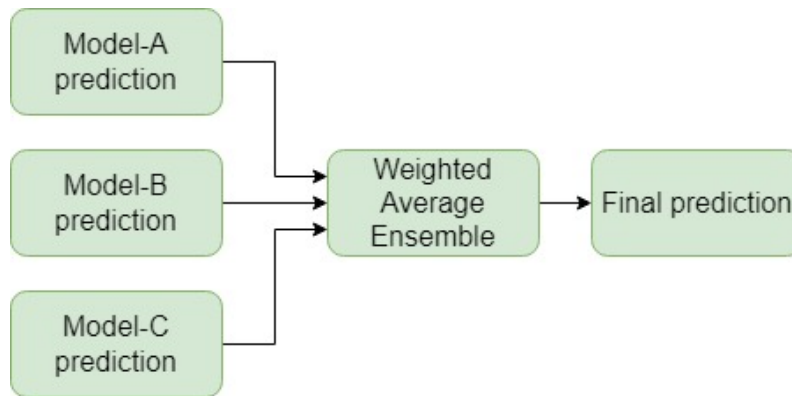


Figure 4.18: Weighted Average Ensemble

In general ensemble method combines the predictions of all three models to give one optimal prediction. When it comes to weighted average ensemble, contribution of each model to the final prediction is weighted by the performance of the model. The formula for weighted average prediction is given in equation 4.8

$$y_i = \frac{(W_a * P_a) + (W_b * P_b) + (W_c * P_c)}{W_a + W_b + W_c} \quad (4.8)$$

Where,

y_i is prediction of ensemble model. W_a , W_b , W_c are weights of model a, b and c respectively. P_a , P_b , P_c are predictions of model a, b, and c respectively.

For finding the optimal weights for each model, the *scipy.optimize* library is used. The following steps explains the procedure to find the optimal weights.

- Start with assigning equal weights (0.3,0.3,0.4) to all three models such that it sums up to 1.
- Calculate the weighted average predictions using the equation 4.8.

- Calculate the MAE between the true value and the obtained predictions.
- The minimize function from *scipy.optimize* is used to find the optimal weights that minimizes the MAE.
 - This function takes arguments such as the ensemble metric (MAE), initial weights, bounds, optimization constraints and the optimization method.
 - The optimization constraints are defined such that the sum of the weights are equal to 1.
 - The bounds for the weights are set to be between 0 and 1.
 - The optimization method is chosen as "Sequential Least Squares Quadratic Programming" (SLSQP). This iteratively updates the weights to minimize the MAE while satisfying the constraints and the bounds. It works by minimizing the sum of squared differences between the predicted and the actual values.
- The algorithm terminates when the obtained solution (weights) converges to a minimum value satisfying the constraints.

In addition to the one-to-one comparison of all three models separately with model-w, predictions obtained using ensemble of all three models are also compared.

Performance evaluation using cross validation

K-fold cross validation is performed to test how well the model generalizes on different datasets. In k-fold cross validation, the entire dataset is divided into k sets. The model is trained on k-1 sub-sets and evaluated on the remaining set. This is repeated for k times, for every combinations of the subsets used for training and testing. This estimates how well the model performs on unseen data. For this purpose, the value of k is chosen as 5. And only LC and numeric data are used to evaluate the model using cross validation. However, this experiment could not be finished on time as each cycle of training on 4 training sets (k-1) and evaluating on 1 test set took more than one hour. And only around 80 epochs got over, that too for only one set. So, a conclusive result on cross validation MAE could not be provided. Nevertheless, whatever the results obtained as of writing this has been mentioned in the appendix.

Feature Importance Assessment

RQ4: Importance of different covariates in contributing to the target value.

The terraclimate dataset consists of 14 parameters in total. It is vital to understand the parameter that has high influence in the frog presence/absence in a location. This importance value will contribute to the research of the impact of global warming and climate change in the migration of frogs. Feature selection means the technique used to select a subset of the relevant features from all the existing features. From the efficiency of the model aspect, the less the features are the more efficient the model is in terms of space and time. And also having irrelevant features can guide the model in a wrong way resulting in worse prediction results. For this experiment an algorithm called Recursive Feature Elimination (RFE) is employed. RFE recursively fits the model and ranks the features based on their importance. The algorithm begins by feeding in all the 14 parameters of terraclimate data and discarding the least important features one by one until the desired number of features remain. In order to rank the importance of features, Random Forest Regressor model is used. The desired number of features is selected as 10. Here the target

value represents the frog count. The importance value of each value are discussed in the results chapter.

Generalization

Inorder to test the proposed model's generalization capacity, this particular experiment is conducted. This experiment will test how well the model performs in predicting the frog count of a location when it comes to real world scenarios and the model's ability to extrapolate its knowledge to unseen environments. Since there are ground truth data available for three different countries, the model will be trained on one country and then predictions will be made on a different country. For this purpose, the model will be trained on data of Australia and evaluated on Costa Rica.

Performance Evaluation of the pre-processed terraclimate data using XGBoost

The terraclimate dataset used for training the fusion model, obtained after performing the pre-processing step mentioned in section 4.2 is evaluated by training using XGBoost technique. The winner of the challenge used XGBoost to evaluate their terraclimate data. So, to compare the two terraclimate data, XGBoost is chosen. XGBoost stands for "Extreme Gradient Boosting" based on gradient boosted regression trees. This algorithm has been used frequently on many kaggle data science competitions. Only one modality of data i.e. tabular data is used here to predict the frog count. The results obtained are discussed in the next chapter and compared with the results of the winner.

Presence/Absence Classification

The task of predicting the presence/absence of frogs in a location is handled as a binary classification problem. The presence points are considered as positive class and the absence points are considered as negative class. The positive class is made up of data points from the original presence dataset. And the negative class is obtained by the pseudo-absence points generated as discussed earlier in section 4.3. The model and data used for training is similar to the frog counting task. Unlike the frog counting task, there are no separate submission data available on which the model can be used to make predictions. So, 20% of the available data points is separated from the training data to finally make the predictions. The results obtained are discussed in the next chapter.

Comparison of Different pseudo-absence data generation method

Since, one of the contribution of this thesis is to provide a credible way to generate the pseudo-absence data, in order to analyse how accurate the proposed method is, comparison with the existing method in the literature [59] is performed. Due to time constraint, only two methods are chosen to compare with and only the land cover type input data is used. The two methods of generating the pseudo-absence data used for comparison purpose are:

- **Random Selection:** From the available data points within the study area, the presence points are separated. From the remaining points, which constitute potential absence points, the pseudo-absence points are randomly chosen. The number of randomly chosen points are selected such that it balances with the presence points.
- **Distance Criteria:** The potential absence points are restricted by the distance threshold. Only the points which are located within the threshold distance are chosen.

The same distance chosen as the threshold in the proposed method are selected here as well. And from those points the pseudo-absence points are randomly selected.

The pseudo-absence points generated by these two methods are used to train the fusion model and the results are compared with the proposed method. The obtained results are explained in the next chapter.

Chapter 5

Results

The results of the experiments discussed in the previous chapter are presented here. Before that the evaluation metric chosen to evaluate the performance of the models used in both the tasks are explained.

5.1 Evaluation Metric

5.1.1 Mean Absolute Error (MAE) for Frog Counting Task

Since the task of predicting the count of frogs is a regression task, it is only logical to use Mean Absolute Error (MAE) as evaluation metric. MAE measures the average absolute difference between the actual and predicted values. In simple terms, MAE tells us how much the predicted value is deviated from the actual value. The larger the MAE is the worse the performance is. And also, MAE was used to evaluate the performance of model-X. So, to get a direct comparison between the proposed model and model-X, MAE is chosen. MAE is calculated using the formula given in equation 5.1

$$MAE = \frac{|(y_i - y_p)|}{n} \quad (5.1)$$

Where,

n is number of observations , y_i represents true value,
and y_p represents predicted value.

For comparison with the model-w, F1 metric is also used to evaluate the model.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5.2)$$

$$Precision = \frac{TP}{TP + FP} \quad (5.3)$$

$$Recall = \frac{TP}{TP + FN} \quad (5.4)$$

Where,

TP is True positives , FP is False positives
and FN is False negatives.

5.1.2 Accuracy and AU-ROC score for Presence/Absence Classification

For the classification task, accuracy is chosen as the evaluation metric. Accuracy is calculated as the ratio of correctly predicted instances to the total instances in the dataset. In addition, the area under ROC curve is also used as evaluation metric. The dataset used for this classification task is balanced to some extent, i.e. in positive class, 6580 sample points are present, whereas in negative class 5589 sample points are present. When it comes to accuracy, it is heavily threshold dependent, so to minimize the influence of threshold, AUC score is also chosen as an evaluation metric.

5.2 Experiment Results

5.2.1 Balancing Dataset

The results obtained before and after balancing the dataset is shown in table 5.1. For comparison purpose, only the model using RGB & Numeric data is shown. The MAE for the model using imbalanced data shows a very high value of around 190 for both training and testing. From figure 5.1, we can observe that after epoch 2, the curve started to saturate and did not go down from there. This shows the inability of the model to learn. After balancing the dataset following the procedure mentioned in section 4.2.3, the model was able to achieve a relatively lower MAE and the model was able to learn well, which is evident from figure 5.2. The MAE obtained after around 500 epochs is around 9 and 29 for the train and test data respectively. As a consequence of balancing the dataset, the model now has more representation of data from the minority range. This eventually resulted in reduced MAE which is expected, as with any balancing technique the model performs better naturally.

| Input Data | Dataset | Train MAE | Test MAE | Submission MAE |
|---------------|-----------------------|-----------|----------|----------------|
| RGB & Numeric | Original (Imbalanced) | 189.0413 | 189.2232 | 208.23 |
| RGB & Numeric | Balanced | 8.44 | 28.82 | 36.25 |

Table 5.1: Results - MAE Comparison between Balanced and Imbalanced Dataset

5.2.2 Performance Comparison

The results obtained by using sliding window and resizing approach on the submission data and also the train and test results is given in table 5.2

From table 5.2, we can notice that the model with input data type of LC & numeric data performs well compared to that of other data types. The second best model is the one which utilizes NDVI & numeric data. The reason for this visible difference in the performance could be due to the nature of the data. As mentioned earlier in section 4.1.1, the Esri 10-meter land cover dataset separates the study area into 10 classes of land types. This type of data can be considered as a form of semantic segmentation. This results in the area being distinguishable from one point to another. This makes the model learn the features easily compared to the other types of data. When it comes to NDVI, particular study area is defined by a value in the range (-1 to 1) which gives us the vegetation health

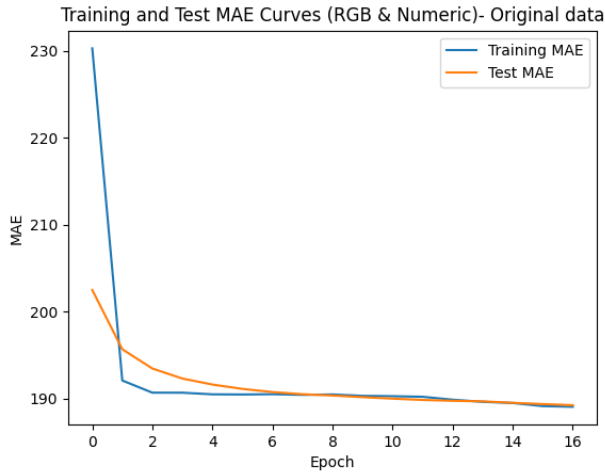


Figure 5.1: Train and Test MAE curves for Im-balanced data

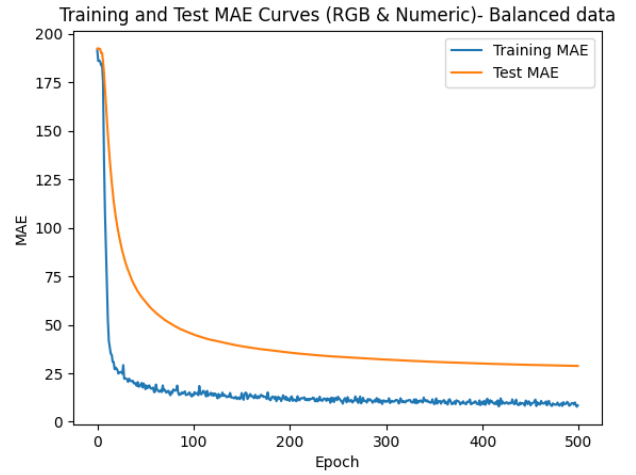


Figure 5.2: Train and Test MAE curves for balanced data

| Input Data | Train MAE | Test MAE | Submission MAE | |
|----------------|-------------|--------------|----------------|--------------|
| | | | Sliding Window | Resizing |
| RGB & Numeric | 8.44 | 28.82 | 39.15 | 36.25 |
| LC & Numeric | 5.06 | 12.08 | 35.42 | 30.18 |
| NDVI & Numeric | 3.99 | 12.17 | 38.34 | 33.74 |

Table 5.2: Train and Test MAE obtained for the frog counting task on three sets of input data and the MAE obtained using sliding window and resizing approach on the submission data

of the area. Similar to the case of Esri dataset, the model can learn the features easily compared to RGB dataset.

Figure 5.3 shows how the pixel values differs. The whole of green patch in the RGB data refers to one class in the land cover data and in the case of NDVI, the same area will have a value close to 1, indicating the high health of the vegetation. In both cases, the models benefits from the nature of data, that results in certain features easily distinguishable.

When the prediction is made on the submission data, which consists of grids of size 225sqkms, the MAE obtained is higher than the test and train MAE. This might be due to the difference in the size of images used in training, which consists of grids of size 30sqkms. So, while predicting using two approaches to cover the size difference, resizing approach yielded better result than sliding window approach. This is because, while performing sliding window approach, the window size might not fit perfectly over the entire image and there will be some part missing or overflow. This could lead to information loss or redundancy, affecting the model's performance. And also due to the fact that the terraclimate data (numeric data) is available for the entire patch of 225sqkms and not for sub-patch of smaller size, the terraclimate data is made up of one set of values for the whole area. All these differences result in the slight performance distinction between the

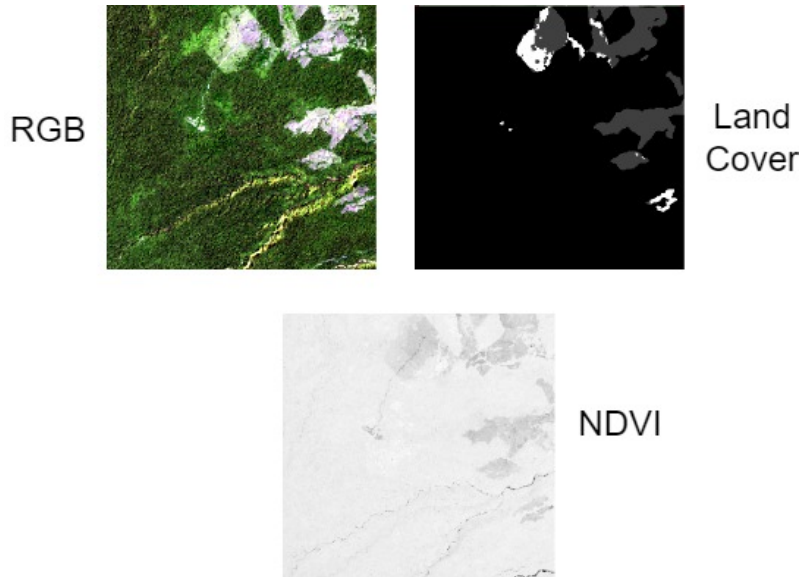


Figure 5.3: Sample patch that shows the distinction between RGB, LC and NDVI data of the same location

sliding window and resizing approach.

To further improve the current result, **log transformation** was applied as discussed in section 4.2.4. The results obtained after applying log transformation are shown in table 5.3. Only test and submission MAE (resizing approach) are shown.

| Input Data | Test MAE | Submission MAE |
|----------------|--------------|----------------|
| RGB & Numeric | 27.72 | 36.09 |
| LC & Numeric | 10.19 | 28.49 |
| NDVI & Numeric | 11.78 | 32.87 |

Table 5.3: Results - After Log Transformation

The test and submission MAE reduced slightly after applying log transformation compared to the results obtained earlier. However, the **inverse of log transformation** has to be taken to get the predictions back to the original scale.

Ensemble Method

The performance differences between the models using three different types of data is tried to overcome by weighted average ensemble. The weights are assigned according to the procedure described in section 4.5.2 under weighted average ensemble. The model using LC is assigned higher weights due to its better performance compared to the other two models. The results obtained on the submission data is given in table 5.4. The predictions obtained using only the resizing approach is taken because of its superior performance obtained compared to the sliding window approach.

Clearly, from the table 5.4, it can be seen that the MAE reduced considerably from the previous results obtained from separate models. The MAE reduced by almost 9, which

| Ensemble | Weights | MAE |
|------------------|---------------------------------|-------|
| Weighted Average | RGB - 0.1, LC - 0.6, NDVI - 0.3 | 21.94 |

Table 5.4: Results - Weighted Average Ensemble

shows the effectiveness of the ensemble approach. Observe figure 5.4, to visualize how the MAE has reduced considerably using the ensemble approach.

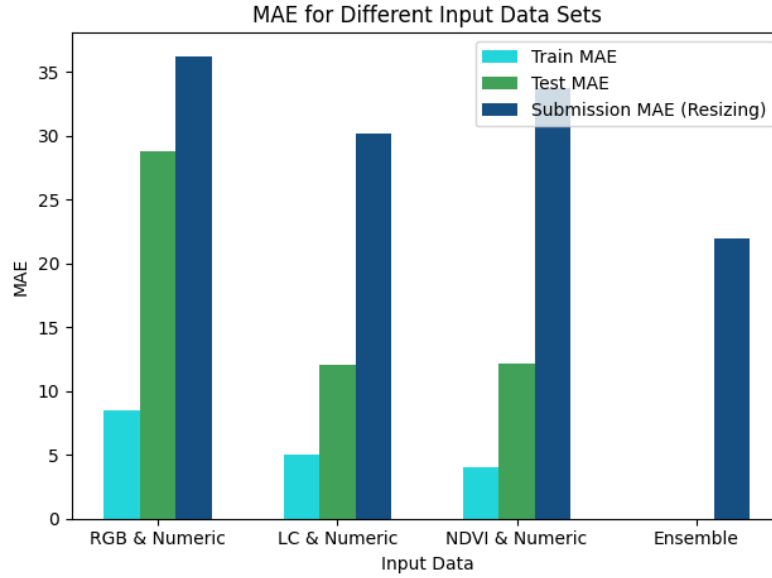


Figure 5.4: Results - Comparison of various methods

5.2.3 Feature Importance Assessment

| S.no | Feature | Importance |
|------|---------|------------|
| 1 | Tmax | 0.55 |
| 2 | Tmin | 0.16 |
| 3 | Pet | 0.08 |
| 4 | Ppt | 0.05 |
| 5 | Vap | 0.05 |
| 6 | Vpd | 0.04 |
| 7 | Soil | 0.03 |
| 8 | Ws | 0.03 |
| 9 | Q | 0.01 |
| 10 | Pdsi | 0 |

Table 5.5: Terraclimate Feature Importance

The results obtained by performing RFE, to rank the features according to their importance is shown in table 5.5. The results obtained are from the fusion model trained using LC & Numeric data. From the table we can infer that almost 80% of the parameters have negligible contribution towards the target variable. So, in order to test the MAE obtained using only the top 6 important features, the model was trained using LC & Numeric data. The trained model was used to get prediction on the submission data. Apart from the MAE, the inference time is also measured while predicting on the submission data to see if the model performs any faster than the original model with all 10 features. The results are shown in table 5.6. Refer table 4.1, for the different feature names.

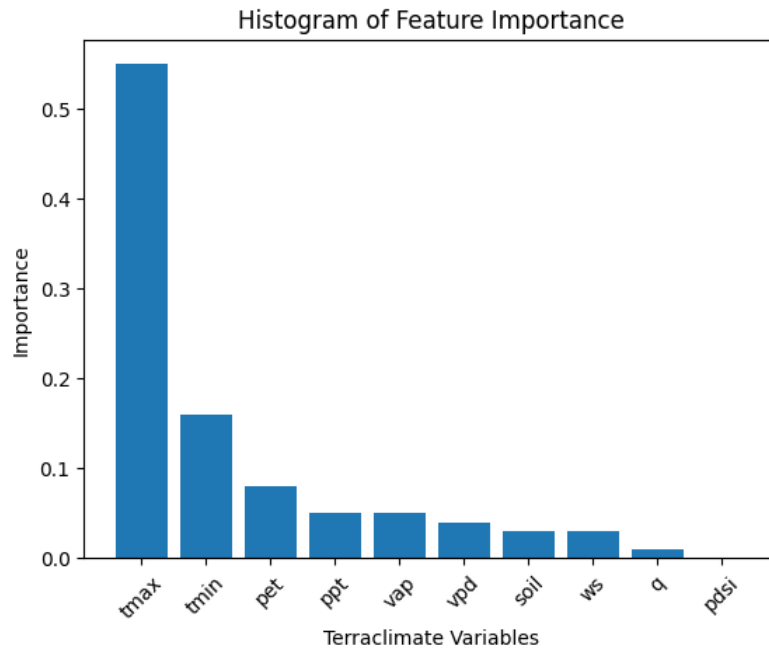


Figure 5.5: Feature Importance - Histogram

| Input Data | No of features | Train MAE | Test MAE | Submission MAE | Inference Time(sec) |
|--------------|----------------|-------------|--------------|----------------|---------------------|
| LC & Numeric | 10 | 5.06 | 12.08 | 30.18 | 0.064 |
| LC & Numeric | 6 | 6.15 | 14.23 | 30.58 | 0.058 |

Table 5.6: Results - MAE for top 6 features

From table 5.6, we can see that there is no significant difference in the MAE between the models using all the features and only the top 6 features. However, it can be seen that there is a slight improvement in the inference time while using only top 6 features. Eventhough the difference is very minute, for larger datasets and more complex model, assessment of the feature importance can lead to a better efficient model in terms of inference time and space.

5.2.4 Generalization

The fusion model was trained only using the Australian dataset, and the trained model was used to predict the frog count in Costa Rica. The results are shown in 5.7. The table shows that similar to the results obtained earlier in section 5.2.2, LC & Numeric input data type produced better performance. And the MAE achieved was comparable with that of the model trained on data of all three countries. This proves that the model is able to generalize well and is able to make good predictions on unseen data. The ensemble method was applied using almost similar weights on all three models. Because, unlike the results shown in section 5.2.2, all three models produced similar MAEs. So, the models are assigned almost equal weights by the optimization procedure described earlier in section 4.5.2. This method yielded even better MAE of 19.85, which is almost 40% reduction compared to the separate models.

| Input Data | Submission MAE | |
|----------------|----------------|--------------|
| | Sliding Window | Resizing |
| RGB & Numeric | 36.78 | 32.57 |
| LC & Numeric | 33.35 | 32.12 |
| NDVI & Numeric | 34.98 | 32.73 |

Table 5.7: Results on Costa Rica using the model trained on Australian data

| Ensemble | Weights | MAE |
|------------------|--|-------|
| Weighted Average | RGB - 0.25 , LC - 0.40 , NDVI - 0.35 . | 19.85 |

Table 5.8: Results of Ensemble method on Costa Rica using the model trained on Australian data

5.2.5 Performance Evaluation of the pre-processed terraclimate data using XGBoost

XGBoost is trained and evaluated on both the test data (training data split into train and test) and the publicly available submission data. However, only the MAE obtained on the training data by the challenge winner is available. Refer table 5.9, for the comparison between proposed dataset and the winner dataset. From the table, it is observed that the proposed dataset slightly outperforms the dataset used by the winner. XGBoost model is used only on terraclimate data for two reasons, the first one is to compare the terraclimate data used by the winner and the one used in this work. The second reason is that XGBoost is more suited for tabular and structured data. However, for images to be used, the features must be extracted first by a suitable CNN model and then use XGBoost for prediction. This is one of the reason why XGBoost is only used on tabular data.

| Model | Evaluation Dataset | MAE (proposed dataset) | MAE (Winner dataset) |
|---------|--------------------|------------------------|----------------------|
| XGBoost | Test data | 10.81 | 10.95 |
| XGBoost | Submission data | 27.16 | Not available |

Table 5.9: Results of XGBoost model trained on terraclimate dataset

5.2.6 Presence / Absence Classification

The classification accuracy and the AU-ROC score achieved in predicting the frog occurrence is shown in table 5.10. LC & Numeric data achieved highest accuracy in classifying the location as presence / absence. The ROC curve obtained for LC and NDVI data type are shown in figure 5.6 and 5.7 respectively. Both the data type achieved similar AUC score. Unlike the frog counting task, the trained model is evaluated on patches of the same size as the training data. So, there is no need for sliding window and resizing approach on the evaluation data.

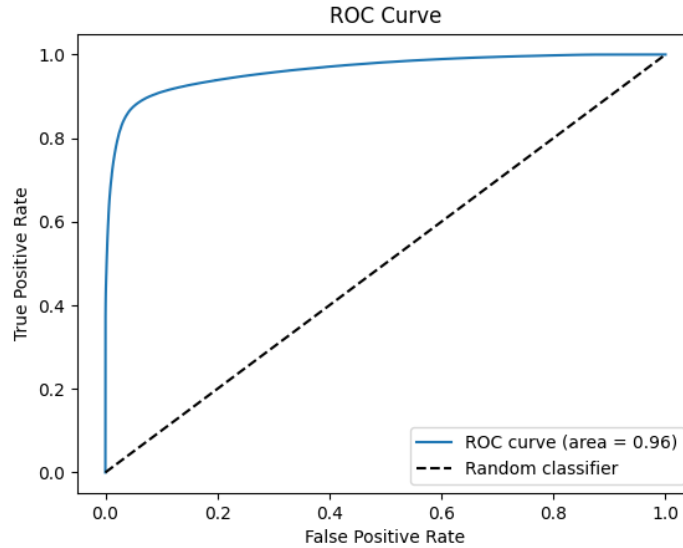


Figure 5.6: AU-ROC Curve for LC & Numeric data

| Method | Train Accuracy (%) | Test Accuracy (%) | Submission Accuracy (%) | AUC Score |
|----------------|--------------------|-------------------|-------------------------|-------------|
| RGB & Numeric | 79.07 | 76.18 | 75.78 | 0.82 |
| LC & Numeric | 91.06 | 90.9 | 89.19 | 0.96 |
| NDVI & Numeric | 90.9 | 90.8 | 88.4 | 0.96 |

Table 5.10: Classification Accuracy for Frog Presence / Absence detection

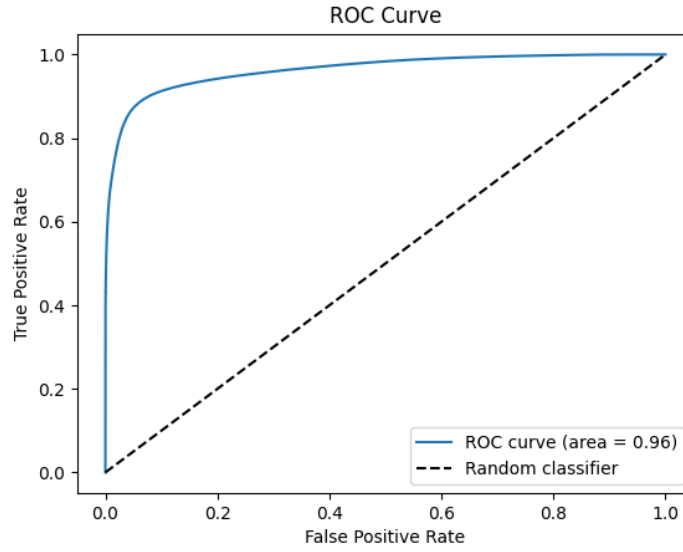


Figure 5.7: AU-ROC Curve for NDVI & Numeric data

5.2.7 Comparison of different pseudo-absence data generation method

The AUC score and the accuracy achieved on the two existing methods of generating pseudo-absence data is compared with the proposed method. Only the LC & Numeric data type are chosen for comparison. The results suggest that the proposed method performs better than the other two method. While a huge difference can be seen between the proposed method and the random selection method, pseudo-absence data generated by the distance criteria method closely follows the proposed method. The reason for this can be attributed to the selection criteria chosen for generating the pseudo-absence data. While the proposed method selects the absence points based on both geographical extent and land cover type, the distance criteria takes into account only the geographical extent. So, there will be some correlation between the two methods. While there is no ground truth data available to evaluate the model, the submission dataset is made up of data generated by all three methods, to ensure fairness in the prediction.

| Method | Train Accuracy (%) | Test Accuracy (%) | Submission Accuracy (%) | AUC Score |
|-------------------|--------------------|-------------------|-------------------------|-------------|
| Proposed Method | 91.06 | 90.09 | 84.87 | 0.90 |
| Random Selection | 72.29 | 70.18 | 65.93 | 0.68 |
| Distance Criteria | 90.47 | 90.16 | 80.18 | 0.88 |

Table 5.11: Pseudo-absence data performance comparison

Chapter 6

Discussion

In this chapter the results obtained in the previous chapter are discussed in detail and compared with the appropriate model.

6.1 Results - Analyses and Discussion

6.1.1 Dataset Balancing

Balancing of dataset resulted in overcoming the issues that usually affect the model that occur due to imbalanced data. These issues were explained in section 4.2.3.

Bias in model performance: Though the model performed slightly lower on the submission data, the balancing method proposed in this work seems to work well, considering the initial very high MAE obtained from imbalanced dataset . Still there is room for improvement in balancing the dataset, considering the relatively lower MAE achieved on lower frog counts.

Generalization: The data available for Australia was very large compared to that of other two countries. This is also an imbalanced data that was discussed in section 4.2.3. However, weighted loss function was used to mitigate this issue. Due to this the model was able to generalize well, which is evident from the results obtained from the generalization experiment.

6.1.2 Performance Comparison

The results obtained using the proposed model is compared with model-w. The MAE obtained by model-w on test data is shown in figure 6.1. For comparison purpose the best performing model (ANN) is taken into consideration along with XGBoost.

| Model | MAE |
|------------------|--------------------|
| ANN | 9.357297531398874 |
| XGBoost | 10.950194889562582 |
| Lasso Regression | 13.941966219142486 |

Figure 6.1: Results - Model-W

Before comparing the performance of the proposed model with model-w, a scatter plot 6.2 is plotted to visualize the predictions made by all the models trained so far. This is

done to better understand where the predictions differ from the actual values (submission data). From the scatter plot, it can be observed that the predicted values were very close to the actual values when the frog count were below 100. And beyond that count the predicted value diverges from the actual value. The model performed better when it comes to lower value of frog count. So, all the trained model was used to predict the lower values of frog count only, which forms about 80% of the submission data.

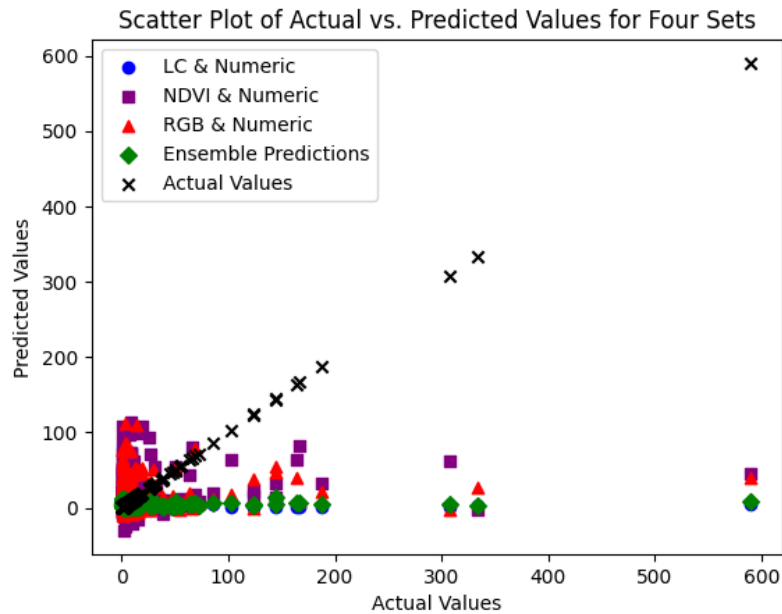


Figure 6.2: Scatter Plot on submission data

The results obtained for lower values of frog count is shown in table 6.3. The results shows a significant reduction in MAE. The corresponding scatter plot is shown in figure 6.5. It shows that the actual and predicted values are much closer to each other. However, the reason for this difference in performance can be attributed again to the dataset. Since there is considerably very low data points that have higher values of frog count, there is a limit for model learning. Eventhough the dataset balancing approach explained earlier resulted in a significant improvement in the performance compared to the original dataset, still it can be observed that there is a big gap when it comes to the distribution of the data.

In order to compare the results obtained with model-w, two approaches are undertaken. The results of model-w shown in figure 6.1, was evaluated on the data used for training and not on the submission data. So, for the comparison to be credible the results obtained by the proposed model on the test data is taken into consideration. Table 6.2 shows the best results obtained by both the models. From the table it can be observed that there is only a slight difference in MAE between both the models. As said earlier, the MAE obtained from the proposed model can be improved further by a better balancing technique for the dataset.

The second approach to compare the two models is to use F1-metric. Eventhough, the task is regression and using f1 score might not be the correct way to evaluate the model, it is used because the leader-board score of this contest uses f1-metric to rank the participants. Figure 6.3 shows the f1 scores of the top ranked participants. The winner achieved a score of 0.42. To evaluate the f1 score of the proposed model, the obtained predictions are separated into two classes. If the predicted value matches exactly with the

actual value, it is of one class and the remaining are of the second class. This way the number of times the model predicted correctly is known. And the leader-board score is evaluated on the submission data. The f1 score obtained using the proposed model using ensemble technique is shown in table 6.1. The best model obtained a score of 0.36, which is slightly lower than the leader.

| Model | F1-Score |
|----------------|----------|
| Ensemble Model | 0.36 |

Table 6.1: F1-Score of Ensemble model
















| Team | Score | Region | Location |
|---|-------|----------|---|
|  Dalyan Ventura | 0.42 | EMEIA |  France |
|  Never Gonna Give You Up | 0.35 | APAC |  Hong Kong |
|  Germano Lima | 0.26 | Americas |  Brazil |
|  Tanla Sadhani | 0.24 | APAC |  Australia |
|  Ze Xuan Ma | 0.24 | APAC |  Singapore |
|  Frogos | 0.24 | APAC |  Australia |
|  Sachin V S | 0.08 | Americas |  Canada |
|  Xiaoyue Yang | 0.07 | APAC |  China |

Figure 6.3: Frog Counting Challenge - Leader Board Scores

| Model | MAE |
|--------------------------------------|-------------|
| Model-W (ANN) | 9.35 |
| Proposed Fusion Model (LC & Numeric) | 10.19 |

Table 6.2: Comparison of Model-W and Proposed model on Test data

6.1.3 Feature Importance Assessment

The results obtained from the feature importance assessment experiment suggest that temperature plays a key role in contributing to the prediction of target variable, followed by evapo-transpiration and precipitation. The results of model trained only with the top 6 features suggest that it simplifies the model by reducing its dimensionality, making it more efficient. By comparing the MAE of both the models, it is clear that only 6 features is more than sufficient to make predictions without sacrificing predictive accuracy. When

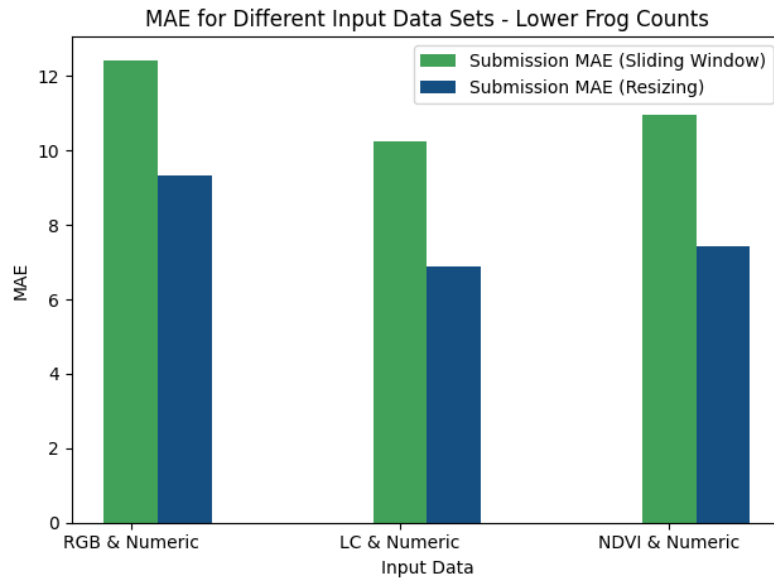


Figure 6.4: Results - MAE obtained for Lower value of Frog count

| Input Data | Submission MAE | |
|----------------|----------------|-------------|
| | Sliding Window | Resizing |
| RGB & Numeric | 12.43 | 9.33 |
| LC & Numeric | 10.25 | 6.90 |
| NDVI & Numeric | 10.97 | 7.43 |

Table 6.3: MAE obtained on Submission data for lower values of frog count

| Dataset | Evaluation Metric | Value | |
|-----------------|-------------------|-------------------------|------------------|
| | | Fusion Model (Proposed) | Model-w (Winner) |
| Submission Data | F1 Score | 0.36 | 0.42 |
| Test Data | MAE | 10.19 | 9.35 |

Table 6.4: Two Comparison Approaches - Overview

looking at the **scalability** of the model, it is important to notice that assessing the feature importance can be beneficial for larger datasets and complex models.

The following key aspects can be analysed in building a SDM for frogs by the results of feature importance assessment.

Variable Selection: Assessing the important features leads the selection of the relevant features to include in the SDM. Temperature and precipitation are identified as the

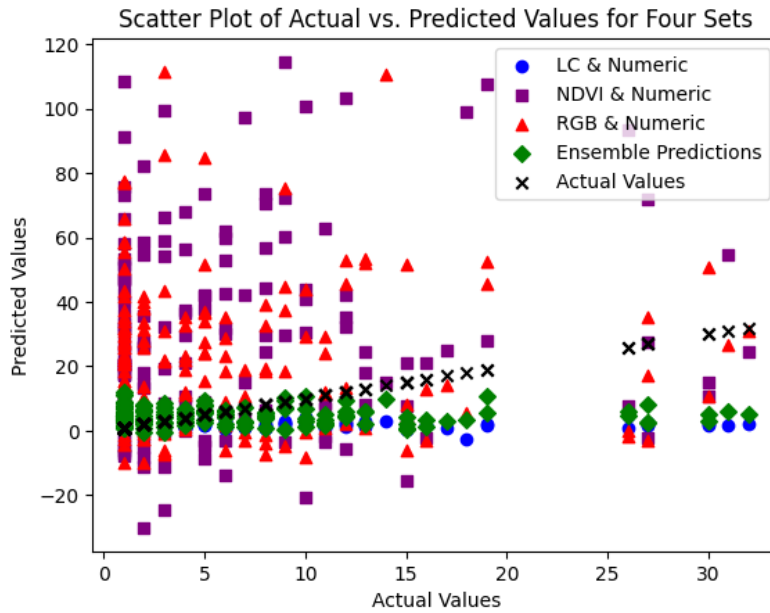


Figure 6.5: Scatter Plot on submission data for lower value of frog count

driving factors in the case of SDMs for frogs.

Biological Insights: The important features obtained provides ecologists and conservationists with valuable insights in understanding the species habitat requirements. In the case of frogs, as mentioned earlier in section 4.1.1 under terraclimate, temperature [28] and precipitation [43] plays an important role in distribution of frogs. These findings correlated with the results obtained through the experiment conducted here.

Climate Change: When it comes to climate change, feature importance can help assess the vulnerability of species to changing environmental conditions. In the case of frogs, by identifying that temperature as the factor that influences the distribution the most, planning and conservation of frogs can be adapted according to the future climate scenarios.

6.1.4 Generalization

The results obtained by the model trained on Australian dataset, in predicting the frog count in Costa Rica suggest that the model generalizes well. The model was able to achieve similar MAE comparable to that of the model trained on all three countries. Australia and Costa Rica are different when it comes to geographical and climate context. To show how different the two countries are different, observe figure 6.6, 6.7 and 6.8. Since, temperature and precipitation are the significant factors that influenced frog count, they are chosen to evaluate how different the two countries are.

From figure 6.6, it can be observed that the maximum and minimum temperature of Australia can go upto 36°celcius and 5°celcius respectively. While, for Costa Rica it can only go upto 32°celcius and 18°celcius.

From figure 6.7, the accumulated precipitation for Australia and Costa Rica is 95mm and 550mm respectively. Both these climatic features shows how varied is the climate between the two countries.

Figure 6.8, shows how the land cover differs between the two countries.

Inspite of such variations between the two countries, the proposed model was able to

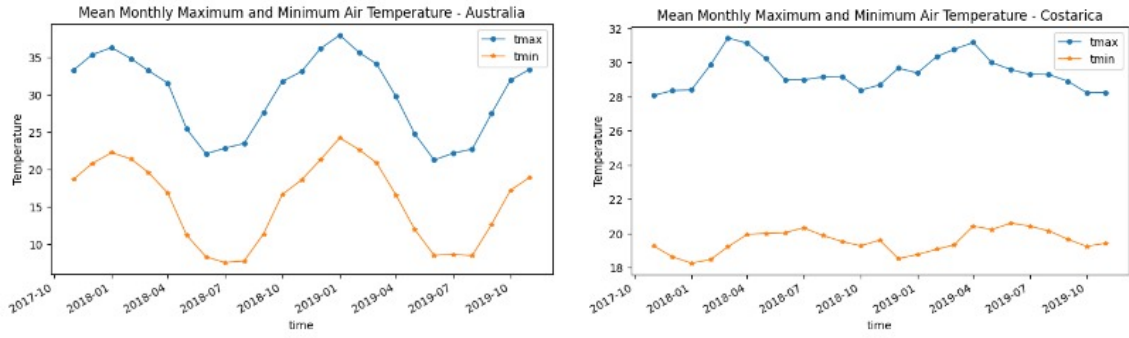


Figure 6.6: Tmax and Tmin of Australia and Costa Rica

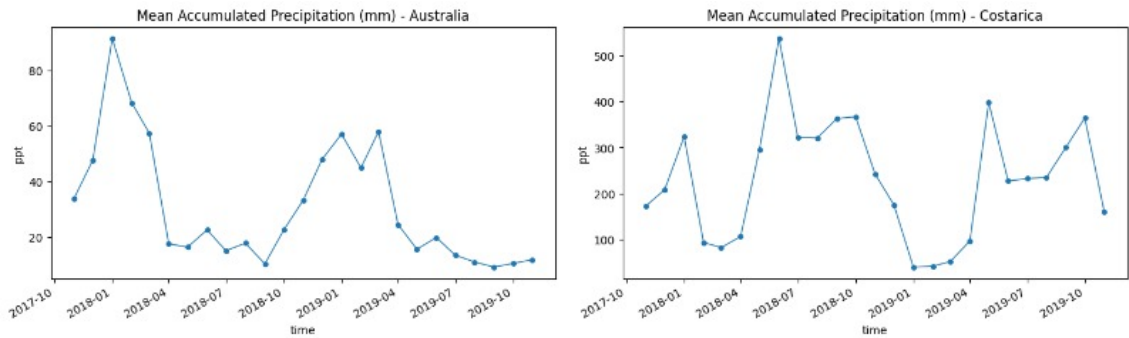


Figure 6.7: Mean Precipitation of Australia and Costa Rica

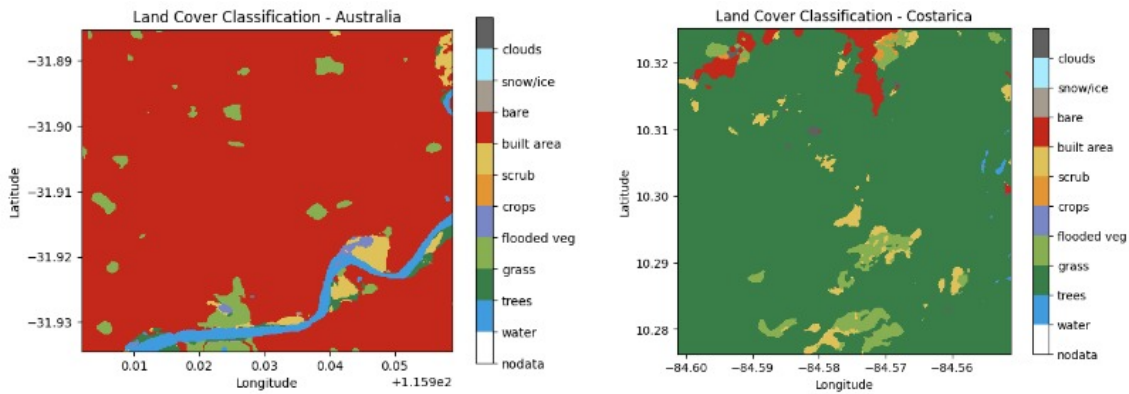


Figure 6.8: Sample Land Cover patches of Australia and Costa Rica

generalize well and resulted in low MAE.

6.1.5 Performance Evaluation using XGBoost

In addition to the comparison of the XGBoost performance with that of the winner, this particular experiment also evaluates the difference between using single modality and multiple modalities of data. using two modalities clearly performs better than the model using single modality. Though the performance is only slightly better, this already shows the potential of using multiple modalities.

| Model | Evaluation Dataset | MAE |
|--------------------------|--------------------------|--------------|
| XGBoost | Submission data | 27.16 |
| Fusion (Ensemble Method) | Submission data | 21.94 |
| XGBoost | Test Data | 10.81 |
| Fusion | Test Data (LC & Numeric) | 10.19 |

Table 6.5: Comparison of XGBoost and Fusion model

6.1.6 Presence / Absence Classification

The results from the presence / absence classification task suggest that LC & Numeric data play a crucial role in identifying the locations with frog presence or absence. The important reason for performing this experiment however is to analyse how effective the pseudo-absence data generated is, in predicting the frog occurrence. Eventhough there is no direct model for comparison, the model was able to achieve good accuracy in dataset that was kept solely for the purpose of evaluation. To put things into perspective, observe the figure 6.9, the presence and absence point are located so close to each other. And figure 6.10, 6.11 shows that the mean temperature and accumulated precipitation of the two locations have not much difference. This shows that the model is able to capture the intricate details between the presence and absence points and is able to classify the location as presence / Absence. From the table 3.1, we can observe the classification accuracy and AUC score of the existing state-of-the-art methods. Though these models cannot be directly compared with the proposed model, the achieved accuracy and AUC score are on par with those models. These comparison will be more credible if the dataset used in those literature are available or the ground truth data for frog absence points are present.

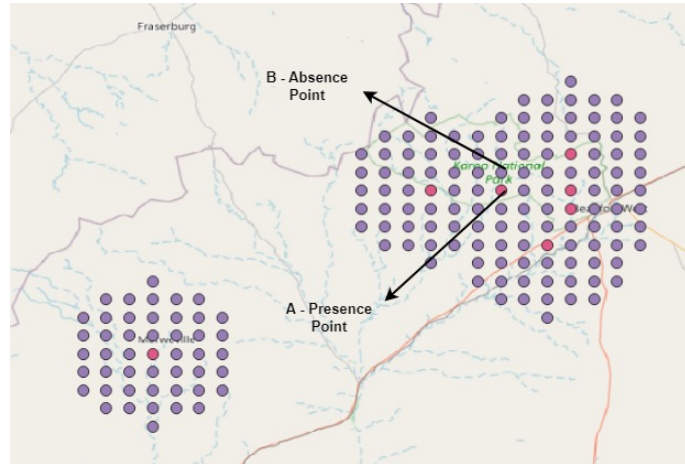


Figure 6.9: Sample Presence and Absence Points of South Africa

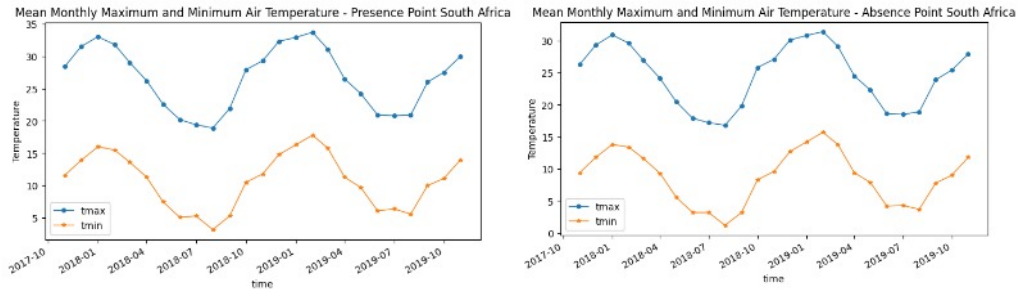


Figure 6.10: Tmax and Tmin of presence and absence point - South Africa

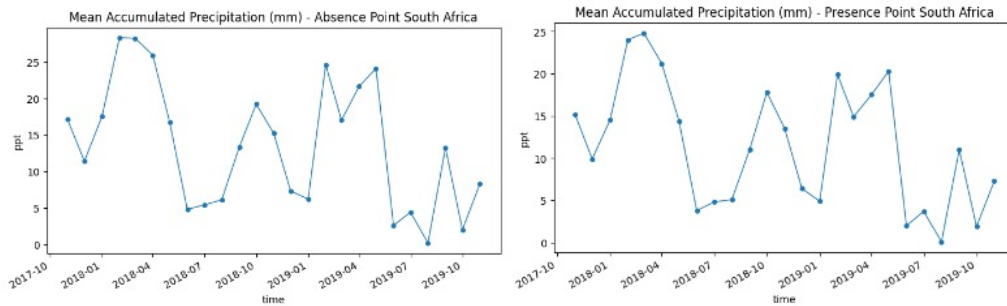


Figure 6.11: Accumulated Precipitation of presence and absence point - South Africa

6.1.7 Comparison of different pseudo-absence generation methods

From the results obtained in the experiment conducted, clearly more the variables are taken into account for generating the pseudo-absence points, better the predictions are. But, by just including more criteria is not enough as some variables may result in generating incorrect pseudo-absence points. For example, from the feature importance assessment, it can be observed that temperature plays an important role in the presence of frogs. So, including it may help in predicting the absence point better. But taking less important features like soil moisture, wind speed and Pdsi etc. can corrupt the data.

6.2 Addressing the Research Questions

RQ1: What are the major limitations of the existing methods for building an SDM in predicting the presence of a species in a particular location?

From the literature review conducted earlier, there were several limitations with the existing SDMs, which resulted in looking beyond them for building SDMs. With the conventional statistical methods, one of the disadvantage is that, MaxEnt approach used presence-only datasets. This could result in the predictions being partial towards only one class. The logistic regression model posed the limitation of being based on correlative data, which could lead to failure in obtaining the actual relationship between the predictor variable and target variable. And also due to the linear nature of the function, complex relationship existing between the covariates and species cannot be represented by them. In the case of CNN based deep learning models, valuable insights about the climatic data are being missed out due to the input data being single modality. Eventhough, CNN-SDMs produced good results, it is only logical to make use of all the available modalities of data to get better result.

RQ2: How multimodal learning can be made use of in building an SDM?

This research question can also be answered by the conducted literature survey. SDMs are characterized by heterogeneous nature of input data. It is important to analyse the spatial pattern of the location. In order to do so, high resolution satellite images are used as inputs. But, climatic variables like temperature and humidity take up the form of tabular data. So, SDMs benefit from having both types of data. This is why multimodal learning can be utilized in building SDMs. Collecting data from both the sources is an important part in building a multimodal based SDM. Combining both types of data can be achieved by fusion architecture. This combines the features extracted from different modalities and represent it together. This approach allows for the representation of information from both modalities in a unified manner.

RQ3:How does the performance of the proposed SDM based on multimodal learning compare to the existing state-of-the-art methods?

The first task is to count the frogs present at a location. This is a regression task, where most of the existing state of the art methods focus on classification. So, the proposed method is compared with the winner (model-w) of the EY open science data challenge. On the test data Model-w obtained a MAE of 9.35. Compared to it, the proposed model obtained a MAE of 10.19 . When it comes to the submission data, the winner of the challenge achieved a F1 score of 0.42, while the proposed model resulted in a score of 0.36, which is placed second in the leader board.

Comparing with model using unimodal data, XGBoost model achieved a MAE of 27.16 on submission data, while the proposed fusion model resulted in a lower MAE of 21.94.

The second task is to identify the presence / absence of frogs at a location. The best accuracy obtained was 89.19% and an AUC score of 0.96 which is on par with that of the existing state-of-the-art methods.

RQ4: How do different covariates contribute to the prediction of the target variable?

From the feature importance assessment, temperature and precipitation are identified as the most important covariate that contributes more to the prediction of target variable. These two covariates make up almost 85% of the contribution among terraclimate variables (tabular data). When it comes to high resolution satellite images, land cover and NDVI patches produced better MAE compared to RGB images. As explained earlier this is due to the semantic segmentation nature of the land cover and NDVI images. However, not all the types of data are used for predicting the frog count, due to time constraints. But, it should produce almost similar performance as compared to the ones used in this work.

RQ5: How is the performance of the model using pseudo-absence points generated by the proposed method compared to the existing ones?

From the experiment conducted to analyse the performance difference between the proposed method of generating pseudo-absence points and a couple of existing methods, the proposed method performed better than the other methods. Random selection method performed worse, because no factor was taken into account while selecting the pseudo-absence point other than the fact that it is not a presence point. On the other hand, the points selected by the distance criteria method performed close to the proposed method.

This is because, there might be some common points between the two methods, as the threshold distance chosen are the same. The model evaluated by both the metrics accuracy (84.87%) and AUC score (0.90) suggest that the proposed method is better at generating pseudo-absence points.

6.3 Limitations

Model Complexity

Adding multiple modalities of data increases the complexity of the fusion model. For example, table 6.6 shows the time it took for training the model for one epoch using two and three modalities of data. While it took under 10 minutes for one epoch for models trained using only two modalities of data, adding an additional data type increased the time almost 10 times. So, ways of efficiently using different modalities of data should be looked into and a mere concatenation will not result in an efficient model.

Data Collection

Though we have input data of multiple modalities available for training, while using the model on unseen locations, obtaining data of multiple modalities of such remote location can be a difficult task. And the reliability of such data is always a question to ponder about.

| Input Data | Train time - 1 epoch (minutes) |
|---------------------|--------------------------------|
| RGB & Numeric | 7 |
| LC & Numeric | 7 |
| NDVI & Numeric | 8.5 |
| RGB, LC and Numeric | 78 |

Table 6.6: Training Time for models using different modalities of data

Prediction ability of higher count of frogs

From the obtained results, one can observe that the model does not perform well on predicting higher frog count. This can be connected to the dataset. Eventhough, balancing of dataset resulted in good performance, due to the lack of data points present for higher counts the learning capacity of the model is limited. Features could not be learned well for higher counts of frog compared to lower counts.

Chapter 7

Conclusion and Future Work

In this chapter, the conclusion of the research is provided and also discusses the future research works.

7.1 Conclusion

The objective of the research is to build a SDM for *anura* using input data from multiple modalities and compare the results with existing method that uses a single modality of data. Though there are already several methods available in building a SDM, it comes with several disadvantages as pointed out in the related works section. And with the growing interest and the effectiveness of multimodal learning, this work could well be a starting point for the less explored area of using multimodal learning to build a SDM. The designed SDM is used for the purpose of both counting frogs and also classifying the area as presence/absence.

By building a fusion architecture that combines images and tabular data, the model is able to achieve performance comparable to that of the winner of the challenge. Though the achieved MAE is slightly less than the model-w, with better representation of the data at higher value of frog count the model can perform much better than the existing methods.

The comparison between XGBoost and fusion model suggest that using multi-modal data performs better than using uni-modal data.

The aim of the presence/absence classification is to evaluate the reliability of the generated pseudo-absence data and observe how effective the model is in learning the distinction between the presence and absence points. The model was able to classify the location as presence/absence with almost 89% accuracy. Though there is no separate data available to evaluate the model, it still was able to capture and make distinction between the presence and absent location on the evaluation data (separated from training data).

Moreover, comparing the proposed method of generating the pseudo-absence data with couple of existing methods, the proposed method performed better and achieved an accuracy of 84.87% and an AUC score of 0.90.

7.2 Future Work

The results obtained are promising and has great scope in using multimodal learning for SDMs. This section describes some of the future works that can be performed to further

increase the prediction accuracy.

Stages of Fusion: The current proposed architecture performs fusion after extracting features from different modalities separately. The features extracted are then passed through ResNet50 and FCNN for images and tabular data respectively. This constitutes the primary learning of the model. This is called **late fusion**.

In the case of **early fusion** [64], the pre-processed data from each different modalities are fused before passing it to a learning algorithm. The image feature vectors and environmental feature vectors are combined at the early stage. This type of fusion is suitable especially when there data from multiple modalities are associated strongly. Which is the case for environmental variables. This is a method worthy of exploring in the future. Apart from this, the kind of fusion can also be explored. This work proposes concatenation of extracted features. But, feature addition is also a technique that can be explored.

Pseudo-Absence data Generation: The method used for generating pseudo-absence data can be studied further to include more details. For example, the current method takes into account only the geographical extent and land cover patches for selecting the pseudo-absence points. Parameters like temperature and precipitation can also be taken into account since they proved to be the most important factor in influencing the frog habitat. Moreover, algorithms like K-means clustering can be used to group points similar to presence points so that the dissimilar points can be chosen as pseudo-absence points.

Including Historical data for better representation: While the provided frog presence dataset includes only the year 2017-2019, covariates of only those period is taken into account for building the SDM. However, including the climatic data of previous years can help in generating additional data. This will result in a better representation of frog counts of higher values. This method can be explored in the future to see if it could result in better predictions.

SDM based on GNN: Similar to the GNN approach explained in section 3.3 in weather forecasting can be applied in building a SDM. Already the dataset has been prepared using the method explained in section 4.4. But, experiments could not be conducted within the time available. This can be a future work which can yield a better result.

Deployment on Edge AI devices: Edge AI is a technology that has gained attention in recent years. It involves deploying deep learning models on low powered micro-controllers. Deploying a multimodal based SDM on a low powered device is a challenging task and is exciting to explore further. **Quantization of neural networks** [29] and **neural network pruning** [63] are some of the techniques that are used to make the neural networks efficient in terms of memory, power consumption and speed.

Bibliography

- [1] Effects of sample size and network depth on a deep learning approach to species distribution modeling | Elsevier Enhanced Reader. URL: <https://reader.elsevier.com/reader/sd/pii/S157495412030087X?token=30A364BBC67F4E0E3D8E5D3AB311424DA159798AAC0C2A91072E9D8ED50F0188F2A49370558C951297A682C> originRegion=eu-west-1&originCreation=20230519215754, doi:10.1016/j.ecoinf.2020.101137.
- [2] An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity / Simon Ferrier and Graham Watson. URL: <https://collection.sl.nsw.gov.au/record/74VvPKwrwj3>.
- [3] Home. URL: <https://www.frogid.net.au/>.
- [4] iNaturalist Research-grade Observations. URL: <https://www.gbif.org/dataset/50c9509d-22c7-4a22-a47d-8c48425ef4a7>, doi:10.15468/ab3s5x.
- [5] Microsoft Planetary Computer. URL: <https://planetarycomputer.microsoft.com/>.
- [6] EY Wavespace Madrid & CT AI. Open Science Data Challenge. URL: <https://challenge.ey.com/challenges/level-3-frog-counting-tool/data-description>.
- [7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. doi:10.1109/TPAMI.2018.2798607.
- [8] Jaya Basnet, Abeer Alsadoon, P. W. C. Prasad, Sarmad Al Aloussi, and Omar Hisham Alsadoon. A Novel Solution of Using Deep Learning for White Blood Cells Classification: Enhanced Loss Function with Regularization and Weighted Loss (EL-FRWL). *Neural Processing Letters*, 52(2):1517–1553, October 2020. doi:10.1007/s11063-020-10321-9.
- [9] Monica Bianchini and Franco Scarselli. On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, August 2014. Conference Name: IEEE Transactions on Neural Networks and Learning Systems. doi:10.1109/TNNLS.2013.2293637.
- [10] Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A Deep Learning Approach to Species Distribution Modelling. In Alexis Joly, Stefanos

- Vrochidis, Kostas Karatzas, Ari Karppinen, and Pierre Bonnet, editors, *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 169–199. Springer International Publishing, Cham, 2018. URL: http://link.springer.com/10.1007/978-3-319-76445-0_10, doi:10.1007/978-3-319-76445-0_10.
- [11] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. URL: <https://www.sciencedirect.com/science/article/pii/S0031320396001422>, doi:10.1016/S0031-3203(96)00142-2.
- [12] John R. Busby. A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology*, 11(1):1–7, 1986. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1442-9993.1986.tb00912.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.1986.tb00912.x>, doi:10.1111/j.1442-9993.1986.tb00912.x.
- [13] Guy Carpenter, Andrew N Gillison, and J Winter. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals.
- [14] D. Richard Cutler, Thomas C. Edwards, Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. Random Forests for Classification in Ecology. *Ecology*, 88(11):2783–2792, 2007. Publisher: Ecological Society of America. URL: <https://www.jstor.org/stable/27651436>.
- [15] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz, and Alexis Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLOS Computational Biology*, 17(4):e1008856, April 2021. Publisher: Public Library of Science. URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008856>, doi:10.1371/journal.pcbi.1008856.
- [16] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, François Munoz, and Alexis Joly. Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge.
- [17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. doi:10.1109/CVPR.2009.5206848.
- [18] Enric Domingo. GeoLifeCLEF 2022 Winning Submission. URL: <https://www.kaggle.com/competitions/geolifeclef-2022-lifeclef-2022-fgvc9/discussion/327055>.
- [19] John M. Drake, Christophe Randin, and Antoine Guisan. Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, 43(3):424–432, 2006. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2664.2006.01141.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2664.2006.01141.x>, doi:10.1111/j.1365-2664.2006.01141.x.
- [20] S. Dupont and J. Luetttin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, September 2000. Conference Name: IEEE Transactions on Multimedia. doi:10.1109/6046.865479.

- [21] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2656.2008.01390.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2008.01390.x>, doi:10.1111/j.1365-2656.2008.01390.x.
- [22] Jane Elith and Janet Franklin. Species Distribution Modeling. January 2017. doi: 10.1016/B978-0-12-809633-8.02390-6.
- [23] K. Ruwani M. Fernando and Chris P. Tsokos. Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, July 2022. Conference Name: IEEE Transactions on Neural Networks and Learning Systems. doi:10.1109/TNNLS.2020.3047335.
- [24] Scott D. Foster and Piers K. Dunstan. The Analysis of Biodiversity Using Rank Abundance Distributions. *Biometrics*, 66(1):186–195, 2010. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2009.01263.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2009.01263.x>, doi:10.1111/j.1541-0420.2009.01263.x.
- [25] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/>.
- [26] Bo-cai Gao. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, December 1996. URL: <https://www.sciencedirect.com/science/article/pii/S0034425796000673>, doi:10.1016/S0034-4257(96)00067-3.
- [27] Abhishek Garg and Rajshekhar Hippargi. Significance of frogs and toads in environmental conservation. February 2007.
- [28] Alyssa A. Gerick, Robin G. Munshaw, Wendy J. Palen, Stacey A. Combes, and Sacha M. O’Regan. Thermal physiology and species distribution models reveal climate vulnerability of temperate amphibians. *Journal of Biogeography*, 41(4):713–723, 2014. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbi.12261>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jbi.12261>, doi:10.1111/jbi.12261.
- [29] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A Survey of Quantization Methods for Efficient Neural Network Inference, June 2021. arXiv:2103.13630 [cs]. URL: <http://arxiv.org/abs/2103.13630>.
- [30] Xu Han, Ming Jia, Yachao Chang, Yaopeng Li, and Shaohua Wu. Directed message passing neural network (D-MPNN) with graph edge attention (GEA) for property prediction of biofuel-relevant species. *Energy and AI*, 10:100201, November 2022. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2666546822000477>, doi:10.1016/j.egyai.2022.100201.
- [31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. arXiv:1512.03385 [cs]. URL: <http://arxiv.org/abs/1512.03385>, doi:10.48550/arXiv.1512.03385.

- [32] Shiwen He, Shaowen Xiong, Yeyu Ou, Jian Zhang, Jiaheng Wang, Yongming Huang, and Yaoxue Zhang. An Overview on the Application of Graph Neural Networks in Wireless Networks. *IEEE Open Journal of the Communications Society*, 2:2547–2565, 2021. Conference Name: IEEE Open Journal of the Communications Society. doi:10.1109/OJCOMS.2021.3128637.
- [33] Jeff Heaton. An Empirical Analysis of Feature Engineering for Predictive Modeling. In *SoutheastCon 2016*, pages 1–6, March 2016. arXiv:1701.07852 [cs]. URL: <http://arxiv.org/abs/1701.07852>, doi:10.1109/SECON.2016.7506650.
- [34] Sha Huang, Lina Tang, Joseph P. Hupy, Yang Wang, and Guofan Shao. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, February 2021. doi:10.1007/s11676-020-01155-1.
- [35] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions.
- [36] Ryan Keisler. Forecasting Global Weather with Graph Neural Networks, February 2022. arXiv:2202.07575 [physics]. URL: <http://arxiv.org/abs/2202.07575>, doi:10.48550/arXiv.2202.07575.
- [37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
- [38] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. GraphCast: Learning skillful medium-range global weather forecasting, August 2023. arXiv:2212.12794 [physics]. URL: <http://arxiv.org/abs/2212.12794>.
- [39] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989. Conference Name: Neural Computation. doi:10.1162/neco.1989.1.4.541.
- [40] Sovan Lek, Alain Belaud, Philippe Baran, Ioannis Dimopoulos, and Marc Delacoste. Role of some environmental variables in trout abundance models using neural networks. *Aquatic Living Resources*, 9(1):23–29, January 1996. URL: <http://www.alr-journal.org/10.1051/alr:1996004>, doi:10.1051/alr:1996004.
- [41] Sovan Lek, Marc Delacoste, Philippe Baran, Ioannis Dimopoulos, Jacques Lauga, and Stéphane Aulagnier. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, 90(1):39–52, September 1996. URL: <https://linkinghub.elsevier.com/retrieve/pii/0304380095001425>, doi:10.1016/0304-3800(95)00142-5.
- [42] Sovan Lek and J. F. Guégan. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120(2):65–73,

August 1999. URL: <https://www.sciencedirect.com/science/article/pii/S0304380099000927>, doi:10.1016/S0304-3800(99)00092-7.

- [43] Gavia F. Lertzman-Lepofsky, Amanda M. Kissel, Barry Sinervo, and Wendy J. Palen. Water loss and temperature interact to compound amphibian vulnerability to climate change. *Global Change Biology*, 26(9):4868–4879, 2020. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.15231>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15231>, doi:10.1111/gcb.15231.
- [44] Yiming Li, Jeremy M. Cohen, and Jason R. Rohr. Review and synthesis of the effects of climate change on amphibians. *Integrative Zoology*, 8(2):145–161, 2013. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1749-4877.12001>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1749-4877.12001>, doi:10.1111/1749-4877.12001.
- [45] Canran Liu, Graeme Newell, and Matt White. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, 42(3):535–548, 2019. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.03188>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03188>, doi:10.1111/ecog.03188.
- [46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, October 2021. IEEE. URL: <https://ieeexplore.ieee.org/document/9710580/>, doi:10.1109/ICCV48922.2021.00986.
- [47] Stéphanie Manel, Jean-Marie Dias, and Steve J. Ormerod. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120(2):337–347, August 1999. URL: <https://www.sciencedirect.com/science/article/pii/S0304380099001131>, doi:10.1016/S0304-3800(99)00113-1.
- [48] Jennifer Miller. Species Distribution Modeling. *Geography Compass*, 4(6):490–509, 2010. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-8198.2010.00351.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-8198.2010.00351.x>, doi:10.1111/j.1749-8198.2010.00351.x.
- [49] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal Deep Learning.
- [50] Mangesh Nichat. Landmark based shortest path detection by using A* Algorithm and Haversine Formula. April 2013.
- [51] Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, 2017. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12757>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12757>, doi:10.1111/ele.12757.

- [52] Trishala K. Parmar, Deepak Rawtani, and Y. K. Agrawal. Bioindicators: the natural indicator of environmental pollution. *Frontiers in Life Science*, 9(2):110–118, April 2016. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/21553769.2016.1162753>. doi:10.1080/21553769.2016.1162753.
- [53] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training Recurrent Neural Networks, February 2013. arXiv:1211.5063 [cs]. URL: <http://arxiv.org/abs/1211.5063>.
- [54] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259, January 2006. URL: <https://www.sciencedirect.com/science/article/pii/S030438000500267X>, doi:10.1016/j.ecolmodel.2005.03.026.
- [55] Steven J. Phillips, Miroslav Dudík, and Robert E. Schapire. A maximum entropy approach to species distribution modeling. In *Twenty-first international conference on Machine learning - ICML '04*, page 83, Banff, Alberta, Canada, 2004. ACM Press. URL: <http://portal.acm.org/citation.cfm?doid=1015330.1015412>, doi:10.1145/1015330.1015412.
- [56] P Pradhyumna, G P Shreya, and Mohana. Graph Neural Network (GNN) in Image and Video Understanding Using Deep Learning for Computer Vision Applications. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1183–1189, August 2021. doi:10.1109/ICESC51422.2021.9532631.
- [57] Tara Rawat and Vineeta Khemchandani. Feature Engineering (FE) Tools and Techniques for Better Classification Performance. May 2019. doi:10.21172/ijiet.82.024.
- [58] Azusa Sawada, Eiji Kaneko, and Kazutoshi Sagi. Trade-offs in Top-k Classification Accuracies on Losses for Deep Learning, July 2020. arXiv:2007.15359 [cs, stat]. URL: <http://arxiv.org/abs/2007.15359>.
- [59] Senait D. Senay, Susan P. Worner, and Takayoshi Ikeda. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE*, 8(8):e71218, August 2013. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3742778/>, doi:10.1371/journal.pone.0071218.
- [60] Sachith Seneviratne. Contrastive Representation Learning for Natural World Imagery: Habitat prediction for 30,000 species. September 2021.
- [61] Sagar Sharma. <https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>.
- [62] Hailu Shiferaw, Woldeamlak Bewket, and Sandra Eckert. Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. *Ecology and Evolution*, 9(5):2562–2574, 2019. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.4919>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4919>, doi:10.1002/ece3.4919.

- [63] Sietsma and Dow. Neural net pruning-why and how. In *IEEE 1988 International Conference on Neural Networks*, pages 325–333 vol.1, July 1988. doi:10.1109/ICNN.1988.23864.
- [64] William C. Sleeman, Rishabh Kapoor, and Preetam Ghosh. Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *ACM Computing Surveys*, 55(7):150:1–150:31, December 2022. URL: <https://dl.acm.org/doi/10.1145/3543848>, doi:10.1145/3543848.
- [65] Imelda Somodi, Nikolett Lepesi, and Zoltán Botta-Dukát. Prevalence dependence in model goodness measures with special emphasis on true skill statistics. *Ecology and Evolution*, 7(3):863–872, January 2017. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5288248/>, doi:10.1002/ece3.2654.
- [66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA, June 2016. IEEE. URL: <http://ieeexplore.ieee.org/document/7780677/>, doi:10.1109/CVPR.2016.308.
- [67] Tina Tirelli and Daniela Pessani. Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus* in piedmont (North-Western Italy). *River Research and Applications*, 25(8):1001–1012, 2009. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/rra.1199>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/rra.1199>, doi:10.1002/rra.1199.
- [68] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, Boston, MA, USA, June 2015. IEEE. URL: <http://ieeexplore.ieee.org/document/7298935/>, doi:10.1109/CVPR.2015.7298935.
- [69] Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00749.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00749.x>, doi:10.1111/j.1467-9868.2010.00749.x.
- [70] Richard L. Wyman. Soil Acidity and Moisture and the Distribution of Amphibians in Five Forests of Southcentral New York. *Copeia*, 1988(2):394–399, 1988. Publisher: [American Society of Ichthyologists and Herpetologists (ASIH), Allen Press]. URL: <https://www.jstor.org/stable/1445879>, doi:10.2307/1445879.
- [71] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, NV, USA, June 2016. IEEE. URL: <http://ieeexplore.ieee.org/document/7780940/>, doi:10.1109/CVPR.2016.571.
- [72] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *2018 IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 1316–1324, Salt Lake City, UT, USA, June 2018. IEEE. URL: <https://ieeexplore.ieee.org/document/8578241/>, doi:10.1109/CVPR.2018.00143.
- [73] Wenchao Xu, Yuxin Pang, Yanqin Yang, and Yanbo Liu. Human Activity Recognition Based On Convolutional Neural Network. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 165–170, August 2018. ISSN: 1051-4651. doi:10.1109/ICPR.2018.8545435.
- [74] Xinchun Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2Image: Conditional Image Generation from Visual Attributes, October 2016. arXiv:1512.00570 [cs]. URL: <http://arxiv.org/abs/1512.00570>.
- [75] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, March 2020. Conference Name: IEEE Journal of Selected Topics in Signal Processing. doi:10.1109/JSTSP.2020.2987728.
- [76] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics*, 12, 2021. URL: <https://www.frontiersin.org/articles/10.3389/fgene.2021.690049>.
- [77] Xiaojuan Zhang, Yongxiu Zhou, Peihao Peng, and Guoyan Wang. A Novel Multimodal Species Distribution Model Fusing Remote Sensing Images and Environmental Features. *Sustainability*, 14(21):14034, January 2022. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute. URL: <https://www.mdpi.com/2071-1050/14/21/14034>, doi:10.3390/su142114034.
- [78] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, January 2020. URL: <https://www.sciencedirect.com/science/article/pii/S2666651021000012>, doi:10.1016/j.aiopen.2021.01.001.
- [79] Stacy L Özdesmi and Uygur Özdesmi. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, 116(1):15–31, March 1999. URL: <https://www.sciencedirect.com/science/article/pii/S0304380098001495>, doi:10.1016/S0304-3800(98)00149-5.

Appendix A

Appendix

A.1 Cross Validation Results

Cross validation for the task of frog counting is performed using k-fold technique. The dataset is divided into 5 parts. In one complete cycle, 4 parts are used as training data and the remaining is used as test data. This one set is run for the required number of epochs. Similarly, the dataset is shuffled, where the previously assigned test set will be included as one of the training set, while one of the training set will be kept as the test set. So, in the end, five different MAE values each corresponding to five different sets will be obtained. The mean of those five values will constitute the validation MAE. However, training and evaluation on only one set could be performed within the time available. The results are shown in A.1

| Set | Epochs | MAE |
|-----|--------|-------|
| 1 | 77 * | 43.26 |
| 2 | - | - |
| 3 | - | - |
| 4 | - | - |
| 5 | - | - |

Table A.1: Cross Validation Results. * Not Completed

A.2 Balancing Approach - K means Clustering

In the dataset balancing step mentioned in section 4.2.3, k-means clustering is used to group similar data points together. An illustration of the feature space is given in figure A.1

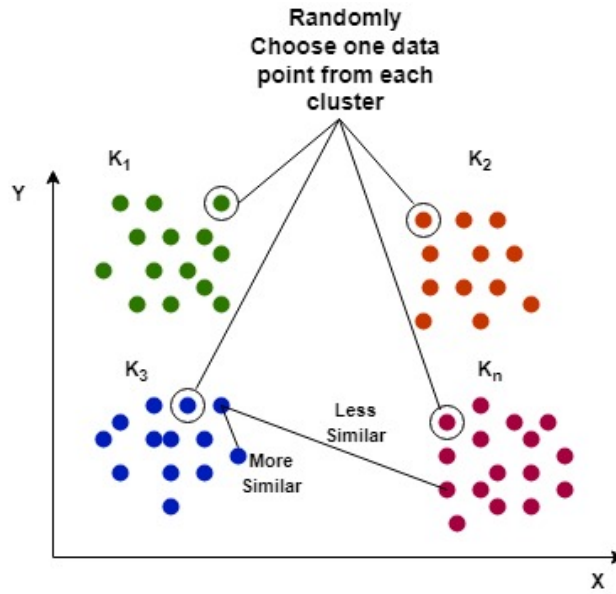


Figure A.1: K-means Clustering - Feature Space

A.3 Terraclimate Variables - Correlation matrix

To know how the parameters in the terraclimate variables interact with each other, a correlation matrix is plotted, refer figure A.2. The correlation matrix gives a value between -1 to 1. -1 means the two variables are inversely related and a value of 1 means the two variables have linear relationship.

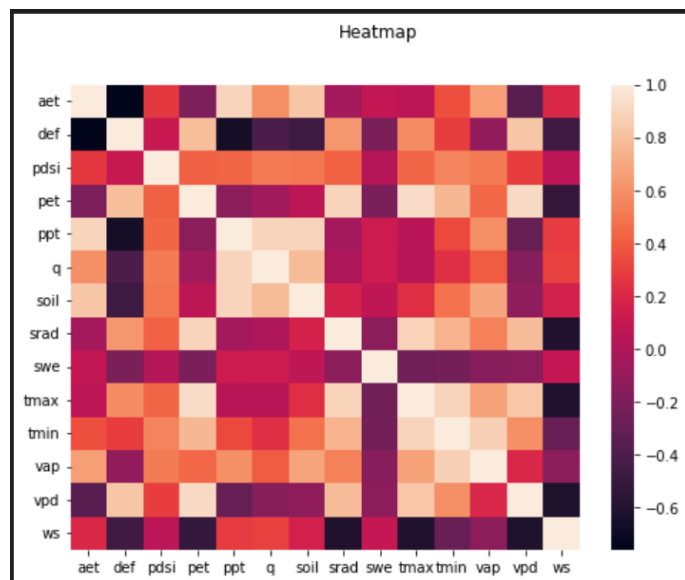


Figure A.2: Heatmap - Correlation matrix of Terraclimate variables