SPECIES DISTRIBUTION MODEL TO PREDICT THE OCCURRENCE OF MALAYAN PARTRIDGE

BY

Darren Leong Chien Hsiung

A REPORT
SUBMITTED TO
Universiti Tunku Abdul Rahman
in partial fulfillment of the requirements
for the degree of
BACHELOR OF COMPUTER SCIENCE (HONOURS)
Faculty of Information and Communication Technology
(Kampar Campus)

June 2025

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Ts Dr Vikneswary a/p Jayapal who has given me this bright opportunity to engage in a project related to Species Distribution Model. It is my first step to establish a career in ecological modelling and biodiversity conservation field. A million thanks to you.

Finally, I must say thanks to my parents and my family for their love, support, and continuous encouragement throughout the course.

ABSTRACT

Climate change has caused several problems in Malaysia such as increase of temperature and change in precipitation patterns. Malayan Partridge (Arborophila campbelli) is a bird species found in Peninsular Malaysia that is facing the threat of habitat loss due to climate changes. Currently, this species is understudied and that leads to less information about the future occurrence of this species. Therefore, this study aims to produce a prediction of current and future occurrence of Malayan Partridge in Peninsular Malaysia with different models Species Distribution Model (SDM) that are Maximum Entropy Model (MaxEnt), Random Forest (RF), Support Vector Machine (SVM), Generalized Linear Model (GLM) and Bioclim. Different pseudoabsence data settings will be implemented to identify the best setting to predict the occurrence of the species. Species occurrence data were collected from public biodiversity databases 19 bioclimatic variables were sourced from WorldClim to predict the current occurrence of the species. A variable selection process will be used to identify the important bioclimatic variables. These variables will be used for the models of SDM. To predict the potential future occurrence of the species, Shared Socioeconomic Pathway (SSP) will be implemented. The performance of the model will be evaluated through Area Under Curve (AUC) and cross-validation techniques. Habitat suitability maps will be produced because of the model to provide visualization.

Area of Study (Minimum 1 and Maximum 2): Machine Learning

Keywords (Minimum 5 and Maximum 10): Species Distribution Model, Habitat Suitability Map, Climate Change, Shared Socioeconomic Pathway, Malayan Partridge

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SDM	Species Distribution Model
RF	Random Forest
GAM	Generalized Addictive Model
MaxEnt	Maximum Entropy Model
GLM	Generalized Linear Model
ANN	Artificial Neural Network
SVM	Support Vector Machine
DNN	Deep Neural Network
CLIP	Contrastive Language-Image Pre-Training
SSP	Shared Socioeconomic Pathway
RCP	Representative Concentration Pathway
AUC	Area Under Curve
ML	Machine Learning
DL	Deep Learning
SHAP	Shapley Additive Explanations

Chapter 1
Project Background

In this chapter, it will introduce the background of the work including project objectives and scope, motivation and contributions.

1.1 Introduction

Climate change is one of the most significant challenges in the world that affects biodiversity and distribution of species. The ecosystems are changing at a high pace due to the rising global temperatures, changing precipitation patterns, and unpredictable weather. Climate changes can lead to extinction of species, changes in distribution and interactions [1]. Therefore, it is important to understand the climatic environmental variables to capture the current and future distribution trends of species.

The Malayan Partridge, also known as Arborophila campbelli, is a bird species in the family of Phasianidae which includes grouse, turkeys, pheasants and partridges. It is a ground-dwelling bird in the montane forest of Peninsular Malaysia. It is a species that thrives and lives in cooler and humid environments, and it is dependent on stable climate condition. However, due to the change of climate, they are losing their habitat due to the shifting altitudinal ranges [2]. As the temperature rises, Malayan Partridge needs to move up to higher locations to seek for better living environments to live in. Due to the limits of mountain ranges, they will have nowhere to go in the end, causing them to become extinct in the future.

To prevent these threats, Species Distribution Model (SDM) is a great choice as it can use algorithms to predict a species occurrence and distribution based on occurrence and environmental data [3]. Choosing the right environmental variables is important to generate a reliable result in SDM. 19 bioclimatic variables are being used as inputs for the model of SDM. However, too many variables may lead to redundancy and overfitting as there might be correlation between each bioclimate variable. Therefore, algorithm such Pearson Correlation, Random Forest (RF), XGBoost and Permutation Importance can be implemented to filter out the important variables first before building the initial model of SDM.

Therefore, the need of understanding and predicting the bioclimatic variables that will affect the occurrence of Malayan Partridge in Peninsular Malaysia is important to conserve this species. It is important to develop a SDM to predict the occurrence of Malayan Partridge to understand the correlation between the bioclimate variables and the occurrence of Malayan Partridge.

1.2 Problem Statement

The Malayan Partridge is a montane forest bird that can be found in highland forest such as Bukit Fraser and Cameron Highlands in Peninsular Malaysia. This species relies heavily on having a dense, untouchable forest environment to survive and thrive. However, the distribution of Malayan Partridge has been seriously threatened by habitat fragmentation, deforestation and climate change due to human activities. The species is under more stress due to the changes of landscape cause my human activities, that may lead to the major drop of its population in Malaysia.

The traditional method is manually recording using camera traps to capture the occurrence of the species is less effective. This method is labor-intensive, time-consuming as the montane forest is a difficult location for humans to access. Also, the manual method may lead to inaccuracies and difficulties in monitoring the occurrence of the species. This is because the camera might not be able to capture every angle in a montane forest that might cause some occurrence to be mis looked at. Other than that, this method may require significant time and effort to set up, making it difficult to set up a large-scale monitoring system. In addition, there is a need for human intervention to record each occurrence species that appear in specific areas. Overall, these manual methods of monitoring Malayan Partridge in Malaysia will be time consuming and error prone.

Currently, there are limited research on the species current and future occurrence patterns despite its importance to conserve this species in Malaysia. It is important to understand how bioclimatic variables and factors can influence the occurrence of this species before planning an effective and useful conservation plan. However, there is less information and knowledge about this area. The lack of predictive system will cause humans to be a step slower in conservation measures and leaving the species in a vulnerable condition.

1.3 Project Objectives

The study must achieve the following objectives.

- To identify the bioclimatic variables that affect the occurrence of Malayan Partridge.
    - This objective is to identify which bioclimatic variables are related to the known habitat locations of Malayan Partridge
    - Having this information can be useful while selecting which bioclimatic variables to implement in SDM.
- To develop Species Distribution Model (SDM) to predict the current and future habitat suitability of Malayan Partridge.
    - Development of SDM that uses different SDM models
    - The model can implement the future data Shared Socioeconomic Pathways (SSP) of bioclimatic variables for future projection for the occurrence of the species.
- To evaluate the accuracy and reliability of the developed models
    - The performance of the developed model will be determined by using performance metrics to ensure the model provides a reliable result.

1.4 Motivation

The Malayan Partridge is a species that is vulnerable to climate changes as it primarily appears and lives in montane forest at higher elevations. Therefore, this species is highly sensitive to any changes to its environment. As the global climate changes. Malayan Partridge might not be able to adapt to the new conditions and lead to extinction. Even though it is important to understand how climate change affect this species, there are currently few research that focus on Malayan Partridge in Malaysia. SDM has proven that it is a useful tool to predict the occurrence of species based on various environmental variables. Therefore, the project is motivated by the needs of understanding how bioclimatic variables can affect the occurrence of Malayan Partridge. Having this information can provide better understanding of the impact on the species and develop conservation measures to protect the species.

1.5 Project Scope

The project is to develop a Species Distribution model using multiple algorithms methods to predict the occurrence of the Malayan Partridge in Malaysia. The aim of the model is to provide a tool that will provide insights on how climate change will affect the current and future occurrence of the species. The scope of the environmental variables will be focused on bioclimatic variables and factors such as land use and vegetation cover will not be included.

This will be achieved by implementing occurrence data of Malayan Partridge and bioclimatic variables as input data in an SDM. The bioclimatic variables will be selected and filtered using correlation filter, Random Forest (RF), XGBoost and Permutation Importance algorithm to identify the important variables for the models SDM. Then different settings of pseudoabsence data will be implemented to identify the best and optimal model settings. After deciding on the best model setting, the results of different models SDM will be compared using the AUC and predicted distribution map for Malayan Partridge. The chosen best models will be evaluated more by predicting the future occurrence of Malayan Partridge using bioclimatic variables with SSP.

The result of the project will identify the best models that can be used to predict the occurrence of Malayan Partridge. These models will be producing the distribution map of the current and future occurrence of the species. This can be used for visualization for any conservation use.

1.6 Contributions

In this project, Species Distribution Model will be developed using different algorithms to predict the current and future occurrence of the Malayan Partridge. Comparing different algorithms can provide a better insight on which algorithms are suitable for predicting the occurrence of the species. The project will also explore the importance of using different pseudoabsence settings for SDM. Having the information and knowledge of future occurrences for the species, it will be able to provide insights into the species' habitats suitability in the future having considered the impact of climate changes. It also provides information for planning conservation measures to protect this species in Malaysia. Other than that, there are limited studies on the impact of climate change on Malayan Partridge in Malaysia. Therefore, this project can help to fill some knowledge gaps about this species.

1.7 Report Organization

The report is structured into 6 chapters. Chapter 1, Introduction provides an overview for the project such as problem statement, project scope and objectives, motivation and contribution. Chapter 2 is the literature review part, provides information about existing research and methodologies related to Species Distribution Model. Chapter 3 shows the system model of the process of developing Species Distribution Model. In chapter 4, it describes the implementation work that has been completed to develop the model. Chapter 5 provides the interpretation of the results produced by SDM. Lastly, Chapter 6 concludes the project findings and future work that can be done.

Chapter 2
Literature Review

2.1 Introduction to Species Distribution Model

Species Distribution Models (SDMs) are essential and important tools used in various fields such as evolution, ecology, conservation and epidemiology to make critical decisions and study biological phenomena [4]. SDM is valuable for cases where experimental approaches are not feasible. By correlating environmental variables with known occurrences of species, SDMs can provide predictive insights of potential distribution species over different geographical areas. Sometimes, biotic information can be included as a predictor variable in SDMs [5]. The presence of a competitor or predator can limit the distribution of a species. For example, the distribution of deer will be affected by the presence of wolves. However, mutualistic relationships, such as a relationship between clownfish and sea anemone can increase the distribution of the species. Clownfish require the protection of anemone from predators, and clownfish can provide nutrients for sea anemone [6]. These variables can be included to the models to improve their accuracy by capturing significant ecological interactions and environmental factors that affect the distribution of species.

SDMs usually can be divided into two main groups that are correlative, and process-based or mechanistic models to predict the distribution of species [7]. Correlative models are designed to make predict species distribution by describing relationships between geographical occurrence and environmental data without explicit consideration of ecological process [8]. Therefore, correlative models will rely on statistical techniques to find patterns and relationship between species occurrence and environmental variables. Models that are under the category of correlative models are heavily dependent on the quality and quantity of data as they will perform the best when there is high quality data available. Poor quality data or incomplete environmental datasets can lead to biased or misleading predictions. Examples of popular correlative models are Maximum Entropy Modelling (MaxEnt), Generalized Linear Models (GLM) and Generalized Addictive Models (GAM).

However, process-based models are significantly different from correlative models. Unlike correlative models, process-based models predict the species distribution based on mathematical functions of ecological processes [7]. These models can determine whether a species can survive based on the ecological processes identified such as physiological tolerances, growth rates and reproduction. Therefore, process-based models predict what area are suitable for the species to survive but do not explain why a species is found in some location and no other location. Therefore, it makes the model become less useful for understanding the relationship between species and environmental variable changes as they don't capture the full complexity of species-environment relationship. However, process-based models are more complex and require more detailed knowledge of an ecological process. Examples of process-based models are Dynamic Energy Budget (DEB) Models, Physiologically Structured Population Models (PSP) and NicheMapper.

Correlative models are used more commonly as it considers the impact of biotic interactions and dispersal constraints on distribution [9]. By having observed occurrence data, correlative models can capture the realized niche of species and make it effective for predicting distribution within a species known range. However, the use of correlative model and process-based model majorly depends on the task [10]. Correlative models can be helpful to estimate distributions based on observed data, but it may struggle to predict distribution in novel or changing environments accurately as the biotic interactions could be different from the observed data. In contrast, process-based models are more suitable for predicting distribution in novel or changing environments, but it may not perform well within a species native range.

2.2 Types of SDM Algorithms

There are several popular algorithms that can be implemented for predicting potential distribution of species using SDM. One of the popular algorithms of SDM is using Maximum Entropy Algorithm (MaxEnt). MaxEnt is widely used in identifying potential distribution for endangered species [11], invasive species management [12], and species conservation planning [13]. Based on the principle of maximum entropy, the algorithm generates prediction by estimating a species' probability distribution throughout a landscape, with the constraints that each environmental variable's predicted value must match its empirical mean. There are several advantages and disadvantages that can be discovered through past research that use this algorithm. One of the advantages of MaxEnt algorithm is that it is one of reliable bioclimatic modelling approaches for presence-only data. The authors in [14] shows that Maxent outperform several other models that are also presence-only methods in terms of predictive accuracy and robustness. However, there are some disadvantages to using a MaxEnt algorithm. MaxEnt algorithm is sensitive to biases in the presence data. [15] have mentioned that MaxEnt is sensitive to sampling bias in the input data that may lead to overfitting. Therefore, MaxEnt may overfit the model and produce inaccurate predictions if the input data is not reflected in the actual environment conditions across the study area.

Another SDM algorithm that is used commonly is Generalized Linear Models (GLM). Based on [16], there are several types of GLM models such as linear regression, logistic regression and Poisson regression. Each type of GLM model can be used in different situations. When working with continuous outcome variables that show a linear connection with predictor variables, linear regression can be implemented. For example, if the amount of sunlight increase (predictor variable) then the height of the plant will be affected by it (continuous outcome variable). In contrast, logistic regression will be suitable for binary outcomes, where the result will only be 2 categories such as yes or no. For Poisson Regression, it will be suitable for counting data, which means that the outcome is the number of times an event occurred in a period. In overall of GLM models, it is a flexible model as it can handle different types of data that makes it versatile for ecological data [17]. However, it would be difficult for GLM models to make predictions if the relationship between variables and predictors is a non-linear relationship.

Generalized Additive Models (GAM) is an extension of GLMs that uses non-linear relationships between response and predictor variables using smoothing functions. The smooth functions will be computed separately for each variable, and it will be added together to form the final model [17]. By doing so it allows GAMs to represent complex relationships more effectively than a single linear term. GAMs are useful when the relationship between species and environmental variables is more complex and not easily fitted by GLMs [18]. One of the advantages of GAMs is it can be model complex and non-linear relationship unlike GLMs. GAMs allow for non-linear relationships by implementing smooth functions such as splines for each predictor. This enables the model to capture patterns in the data that might be unnoticed by linear models [16]. Besides, it can improve the predictive performance compared to linear models. GAMs can offer better predictive performance, especially when the underlying data relationships are complex. GAMs face a risk of overfitting when there are too many knots or overly flexible spline. Therefore, penalization methods such as ridge or lasso are often needed to prevent overfitting [16]. Also, it is important to choose the correct amount of smoothing as incorrect choices may lead to either underfit or overfit of the data.

Random Forest is a classifier that consists of many decision trees that implements Breiman's random forest algorithm for classification and regression [19]. This algorithm will create a collection of decision trees where each tree is trained on a random subset of training data. Each tree will be trained using a technique known as bootstrapping. It will be used to make sure each tree is trained on random subsets of data. This technique can improve model variety while reducing overfitting. Random Forest is one of the most dependable learning algorithms that performs well in predicting species distribution [20]. This is because, due to the averaging of multiple trees and the random feature selection, it is less prone to overfitting compared to individual decision tree. Additionally, random forest can make predictions even when some data are lacking using surrogate split. However, it is difficult to interpret the overall model due to the complexity and large number of decision trees. Besides, training many trees may be time-consuming and expensive as it requires many memories to operate on.

Artificial neural networks (ANN) are also a SDM algorithm that can be used to predict distribution of species. Bionic neural networks served as an inspiration for ANN computation models. It can represent complex and non-linear interactions and is made from interconnected nodes layered one over the other. ANN contains a hidden layer where each node in the hidden layer receives data from each input, sums the input, adds a constant and then convert the result into a fixed function [16]. The quantity of weight decay in the links and the quantity of hidden neurons will determine how accurate

ANNs are. ANNs are very flexible that is enough to approximate any smooth function [16]. It can easily adapt to various types of data and relationships, making it suitable for a diverse SDM scenarios. However, the amount of data needed to make accurate predictions is large and the training can be complex and intensive.

2.2.1 Comparison Table of SDM Algorithms

SDM Algorithm	Advantage	Disadvantages
Maximum Entropy (MaxEnt)	Effective for presence-only data, means that it is useful when occurrence data is available	Sensitive to biases in presence data, which may lead to overfitting
Generalized Linear Models (GLM)	Flexible as it can handle different types of data (continuous, binary, count)	Struggle with non-linear relationship between variables and predictors
Generalized Additive Models (GAM)	Handles non-linear relationships through smoothing functions	Risk of overfitting with too many knots
Random Forest	Ability to reduce overfitting through the averaging of multiple decision trees and random feature selection	Difficult to interpret due to its complex and large model
Artificial Neural Networks (ANN)	Flexible and capable of modelling complex and non-linear interactions	Require large amounts of data to produce accurate predictions

Table 2.2.1.1 Comparison Table of SDM algorithms

2.3 Species Distribution Models with Deep Learning

Deep learning is one of the very powerful tools in Species Distribution Model (SDM) as it provides features to capture complex relationships between the species occurrence and environmental data. Previously, deep learning is commonly applied in other areas such as finance and business. The trend of using deep learning for the field of ecology has just started. As compared to other sectors, datasets can be easily generated to produce large datasets that will be needed for deep learning compared to ecological data. However, with the advancement of technology, there are efforts carried out to create large ecological datasets [21]. When a large dataset is being implemented for SDM, the performance of Deep Learning has the potential to outperform other Machine Learning approaches [22].

Deep Neural Network (DNN) is one of the most common usages of deep learning for SDM. This is due to the ability of DNN to combine different types of data such as image-like pixel data, numeric and categorical variables [22]. This allows them to have the ability to adapt to the multilayers of the ecological environment. One of the examples of usage of DNN in SDM is to predict the bark beetle outbreaks [22]. In the results of the model, it shows higher accuracy compared to using machine learning models to predict the same thing.

Even though deep learning usually requires many data sets, there are ways that have been developed to improve the accuracy of zero-sample species recognition. For example, zero-shot is a great example that allows models to recognize species that they have never been trained on with related occurrence data. This is particularly useful when there is a need to predict a species that are rare. One example approach proposed to implement CLIP-driven zero-shot species recognition method to recognize species [23]. Normally, CLIP is being used for general images and text pairing tasks [24]. However, the paper proposed to use CLIP together with zero-shot to enable models to classify a species based on its textual characteristics without having any visual examples.

Despite having a large advantage of using deep learning, there are currently limited studies of deep learning SDM. It might face challenges in ecological implementation such as using it to predict a species habitat suitability.

2.4 Environmental variables that affect bird distribution

The distribution of birds including Malayan Partridge is influenced by multiple different environmental variables that determine the suitability of their habitat to thrive in. One of the most important environmental variables that affect the distribution of birds is bioclimate variables. Birds are often restricted to certain temperatures and precipitation ranges that cause them to be more sensitive to climate changes. According to [25], temperature and precipitation patterns were 2 of the most significant predictors of bird diversity. The study shows that if an area has more variable climates, it will lead to a lower bird diversity. As climate changes, birds are forced to shift to higher elevations to search for a more suitable living environment. However, for birds like Malayan Partridge, it will be difficult for them to move upwards as they are already located at a higher altitude environment.

Other than that, vegetation also plays a big role in providing habitats and food resources for birds and different species of birds may require different vegetation type and structures as their habitat. Studies show that vegetation is the key factor of bird distribution and habitat suitability [25] [26]. Even though there are other environmental factors that also contribute to the distribution of species, vegetation is still the most influential factor. [25] highlighted that birds prefer forested landscapes instead of human-modified areas as forested landscapes have a better structural complexity and diverse food resources compared to human-modified areas. This shows that deforestation and modification of habitats will lead to the loss of bird's distribution.

Lastly, human activities such as deforestation and urbanization can lead to the loss of habitats of birds. Research in [27] found out that bird's species is affected by human disturbance and activities such as reducing the forest cover for construction purposes. [26] also noted that development near the forest negatively impacted the distribution of birds due to the loss of habitats and increased human activity and disturbance.

While there are broad options for environmental variables, this study focuses on bioclimatic variables only. This is because there is high-resolution bioclimatic data available compared to other variables. Bioclimatic variables are always impactful to predict the occurrence of the species as climate can directly affect the species' survival. Although other factors may be useful, they are not included in this model of the project.

2.5 Applications of SDMs in predicting bird distribution

SDM has been successfully used in predicting the distribution of birds across different habitats and areas. It can offer insights into suitability, ranges of shifts and environmental threats. For Malayan Partridge, SDM can be useful in mapping its current distribution, understanding the environmental variables that affect the species and guiding conservation actions.

In terms of SDM implementation for birds, there are several research and studies used SDM to identify the current and potential area where the species is likely to occur in. For example, the author [28] implemented MaxEnt modelling to identify the distribution of common pheasant in Iran, proving that the distance to agricultural lands and forest were the key factors for the suitability of habitat for the common pheasant. The study also shows that habitat fragmentation and the changes of land caused the pheasant to move into higher altitudes that may cause their survival long term as it might not be their ideal living environment. Other than that, another study [29] found that montane birds in Malaysia were highly sensitive to land-use changes. The study showed the importance of conversing the montane regions in Malaysia to ensure the suitability of habitat for montane birds such as Malayan Partridge to survive in the future.

Other than that, several studies have used SDMs to demonstrate how climate change can threaten the bird’s species particularly for those in a higher altitude environment. A study [30] implemented SDM to predict the future distributions of threaten birds on the Qinghai-Tibet Plateau in China, and finding out that almost 80% of species would lose habitats under warming and temperature increase conditions. The study particularly highlighted Galliformes as the highest risk due to their limited dispersal abilities. For example, most Galliformes have weak flight abilities that majorly affect the ability to sustain long-distance travel. Similarly, another study [31] used MaxEnt modelling with RCP4.5 and RCP8.5 climate projections to predict the future distribution of 3 different Galliformes species in Pakistan, and result shows that all three species will lose over 25% of their habitat by 2050 and shift to higher elevations, to find for cooler living environments. However, mountains have limited heights, meaning that once the species reach the maximum height, they will run out of suitable habitats.

SDMs also play a huge role in conservation planning and planning for protected areas for different species. SDMs can make predictions for habitat suitability for each species, helping in making decisions to establish protected areas or implementing restoration projects. For example, research by [32] has conducted a diversity assessment in China using SDM combined with Zonation algorithms to identify conservation priority area for Galliformes species to 3 future climate change scenarios. The study shows that over 80% of the priority conservation areas that are suitable for Galliformes species are not formally protected, showing that there is mismatch between the protected areas and actual habitat for the species.

SDM is a valuable tool to predict the distribution of birds, but there are several limitations and disadvantages. Firstly, data limitations are a major issue when using SDM to predict distribution of birds. This is because SDM algorithms rely on occurrences data of a specific species. For example, in the research by [28], many Galliformes species that have limited, or few occurrence records have caused bias in the data predictions. In many cases, occurrences data are incomplete due to the lack of survey and research in the specific area of the species. As a result, it will lead to bias in SDM [33], only reflecting the species distribution in areas that are well-studied and researched.

Also, model uncertainty exists when using SDMs to predict the distribution of a species. This is because different SDMs such as MaxEnt, GAMs or Random Forest will produce different predictions due to their differences in their algorithms and data handling. However, this can be solved by doing ensemble modelling [32]. Ensemble modelling combines multiple SDMs to improve predictive ability and reduce uncertainty by using averaging predictions and model weighting. The study highlighted that ensemble approaches provide a more robust prediction by mitigating the bias of each SDMs.

As Malayan Partridge shares similar living conditions as montane birds and pheasant, applying SDMs is an effective way to predict its priority habitats and make conservation decisions in those areas based on previous research. The research aligns with the trends that can be seen by Malayan Partridge, showing that climate change may affect its habitat and distribution. SDMs can predict how the distribution will shift with different climate scenarios. Due to the rising temperature in hill stations such as Cameron Highland and Fraser Hill, SDMs can help to identify which area will remain suitable for Malayan Partridge after climate changes. For Malayan Partridge, SDMs can help to determine whether existing protected areas are covering their habitat and whether any new conservation zones need to be established. By doing so, it can ensure that there are no resources waste on non-suitable habitat areas, instead it can focus on expanding protected zones and improve habitat environment on the suitable habitat areas.

2.6 Limitation of Previous Works

Even though Species Distribution Model has been a common method used for predicting bird species and conservation measures, there are no research can find to study the species of Malayan Partridge in Malaysia. Therefore, there is less understanding of the species relationship with different bioclimate variables. Without this information, it makes it difficult to prepare any conservative measures to protect the species in the future. Other than that, most study [28][30][31] to develop SDM for bird's species using only MaxEnt model since it is the most common method for SDM. Developing only 1 model to predict a species may limit the prediction accuracy and lack of comparison with other models. Different SDM algorithms such as RF, GLM and GAM can produce different predictions and results as they have different ways of handling data and relationship between variables. By evaluating multiple models, it allows a more thorough and complete analysis of understanding the prediction of species distribution. Therefore, this project will address these limitations by analyzing the bioclimatic variables that affect the species of Malayan Partridge and implementing multiple SDM algorithms for comparison. This approach can provide a more complete result to identify the current and future distribution of Malayan Partridge.

Chapter 3
System Model

3.1 Research Methodology

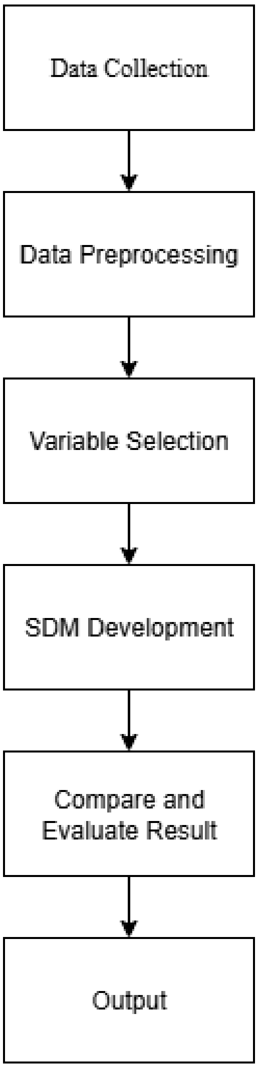


Figure 3.1.1 Overview of Research Development Workflow

3.1.1 Data Collection:

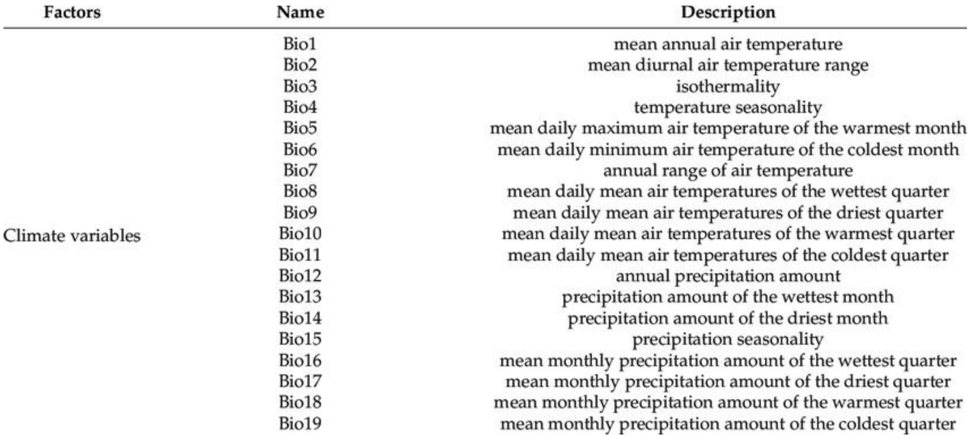
This is the first step in the process of developing a SDM, collecting necessary data needed for the input of the model. For this project, there are 3 types of data that are required to ensure that the model works.

1. Occurrence data:

The occurrence data needed in this project is the records of Malayan Partridge presence in Malaysia. This data is acquired from the website of eBird [34] that contains all the records captured for Malayan Partridge. Having accurate data of the species is important as it is one of the inputs for SDM.

2. Bioclimatic variables data:

   These are the environmental variables that are needed for SDM. There is a total of 19 bioclimatic variables available starting from bio1 to bio19. The data of bioclimatic variables are being sourced from WorldClim [35] as they provide high-resolution global climate datasets that are suitable as an input for SDM. The data will be using spatial resolutions of 30 seconds (~1km^2).



Factors	Name	Description
Climate variables	Bio1	mean annual air temperature
	Bio2	mean diurnal air temperature range
	Bio3	isothermality
	Bio4	temperature seasonality
	Bio5	mean daily maximum air temperature of the warmest month
	Bio6	mean daily minimum air temperature of the coldest month
	Bio7	annual range of air temperature
	Bio8	mean daily mean air temperatures of the wettest quarter
	Bio9	mean daily mean air temperatures of the driest quarter
	Bio10	mean daily mean air temperatures of the warmest quarter
	Bio11	mean daily mean air temperatures of the coldest quarter
	Bio12	annual precipitation amount
	Bio13	precipitation amount of the wettest month
	Bio14	precipitation amount of the driest month
	Bio15	precipitation seasonality
	Bio16	mean monthly precipitation amount of the wettest quarter
	Bio17	mean monthly precipitation amount of the driest quarter
	Bio18	mean monthly precipitation amount of the warmest quarter
	Bio19	mean monthly precipitation amount of the coldest quarter

Figure 3.1.1.1 Description of Bio1 to Bio19 [36]

3. Future bioclimatic variables data:

   This data is like normal bioclimatic variables, but it is the data that represents the future climate scenarios based on Shared Socioeconomic Pathway (SSP) of different levels. SSP is a framework used in climate change that shows the possible futures climate based on socioeconomic development, population, economy and technology. These factors will affect the greenhouse gas emission that led to climate changes. Table 3.1.1.2 shows the different levels of SSP. For example, SSP 1 refers to the future socioeconomic, the higher the worse. 2.6 refers to Representative Concentration Pathway (RCP), the greenhouse gas trajectory, the higher the value, the higher the emission of gas. The level of SSP will be used is SSP2-45 and will be using future climate data for the year 2021 to 2040 and 2041 to 2060. Having this data allows the model to predict future potential suitable habitats for the occurrence of Malayan Partridge based on future climate projections.

SSP - RCP	**Interpretation**
SSP 1 -2.6	Low emission and warming
SSP 2 – 4.5	Medium emission and warming
SSP 3 – 7.0	High emission and warming
SSP 5 – 8.5	Very High emission and warming

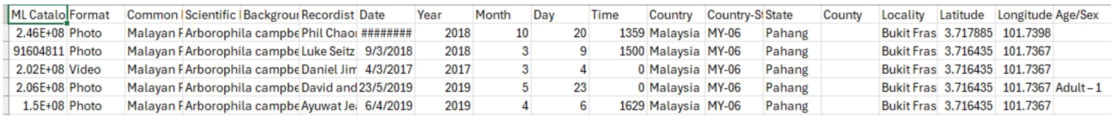
Table 3.1.1.1 Different Values of Shared Socioeconomic Pathways [37]

3.1.2 Data Preprocessing:

After the required data has been obtained, it needs to be processed before it can be the input for the model. Other than that, there is data that needs to be generated for the models of SDM. There will be four main steps that need to be taken.

1. Cleaning occurrence data:

   This process involves cleaning the occurrence data of Malayan Partridge. This involves removing duplicate data and removing any rows with empty data. Also, there are many different columns that consist of data that are not required for the input of SDM such as date and time. Figure 3.1.2.1 shows some of the columns of the original dataset for Malayan Partridge. Most of the columns will be removed except for species name, longitude and latitude. This will be able to avoid any confusion for the model.



ML Catalo	Format	Common I	Scientific I	Backgrour	Recordist	Date	Year	Month	Day	Time	Country	Country-S	State	County	Locality	Latitude	Longitude	Age/Sex
2.46E+08	Photo	Malayan F	Arborophila campbe		Phil Chaoı	########	2018	10	20	1359	Malaysia	MY-06	Pahang		Bukit Fras	3.717885	101.7398	
91604811	Photo	Malayan F	Arborophila campbe		Luke Seitz	9/3/2018	2018	3	9	1500	Malaysia	MY-06	Pahang		Bukit Fras	3.716435	101.7367	
2.02E+08	Video	Malayan F	Arborophila campbe		Daniel Jir	4/3/2017	2017	3	4	0	Malaysia	MY-06	Pahang		Bukit Fras	3.716435	101.7367	
2.06E+08	Photo	Malayan F	Arborophila campbe		David and	23/5/2019	2019	5	23	0	Malaysia	MY-06	Pahang		Bukit Fras	3.716435	101.7367	Adult – 1
1.5E+08	Photo	Malayan F	Arborophila campbe		Ayuwat Je	6/4/2019	2019	4	6	1629	Malaysia	MY-06	Pahang		Bukit Fras	3.716435	101.7367	

Figure 3.1.2.1 Columns of the Original Occurrence Dataset

2. Prepare environmental variables:

   The bioclimatic variables data that have been obtained includes data for the whole world. To speed up the processing time of the model, the data needs to be trimmed and align with the proposed study area that is Malaysia. Other than that, the future bioclimatic variable data might include multiple bands together, therefore it must be extracted to become individual band of bioclimatic variables data.

3. Data transformation:

   The bioclimatic variables data comes originally as a GeoTIFF (.tif) file from WorldClim. However, ASCII raster format (.asc) file is more suitable to implement in the model of SDM. Therefore, the data format needs to be transformed to smoothen the prediction process.

4. Generating pseudoabsence data:

   Since there are no true absence data points for Malayan Partridge can be found, pseudoabsence data are generated to be implement in the models for SDM. Currently, there are no specific studies that show that is the best pseudoabsence settings for SDM. However, the number of pseudoabsence data can affect the performance of the models [38]. Therefore, multiple pseudoabsence settings will be generated to be implemented in the models to identify which settings is the most suitable to predict the occurrence of Malayan Partridge. A 10km buffer around the presence points of the species will be implemented when generating pseudoabsence points. By doing so it can reduce the risk of putting pseudoabsence points at the locations that might be suitable for the species.

3.1.3 Variable Selection:

In this step, the goal is to identify which bioclimatic variables will be used for the development of SDM. This is because the processing time of the model will be long if the environmental variables are too much. Correlation filters will be done to identify and remove the correlation variables. Then, the remaining variables will be implemented in 3 different methods to rank the importance of each variable towards Malayan Partridge. The variables that have at least 2 votes from the methods will be used as the final variables for SDM.

1. Pearson Correlation:

   Pearson Correlation is a feature selection method that focuses on removing variables that are highly correlated with each other. Pearson correlation coefficient, r is used to measure how much the two variables can represent each other [39]. The higher the value of r, the higher the correlation between two of the variables. The threshold implemented for this variable selection method is |r|> 0.8, if this condition is met, the variable will be removed from the list of bioclimatic variables that will be used for Species Distribution Model. Formula of r is given to calculate the value for correlation analysis. Value r is being implemented to find out the collinearity between bioclimatic variables.

Formula for Pearson Correlation in equation 3.1:

$$|r| = \frac{n\,\Sigma xy \;-\; (\Sigma x)(\Sigma y)}{\sqrt{(n\,\Sigma x^2 \;-\; (\Sigma x)^2)(n\,\Sigma y^2 \;-\; (\Sigma y)^2)}} \tag{3.1}$$

where:

x = values of first variable,

y = values of second variable

n = number of paired variables observed

2. Random Forest:

   Random Forest is an ensemble of decisions trees built on random subsets of the data and predictors. Each tree will provide votes for a prediction. It will rank predictors by their contribution across all trees. It will take the majority vote that is provided in the tree. Random Forest is a great option to perform variable selection as it can capture complex interactions between features very well. This is important as the interaction between species and bioclimate are often complex and non-linear.

3. XGBoost:

   XGBoost, Extreme Gradient Boosting is like Random Forest as both are tree-based methods. It differs from Random Forest as it will build model sequentially, where each new model focuses on errors of the previous models. The sequential process allows the model to improve its overall performance. XGBoost will be able to provide an important score for each bioclimate variable on how frequently it was used in the decision trees. The importance of a score will help to identify which variables are more important.

4. Permutation Importance:

   Permutation importance is a model-agnostic method that measures how much the model performance is affected when the value of the variable is being shuffled randomly. If the variable is important, shuffling the values of the variable will cause the performance of the model to drop significantly. Therefore, this method is particularly suitable for variable selection as it can identify whether the selected bioclimate variables are truly important and contribute to the performance of the model.

5. Voting Process:

   A voting process will be conducted with a threshold of more than 2. If 2 or more models select a variable, it will be used as part of the final Species Distribution Model.

3.1.4 SDM Development:

In this project, it will implement five different algorithms to develop Species Distribution Model to predict the occurrence of Malayan Partridge. Having five different algorithms allows for comparison and identifying which algorithms are the most suitable to predict the occurrence of Malayan Partridge. These models will use the final variables identified from the variable selection process. The models will apply different pseudoabsence settings generated from the data preprocessing part. Below are the selected algorithms:

1. Maximum Entropy (MaxEnt):

   MaxEnt is a presence-only model that is commonly used for SDM research [11][12]. This is because it does not require absence data of the species that are difficult to obtain in the real world. This method will estimate the probability of distribution of a species that is closest to uniform with the environmental conditions of the presence points for the species. This means that it will start with an assumption that the species will occur everywhere but then takes account of the known occurrence and the environmental variables to identify which area is suitable.

2. Random Forest (RF):

   RF is an ensemble method that is built from multiple decision trees [16]. It will use a voting process to identify which variable is suitable for the species and make predictions for presence or absence. It is also commonly used for SDM as it takes account of complex and non-linear relationships.

3. Support Vector Machine (SVM):

   SVM is a classifier method that finds the best boundary that separates the presence and absence of the species. It will use high dimensions to make separation for predictions. SVM also considers complex and non-linear relationships well that is suitable for SDM [40].

4. Generalized Linear Model (GLM):

   GLM is a model that assumes linear relationship between environmental variables and the species [19]. However, it can use logistic function to help capture the non-linear relationship while making predictions for a species distribution.

5. Bioclim:

   Bioclim is an envelope method that focuses on presence only data [41]. It will identify the minimum and maximum values of the bioclimate variables for the known areas for the species. Then it will use these values to predict whether the area is suitable for a species to occur.

These models are chosen because they can provide a balance between machine learning models and traditional statistical approaches. While machine learning is the main study, additional methods can provide more insight into how these models perform in Species Distribution Model. GLM is a statistical model that assumes linear relationships while Bioclim is an envelope model that works by defining the species range of environmental variables. These methods can provide a clearer interpretation and baselines for prediction. On the other hand, MaxEnt, RF and SVM are more modern machine learning methods that can capture complex and non-linear relationships very well. MaxEnt is slightly different compared to RF and SVM as it uses presence only data for prediction while RF and SVM require presence and absence data to provide a reliable result for SDM.

3.1.5 Compare and Evaluate Result

The five models will be developed with different pseudoabsence settings. To evaluate the performance of each model, it will be using train AUC and test AUC. Train AUC shows how well the model fits on the trained data, while test AUC shows how well the model can generalize unseen and new data. The purpose of using train AUC and test AUC is to identify whether the model has any overfitting signs. If a model train AUC performs very well but test AUC performs poorly, it shows that the model memorizes the patterns instead of understanding it. This approach can ensure the selected models perform well while having robust and reliable predictions.

Other than that, the prediction map for the distribution of Malayan Partridge will be compared with the actual occurrence of the species. Having a high train and test AUC does not mean the model can perform well in generating an ecological sense prediction map for the species occurrence. Also, the model must be good in predicting as it will affect the prediction for future distribution of the species. By combining statistical evaluation and visual validation of distribution maps, it ensures that the best models selected make sense for SDM applications.

After deciding on a best pseudoabsence settings, the performance of algorithms will be compared with each other to decide the best models. The best models will be used for predicting the species of Malayan Partridge, the models will be implemented to predict the future distribution of the species. It will be using Shared Socioeconomic Pathway 2-4.5 that represents moderate gas emissions for the year 2021 to year 2040 and year 2041 to year 2060.

3.1.6 Output

The final output of this project will identify the best pseudoabsence data settings for the SDM to predict the occurrence of Malayan Partridge. Then, the best models will be chosen from the best pseudoabsence data settings that have been identified. The best models will generate a habitat suitability map for the current and future occurrence of Malayan Partridge. Other than that, the project will identify which bioclimates are the most important for the prediction of Malayan Partridge in Malaysia.

Chapter 4
System Implementation and Results

4.1 Hardware

The hardware involved in this project is a desktop. The desktop will be used throughout the development process, testing process and model evaluation of SDM. The specification of the desktop is stated in Table 4.1.1.

Description	**Specifications**
Processor	Intel Core i5-8400 CPU @ 2.80 GHz
Operating System	Windows 10 Pro 64-bit
Graphic	NVIDIA GeForce GTX 1070 Ti
Memory	16 GB DDR4 RAM
Storage	512 GB SSD

Table 4.1.1 Specifications of Desktop

4.2 Software

1. MaxEnt

   MaxEnt (Maximum Entropy) is a machine learning algorithm that can be used for Species Distribution Modelling based on presence-only data. It was chosen for this project as it is effective to predict the occurrence of Malayan Partridge.

2. Google Colab

   Google Colab is the main software that I use to do coding for doing Random Forest for variable selection of bioclimatic variables.

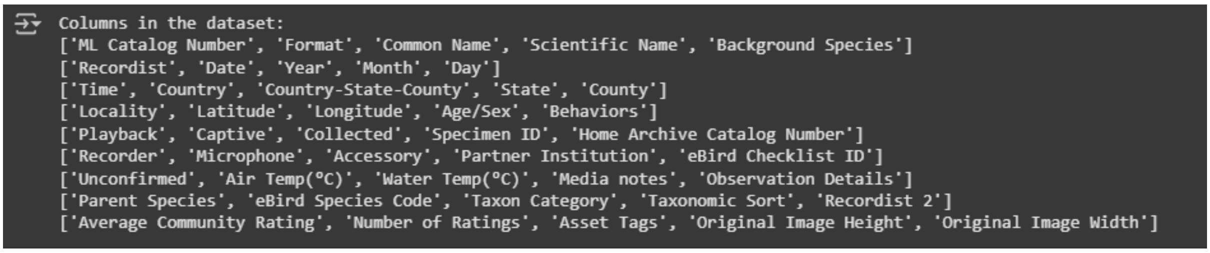
3. QGIS

   QGIS is an open-source Geographic Information System (GIS) used for spatial analysis and visualizing environmental layers and occurrence data. It was used to process my data for Species Distribution Model.

4.3 Data Preprocessing Implementation

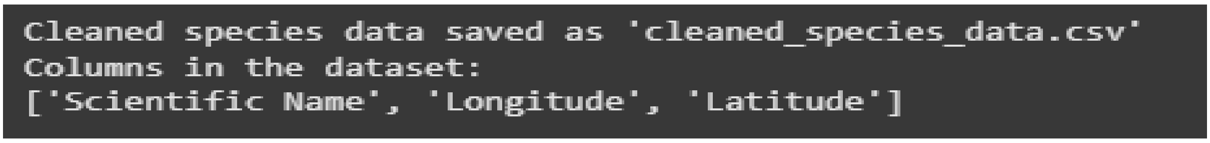
4.3.1 Occurrence data

The occurrence data of Malayan Partridge is being obtained, and this is the initial version of the data. There will be many columns that are not needed for the development of SDM as it only needs species name, longitude and latitude. Figure 4.3.1.1 shows the initial columns available in the dataset. Therefore, there is a need to reduce the number of columns. Figure 4.3.1.3 shows the coding for this process, and it also involves clearing any duplicate occurrence to ensure that there is no bias during data modelling. Figure 4.3.1.2 shows the results after cleaning the dataset.



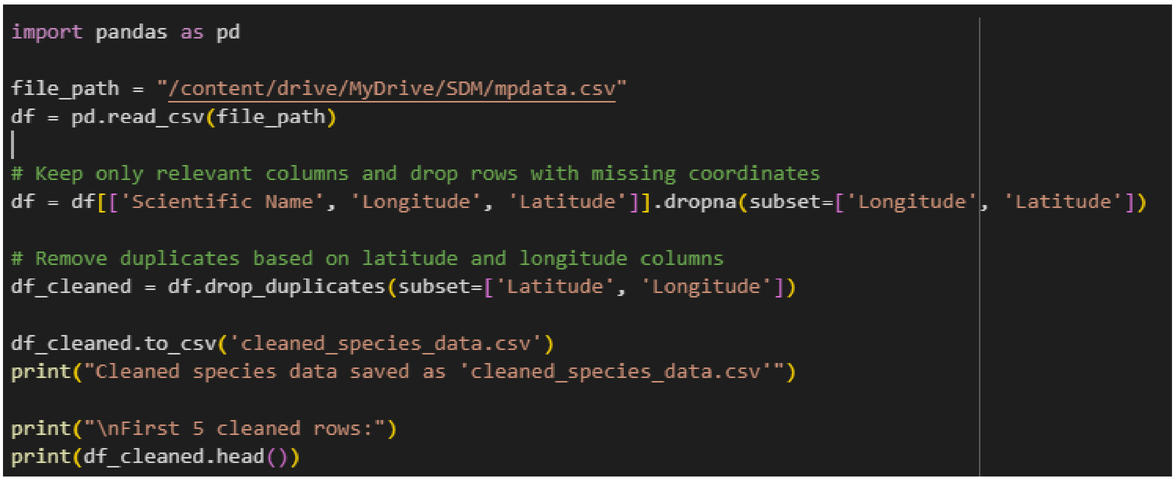
```
Columns in the dataset:
['ML Catalog Number', 'Format', 'Common Name', 'Scientific Name', 'Background Species']
['Recordist', 'Date', 'Year', 'Month', 'Day']
['Time', 'Country', 'Country-State-County', 'State', 'County']
['Locality', 'Latitude', 'Longitude', 'Age/Sex', 'Behaviors']
['Playback', 'Captive', 'Collected', 'Specimen ID', 'Home Archive Catalog Number']
['Recorder', 'Microphone', 'Accessory', 'Partner Institution', 'eBird Checklist ID']
['Unconfirmed', 'Air Temp(°C)', 'Water Temp(°C)', 'Media notes', 'Observation Details']
['Parent Species', 'eBird Species Code', 'Taxon Category', 'Taxonomic Sort', 'Recordist 2']
['Average Community Rating', 'Number of Ratings', 'Asset Tags', 'Original Image Height', 'Original Image Width']
```

Figure 4.3.1.1 Columns of Original Dataset



```
Cleaned species data saved as 'cleaned_species_data.csv'
Columns in the dataset:
['Scientific Name', 'Longitude', 'Latitude']
```

Figure 4.3.1.2 Columns of Updated Dataset



```python
import pandas as pd

file_path = "/content/drive/MyDrive/SDM/mpdata.csv"
df = pd.read_csv(file_path)

# Keep only relevant columns and drop rows with missing coordinates
df = df[['Scientific Name', 'Longitude', 'Latitude']].dropna(subset=['Longitude', 'Latitude'])

# Remove duplicates based on latitude and longitude columns
df_cleaned = df.drop_duplicates(subset=['Latitude', 'Longitude'])

df_cleaned.to_csv('cleaned_species_data.csv')
print("Cleaned species data saved as 'cleaned_species_data.csv'")

print("\nFirst 5 cleaned rows:")
print(df_cleaned.head())
```

Figure 4.3.1.3 Coding for Processing Occurrence Data

4.3.2 Bioclimatic Variables

Initial bioclimatic variables downloaded from WorldClim contain the data for the entire world. However, the target area of this project is only Malaysia. Having data for the whole world not only will slow down the processing time, also may cause unnecessary noise and complexity for the model. Therefore, the data need to be clipped according to the target area that is Malaysia, and it can be done using QGIS, a GIS software. Figure 4.3.2.1 shows the initial bioclimatic data and Figure 4.3.2.2 shows the bioclimatic data after being clipped to only Malaysia. This process needs to be done for the current and future bioclimatic variables data.

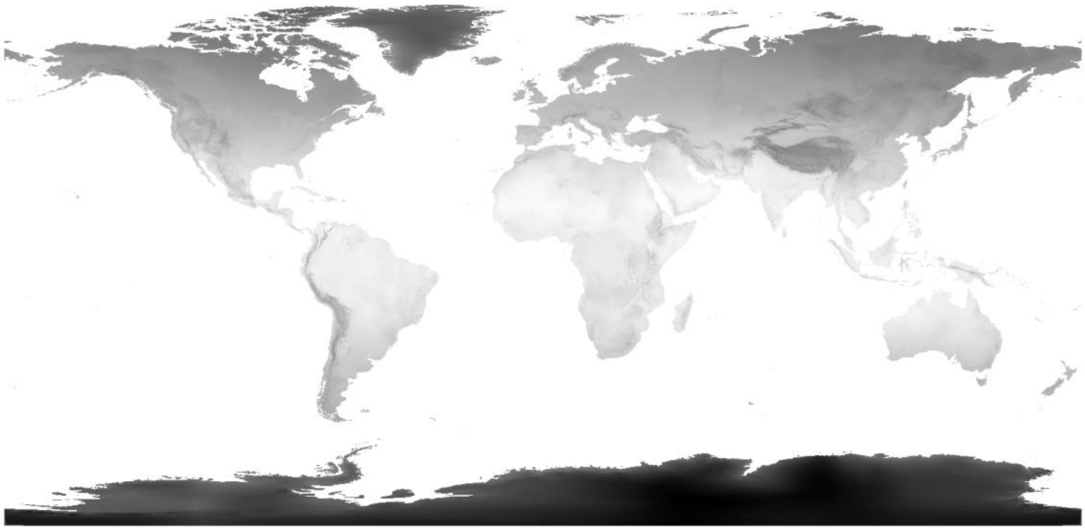
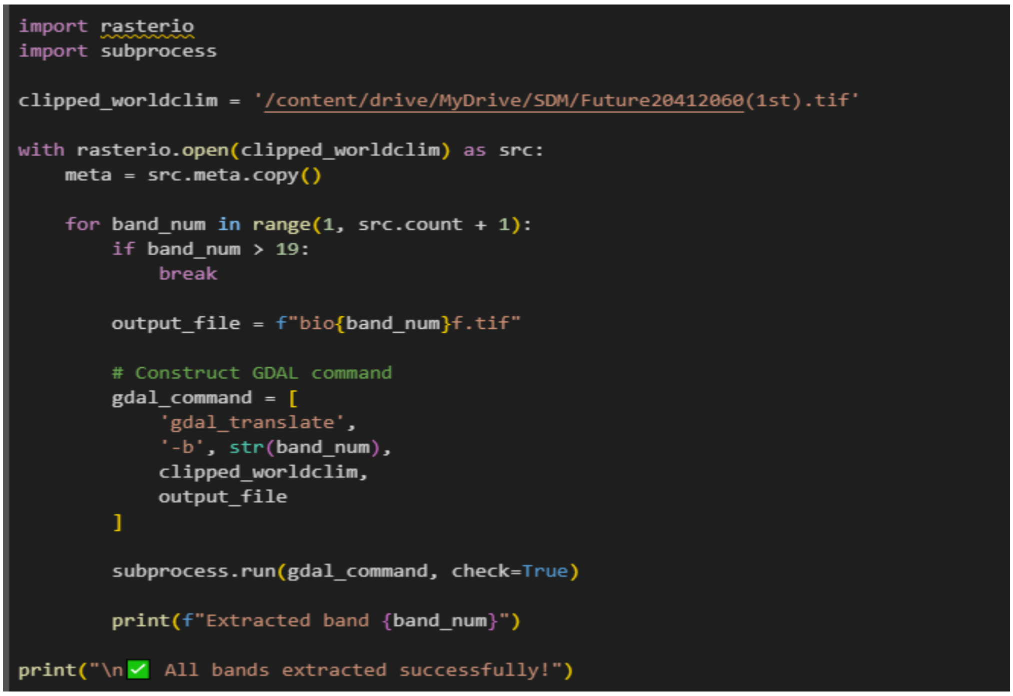


Figure 4.3.2.1 Initial Bioclimatic Variable Data

Figure 4.3.2.2 Clipped Bioclimatic Variable Data

For future bioclimatic variables obtained from WorldClim, all the bioclimatic variables are included in one GeoTIFF (tif) file. Therefore, it means there are 19 bands combined, each band represents a bioclimate variable, for example band 1 means bio1. To implement it in MaxEnt, there is a need to extract these into 19 individual bands. Figure 4.3.2.3 shows the coding for this process. GDAL translation needs to be utilized to complete the extraction process. Coding involves opening the raster file of the future bioclimate data with a loop. Then it will use GDAL command to complete the process of extraction to the notebook.



```python
import rasterio
import subprocess

clipped_worldclim = '/content/drive/MyDrive/SDM/Future20412060(1st).tif'

with rasterio.open(clipped_worldclim) as src:
    meta = src.meta.copy()

    for band_num in range(1, src.count + 1):
        if band_num > 19:
            break

        output_file = f"bio{band_num}f.tif"

        # Construct GDAL command
        gdal_command = [
            'gdal_translate',
            '-b', str(band_num),
            clipped_worldclim,
            output_file
        ]

        subprocess.run(gdal_command, check=True)

        print(f"Extracted band {band_num}")

print("\n✅ All bands extracted successfully!")
```

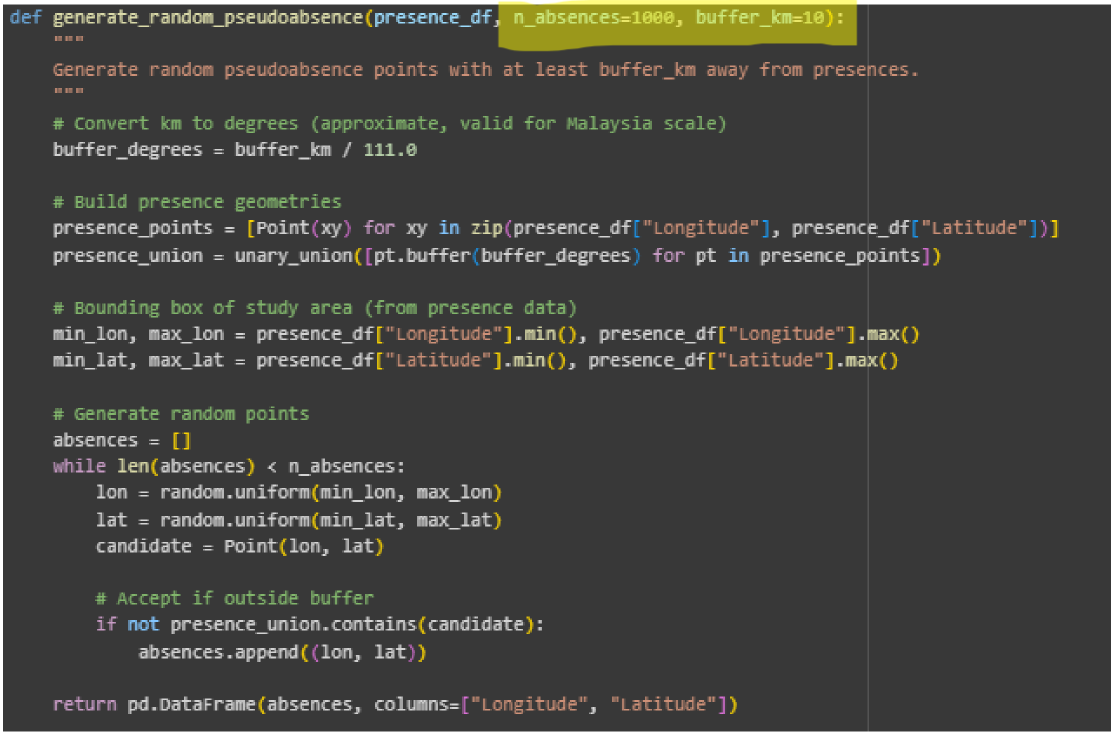
Figure 4.3.2.3 Coding for Band Extraction Process

4.3.3 Generating Pseudoabsence data

Since there are no true absence data for Malayan Partridge can be found, pseudoabsence data is generated for models that required such as RF, SVM and GLM. The buffer is set at 10km as per mentioned before in Section 3.1.2 Data Preprocessing. Therefore, the pseudoabsence data will generate at 10km away from the known presence location. The buffer distance and the number of pseudoabsence data can be set at the highlighted part of Figure 4.3.3.1. There will be 5 different pseudoabsence data generated in this process. By doing so, it can identify which pseudoabsence data settings will be suitable for developing SDM to predict the occurrence of Malayan Partridge.

Below are the different amounts of pseudoabsence data generated, with the number of presence data of 64:

- 1 to 1 ratio of pseudoabsence data (64 presence to 64 pseudoabsence)
- 1 to 2 ratio of pseudoabsence data (64 presence to 128 pseudoabsence)
- 1 to 4 ratio of pseudoabsence data (64 presence to 256 pseudoabsence)
- 1000 pseudoabsence data
- 10000 pseudoabsence data



```
def generate_random_pseudoabsence(presence_df, n_absences=1000, buffer_km=10):
    """
    Generate random pseudoabsence points with at least buffer_km away from presences.
    """
    # Convert km to degrees (approximate, valid for Malaysia scale)
    buffer_degrees = buffer_km / 111.0

    # Build presence geometries
    presence_points = [Point(xy) for xy in zip(presence_df["Longitude"], presence_df["Latitude"])]
    presence_union = unary_union([pt.buffer(buffer_degrees) for pt in presence_points])

    # Bounding box of study area (from presence data)
    min_lon, max_lon = presence_df["Longitude"].min(), presence_df["Longitude"].max()
    min_lat, max_lat = presence_df["Latitude"].min(), presence_df["Latitude"].max()

    # Generate random points
    absences = []
    while len(absences) < n_absences:
        lon = random.uniform(min_lon, max_lon)
        lat = random.uniform(min_lat, max_lat)
        candidate = Point(lon, lat)

        # Accept if outside buffer
        if not presence_union.contains(candidate):
            absences.append((lon, lat))

    return pd.DataFrame(absences, columns=["Longitude", "Latitude"])
```

Figure 4.3.3.1 Coding for Generating Pseudoabsence Data

4.4 Variable Selection

The variable selection process starts with correlation filter using Pearson Correlation. This step will identify and remove the correlated variables to avoid redundancy of information provided by the bioclimate variables. Threshold of |r| > 0.8 is applied, therefore any correlated value more than 0.8 will be removed. Figure 4.4.1 shows the correlation heatmap that visualizes how correlated each variable is. Referring to bio1 in Figure 4.4.1, it shows that it is highly correlated with bio3, bio5, bio6, bio8, bio9, bio10, bio11, and bio15. Therefore, these variables can be excluded from the final variable list. The same process will be applied to the rest of the variables. Results of correlation filter provides the output that removes 12 highly correlated variables that are bio10, bio11, bio13. bio16, bio17, bio18, bio3, bio4, bio5, bio6, bio8, bio9. The remaining variables are bio1, bio12, bio14, bio15, bio19, bio2, bio7 and will be used for variable selection techniques in the next step.

Next, three variable selection methods will be implemented to identify the final variables that will be used for SDM. Thresholds are implemented during this process. For RF and XGBoost, it will only select variables that account for 90% of total importance for the model. For permutation importance, it will consider variables that have value more than 0 as it means the variables some importance for the model. Variables that meet these criteria will receive 1 vote from each model. Figure 4.4.2 shows the results of the variable selection process. The final variables that will be implemented in SDM are bio1, bio12, bio15, bio19, bio2, and bio7.

Based on the final variables, it shows that the variables are mainly related to the temperature and precipitation patterns.

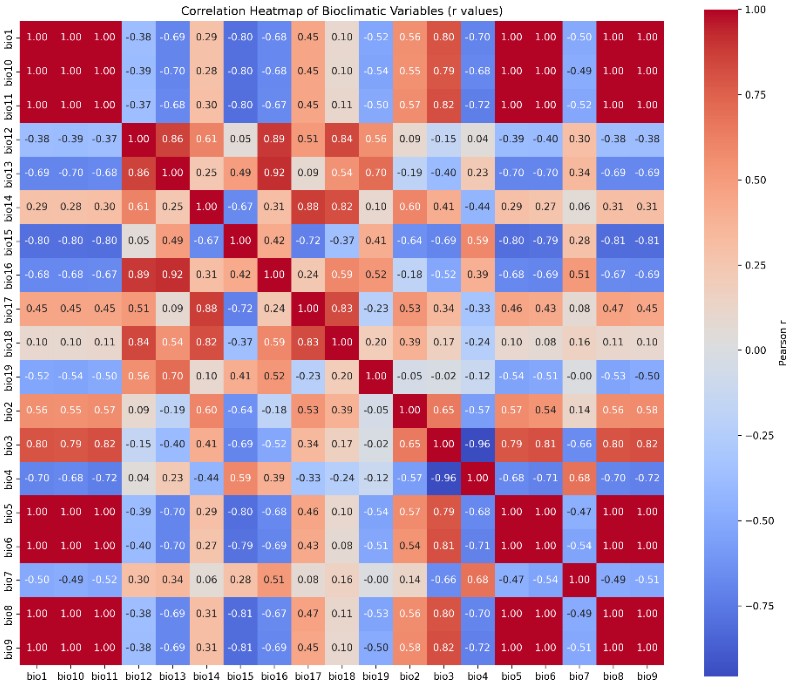
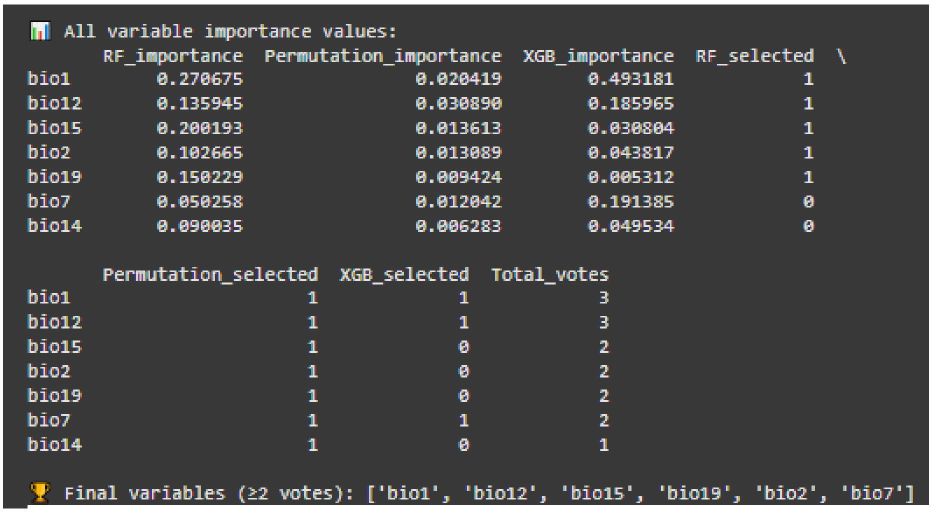


Figure 4.4.1 Heatmap for the Correlation between Bioclimate Variables



```
📊 All variable importance values:
       RF_importance  Permutation_importance  XGB_importance  RF_selected  \
bio1        0.270675                0.020419        0.493181            1
bio12       0.135945                0.030890        0.185965            1
bio15       0.200193                0.013613        0.030804            1
bio2        0.102665                0.013089        0.043817            1
bio19       0.150229                0.009424        0.005312            1
bio7        0.050258                0.012042        0.191385            0
bio14       0.090035                0.006283        0.049534            0

       Permutation_selected  XGB_selected  Total_votes
bio1                      1             1            3
bio12                     1             1            3
bio15                     1             0            2
bio2                      1             0            2
bio19                     1             0            2
bio7                      1             1            2
bio14                     1             0            1

🏆 Final variables (≥2 votes): ['bio1', 'bio12', 'bio15', 'bio19', 'bio2', 'bio7']
```

Figure 4.4.2 Results of Variable Selection Process

4.5 Model Development

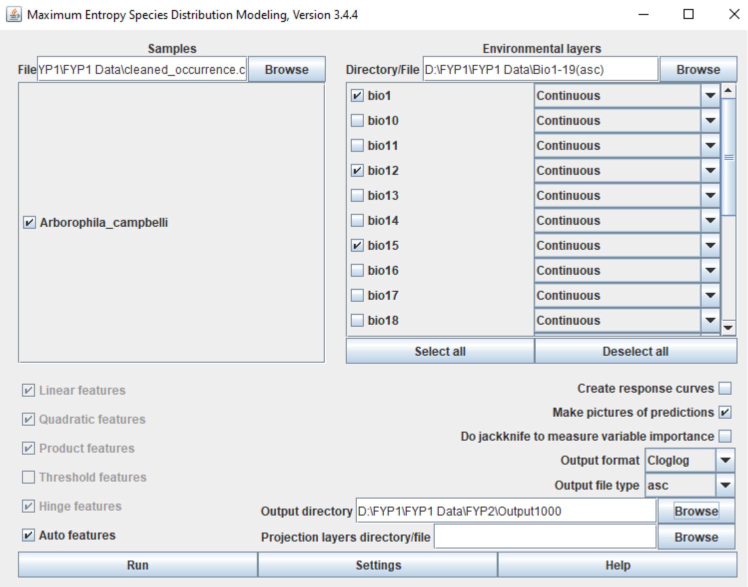
The five models will train using different amount of pseudoabsence data that was generated from the data preprocessing part. The input data for the models are the occurrence data and pseudoabsence data of Malayan Partridge and the final bioclimate variables set. The models will develop using different pseudoabsence settings mentioned in section 4.3.3. Figure 4.5.1 shows the final variables that will be implemented in the models for SDM.

Bioclimate Variables	**Interpretation**
Bio1	Annual Mean Temperature
Bio2	Mean Diurnal Range
Bio7	Temperature Annual Range
Bio12	Annual Precipitation
Bio15	Precipitation Seasonality
Bio19	Precipitation of Coldest Quarter

Table 4.5.1 Final Variables for the Input of SDM

4.5.1 Development of MaxEnt model

In terms of development of model, MaxEnt differs from other models. MaxEnt uses the existing MaxEnt software. Figure 4.5.1.1 shows the MaxEnt application interface used to develop the MaxEnt model. The model also configured with 5-fold and different background points that is equivalent to the pseudoabsence points.


Maximum Entropy Species Distribution Modeling, Version 3.4.4

Samples

File YP1\FYP1 Data\cleaned_occurrence.c

☑ Arborophila_campbelli

Environmental layers

Directory/File D:\FYP1\FYP1 Data\Bio1-19(asc)

☑ bio1 Continuous
☐ bio10 Continuous
☐ bio11 Continuous
☑ bio12 Continuous
☐ bio13 Continuous
☐ bio14 Continuous
☑ bio15 Continuous
☐ bio16 Continuous
☐ bio17 Continuous
☐ bio18 Continuous

☑ Linear features
☑ Quadratic features
☑ Product features
☐ Threshold features
☑ Hinge features
☑ Auto features

Create response curves ☐
Make pictures of predictions ☑
Do jackknife to measure variable importance ☐
Output format Cloglog
Output file type asc
Output directory D:\FYP1\FYP1 Data\FYP2\Output1000
Projection layers directory/file

Figure 4.5.1.1 MaxEnt Interface

4.5.2 Development of RF, GLM and SVM models

Three of these models will utilize GridSearchCV to try to find the best parameters for each model. These models will take the occurrence data of Malayan Partridge and the data with the points of bioclimates as input. The models' initial starting parameters can be referred to in Figure 4.5.2.1.

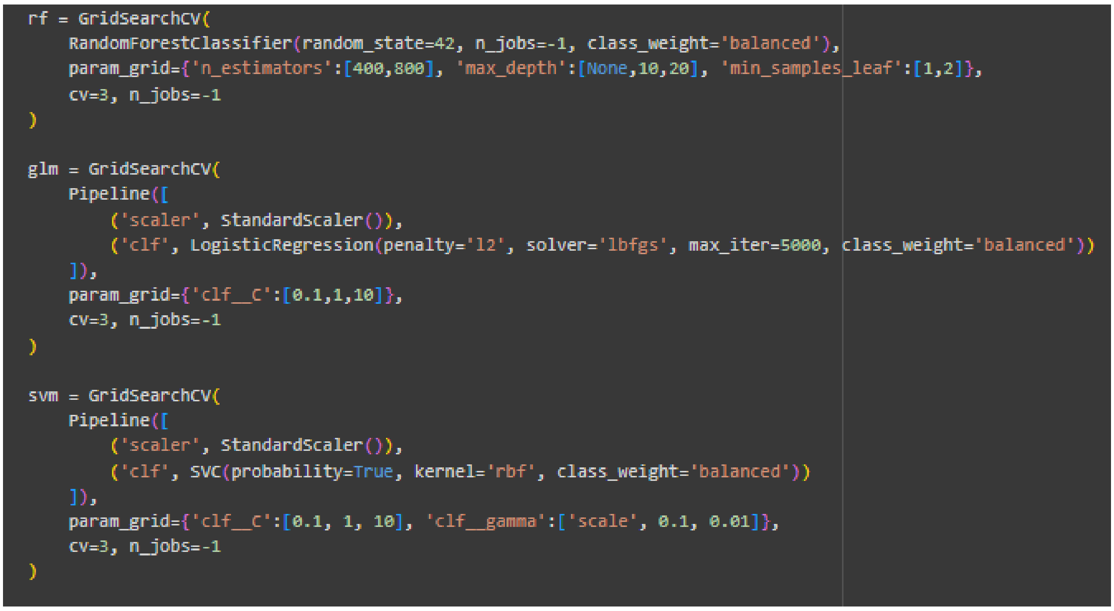
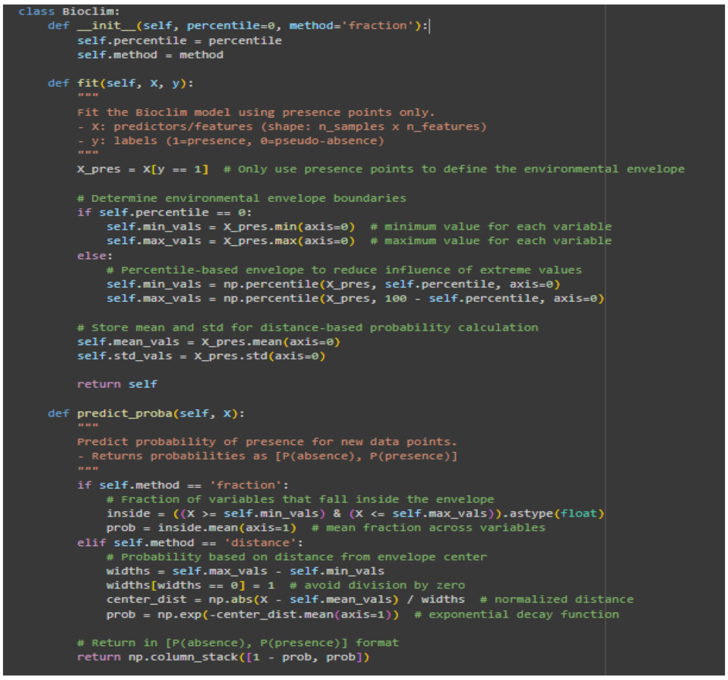

```
rf = GridSearchCV(
    RandomForestClassifier(random_state=42, n_jobs=-1, class_weight='balanced'),
    param_grid={'n_estimators':[400,800], 'max_depth':[None,10,20], 'min_samples_leaf':[1,2]},
    cv=3, n_jobs=-1
)

glm = GridSearchCV(
    Pipeline([
        ('scaler', StandardScaler()),
        ('clf', LogisticRegression(penalty='l2', solver='lbfgs', max_iter=5000, class_weight='balanced'))
    ]),
    param_grid={'clf__C':[0.1,1,10]},
    cv=3, n_jobs=-1
)

svm = GridSearchCV(
    Pipeline([
        ('scaler', StandardScaler()),
        ('clf', SVC(probability=True, kernel='rbf', class_weight='balanced'))
    ]),
    param_grid={'clf__C':[0.1, 1, 10], 'clf__gamma':['scale', 0.1, 0.01]},
    cv=3, n_jobs=-1
)
```

Figure 4.5.2.1 Coding Implementation for RF, GLM and SVM Models

4.5.3 Development of Bioclim model

Bioclim model uses only the presence points of the occurrence of Malayan Partridge and bioclimate variable as input. It calculates the minimum and maximum value of the bioclimate variable where the species occurs to define the environmental envelope. Then it will calculate the probability of the species' presence by checking whether the condition of the area matches the envelope. This method does not require any parameters tuning as it relies directly on the environment of the species. Figure 4.5.3.1 shows the coding implementation for the development of Bioclim model.



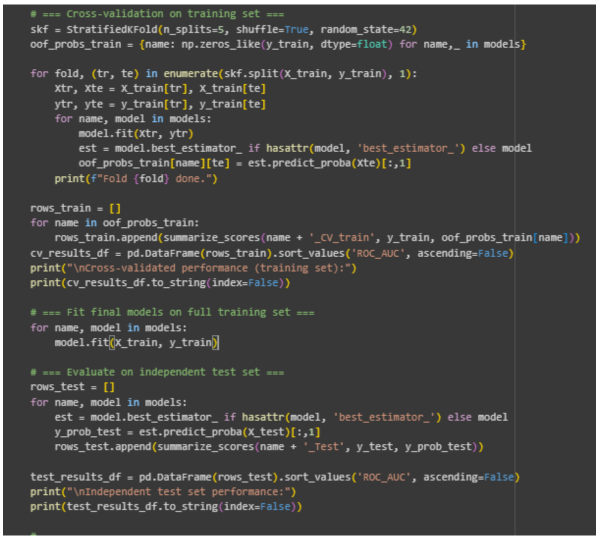
```
class Bioclim:
    def __init__(self, percentile=0, method='fraction'):
        self.percentile = percentile
        self.method = method

    def fit(self, X, y):
        """
        Fit the Bioclim model using presence points only.
        - X: predictors/features (shape: n_samples x n_features)
        - y: labels (1=presence, 0=pseudo-absence)
        """
        X_pres = X[y == 1]  # Only use presence points to define the environmental envelope

        # Determine environmental envelope boundaries
        if self.percentile == 0:
            self.min_vals = X_pres.min(axis=0)  # minimum value for each variable
            self.max_vals = X_pres.max(axis=0)  # maximum value for each variable
        else:
            # Percentile-based envelope to reduce influence of extreme values
            self.min_vals = np.percentile(X_pres, self.percentile, axis=0)
            self.max_vals = np.percentile(X_pres, 100 - self.percentile, axis=0)

        # Store mean and std for distance-based probability calculation
        self.mean_vals = X_pres.mean(axis=0)
        self.std_vals = X_pres.std(axis=0)

        return self

    def predict_proba(self, X):
        """
        Predict probability of presence for new data points.
        - Returns probabilities as [P(absence), P(presence)]
        """
        if self.method == 'fraction':
            # Fraction of variables that fall inside the envelope
            inside = ((X >= self.min_vals) & (X <= self.max_vals)).astype(float)
            prob = inside.mean(axis=1)  # mean fraction across variables
        elif self.method == 'distance':
            # Probability based on distance from envelope center
            widths = self.max_vals - self.min_vals
            widths[widths == 0] = 1  # avoid division by zero
            center_dist = np.abs(X - self.mean_vals) / widths  # normalized distance
            prob = np.exp(-center_dist.mean(axis=1))  # exponential decay function

        # Return in [P(absence), P(presence)] format
        return np.column_stack([1 - prob, prob])
```

Figure 4.5.3.1 Coding Implementation for Bioclim

4.6 Evaluation and Comparison

4.6.1 Test AUC and Train AUC

The model evaluation focuses on three main steps that are cross validation, test AUC and train AUC. First it will use 5-fold cross-validation on the training data by splitting it into parts to let the model train on some folds and test on the remaining fold. It can help to check how the model performs and avoid overfitting. After that the model will be trained on the full training set. Lastly, the model will test on an independent set to identify how well the model can perform to new unseen data.

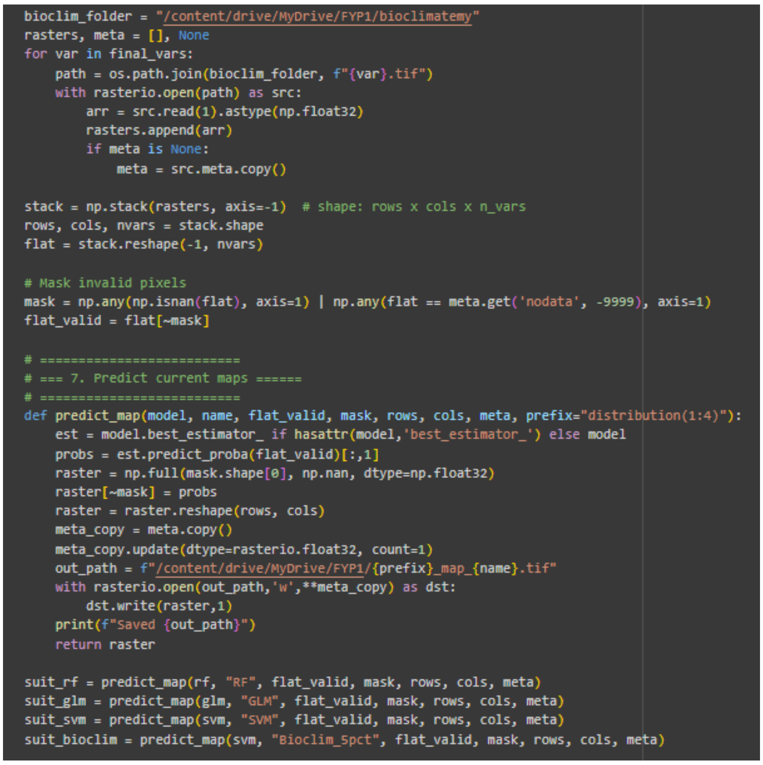


```
# === Cross-validation on training set ===
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
oof_probs_train = {name: np.zeros_like(y_train, dtype=float) for name,_ in models}

for fold, (tr, te) in enumerate(skf.split(X_train, y_train), 1):
    Xtr, Xte = X_train[tr], X_train[te]
    ytr, yte = y_train[tr], y_train[te]
    for name, model in models:
        model.fit(Xtr, ytr)
        est = model.best_estimator_ if hasattr(model, 'best_estimator_') else model
        oof_probs_train[name][te] = est.predict_proba(Xte)[:,1]
    print(f"Fold {fold} done.")

rows_train = []
for name in oof_probs_train:
    rows_train.append(summarize_scores(name + '_CV_train', y_train, oof_probs_train[name]))
cv_results_df = pd.DataFrame(rows_train).sort_values('ROC_AUC', ascending=False)
print("\nCross-validated performance (training set):")
print(cv_results_df.to_string(index=False))

# === Fit final models on full training set ===
for name, model in models:
    model.fit(X_train, y_train)

# === Evaluate on independent test set ===
rows_test = []
for name, model in models:
    est = model.best_estimator_ if hasattr(model, 'best_estimator_') else model
    y_prob_test = est.predict_proba(X_test)[:,1]
    rows_test.append(summarize_scores(name + '_Test', y_test, y_prob_test))

test_results_df = pd.DataFrame(rows_test).sort_values('ROC_AUC', ascending=False)
print("\nIndependent test set performance:")
print(test_results_df.to_string(index=False))
```

Figure 4.6.1.1 Coding Implementation for Test and Train AUC

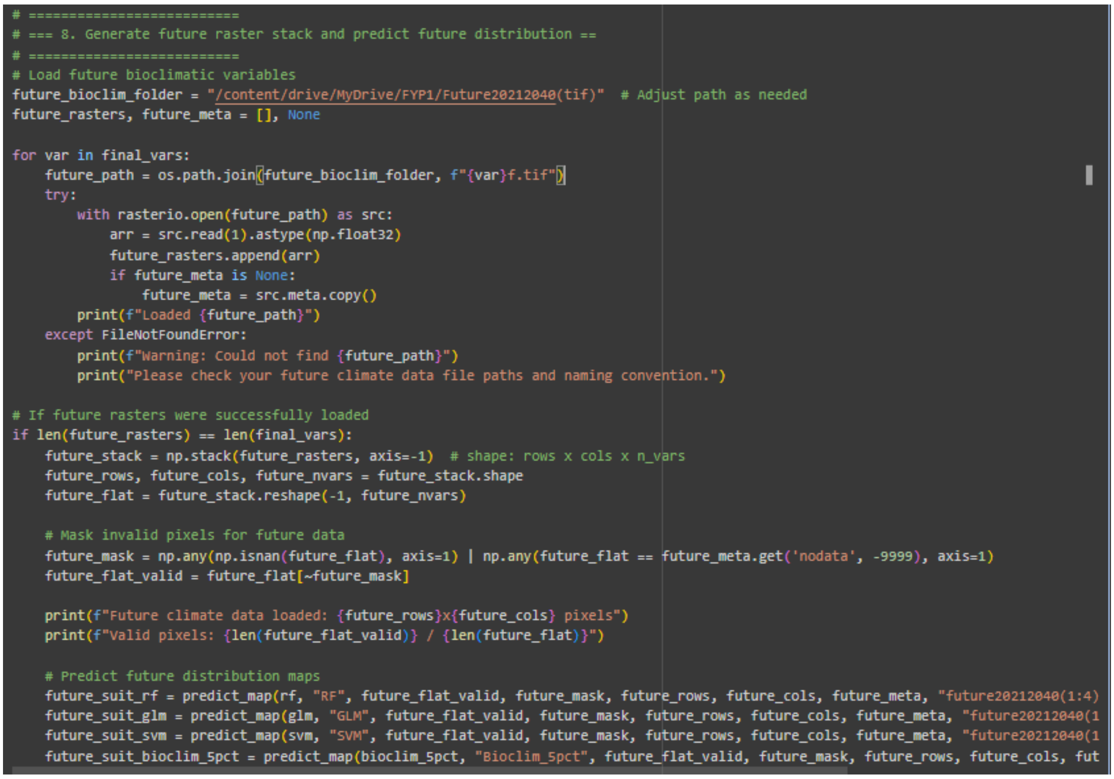
4.6.2 Generating Distribution Map

The distribution maps for Malayan Partridge were generated by stacking the raster layers for the bioclimatic variable in the study area. Each layer has a bioclimatic variable and combine each layer into a multi-dimensional array. The trained models of RF, GLM, SVM and Bioclim will estimate the probability of Malayan Partridge for each pixel and producing the map of habitat suitability for the species. Figure 4.6.2.1 shows the coding implementation for generating current distribution map for Malayan Partridge. To identify the future potential habitat suitability for Malayan Partridge, future suitability map were generated using Shared Socioeconomic Pathways 2-4.5 for the year 2021 to year 2040 and year 2041 to year 2060. Figure 4.6.2.2 shows the coding implementation for generating future distribution suitability map for Malayan Partridge.



```
bioclim_folder = "/content/drive/MyDrive/FYP1/bioclimatemy"
rasters, meta = [], None
for var in final_vars:
    path = os.path.join(bioclim_folder, f"{var}.tif")
    with rasterio.open(path) as src:
        arr = src.read(1).astype(np.float32)
        rasters.append(arr)
        if meta is None:
            meta = src.meta.copy()

stack = np.stack(rasters, axis=-1)  # shape: rows x cols x n_vars
rows, cols, nvars = stack.shape
flat = stack.reshape(-1, nvars)

# Mask invalid pixels
mask = np.any(np.isnan(flat), axis=1) | np.any(flat == meta.get('nodata', -9999), axis=1)
flat_valid = flat[~mask]

# ==========================
# === 7. Predict current maps ======
# ==========================
def predict_map(model, name, flat_valid, mask, rows, cols, meta, prefix="distribution(1:4)"):
    est = model.best_estimator_ if hasattr(model,'best_estimator_') else model
    probs = est.predict_proba(flat_valid)[:,1]
    raster = np.full(mask.shape[0], np.nan, dtype=np.float32)
    raster[~mask] = probs
    raster = raster.reshape(rows, cols)
    meta_copy = meta.copy()
    meta_copy.update(dtype=rasterio.float32, count=1)
    out_path = f"/content/drive/MyDrive/FYP1/{prefix}_map_{name}.tif"
    with rasterio.open(out_path,'w',**meta_copy) as dst:
        dst.write(raster,1)
    print(f"Saved {out_path}")
    return raster

suit_rf = predict_map(rf, "RF", flat_valid, mask, rows, cols, meta)
suit_glm = predict_map(glm, "GLM", flat_valid, mask, rows, cols, meta)
suit_svm = predict_map(svm, "SVM", flat_valid, mask, rows, cols, meta)
suit_bioclim = predict_map(svm, "Bioclim_5pct", flat_valid, mask, rows, cols, meta)
```

Figure 4.6.2.1 Coding Implementation for Generating Current Distribution Map



```
# ==========================
# === 8. Generate future raster stack and predict future distribution ==
# ==========================
# Load future bioclimatic variables
future_bioclim_folder = "/content/drive/MyDrive/FYP1/Future20212040(tif)"  # Adjust path as needed
future_rasters, future_meta = [], None

for var in final_vars:
    future_path = os.path.join(future_bioclim_folder, f"{var}f.tif")
    try:
        with rasterio.open(future_path) as src:
            arr = src.read(1).astype(np.float32)
            future_rasters.append(arr)
            if future_meta is None:
                future_meta = src.meta.copy()
        print(f"Loaded {future_path}")
    except FileNotFoundError:
        print(f"Warning: Could not find {future_path}")
        print("Please check your future climate data file paths and naming convention.")

# If future rasters were successfully loaded
if len(future_rasters) == len(final_vars):
    future_stack = np.stack(future_rasters, axis=-1)  # shape: rows x cols x n_vars
    future_rows, future_cols, future_nvars = future_stack.shape
    future_flat = future_stack.reshape(-1, future_nvars)

    # Mask invalid pixels for future data
    future_mask = np.any(np.isnan(future_flat), axis=1) | np.any(future_flat == future_meta.get('nodata', -9999), axis=1)
    future_flat_valid = future_flat[~future_mask]

    print(f"Future climate data loaded: {future_rows}x{future_cols} pixels")
    print(f"Valid pixels: {len(future_flat_valid)} / {len(future_flat)}")

    # Predict future distribution maps
    future_suit_rf = predict_map(rf, "RF", future_flat_valid, future_mask, future_rows, future_cols, future_meta, "future20212040(1:4)
    future_suit_glm = predict_map(glm, "GLM", future_flat_valid, future_mask, future_rows, future_cols, future_meta, "future20212040(1
    future_suit_svm = predict_map(svm, "SVM", future_flat_valid, future_mask, future_rows, future_cols, future_meta, "future20212040(1
    future_suit_bioclim_5pct = predict_map(bioclim_5pct, "Bioclim_5pct", future_flat_valid, future_mask, future_rows, future_cols, fut
```

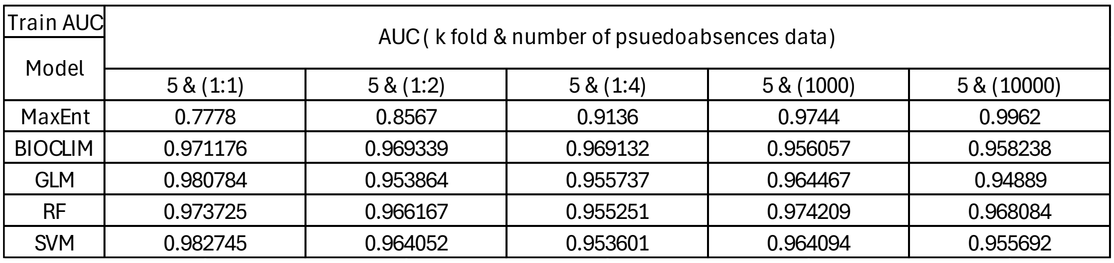
Figure 4.6.2.2 Coding Implementation for Generating Future Habitat Suitability Map

Chapter 5

Model Evaluation and Discussion

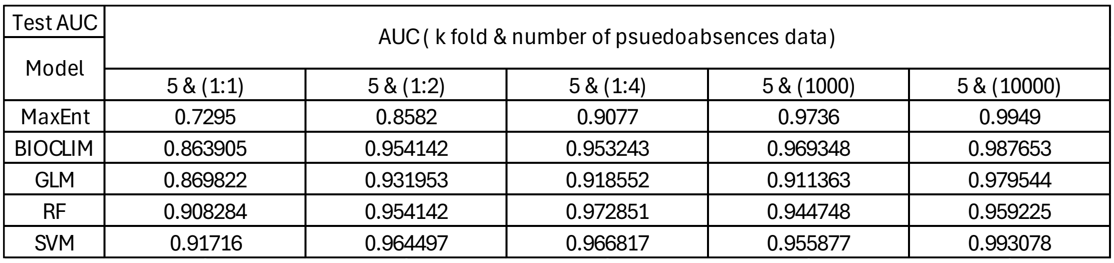
5.1 Results of Different Pseudoabsence settings

Five different algorithms are applied to model the distribution of current species for Malayan Partridge in Malaysia. To find the optimum settings for the model, these models are being trained with 5-fold cross validation and different numbers of pseudo absence data. Cross-validation is a method to identify how well a model generalizes independent data. It is an efficient way to reduce overfitting and ensure that models perform well on unseen data. In terms of the number of pseudo absence data, models are tested with different amounts to examine how the model performance is affected with more or fewer pseudo absence data. The dataset is also being divided into training and testing datasets. Train test split will be able to identify whether there is any overfitting for the model. If the testing AUC is significantly lower than the training AUC, it indicates that the model is overfitting as the model can learn patterns specific to the data but not able to generalize to new and unseen data. Figure 5.1.1 shows the train AUC for each of the models under different pseudoabsence settings and Figure 5.1.2 shows the test AUC for each of the models under different pseudoabsence settings.



Train AUC	AUC (k fold & number of psuedoabsences data)				
Model					
	5 & (1:1)	5 & (1:2)	5 & (1:4)	5 & (1000)	5 & (10000)
MaxEnt	0.7778	0.8567	0.9136	0.9744	0.9962
BIOCLIM	0.971176	0.969339	0.969132	0.956057	0.958238
GLM	0.980784	0.953864	0.955737	0.964467	0.94889
RF	0.973725	0.966167	0.955251	0.974209	0.968084
SVM	0.982745	0.964052	0.953601	0.964094	0.955692

Table 5.1.1 Results of Train AUC



Test AUC	AUC (k fold & number of psuedoabsences data)				
Model					
	5 & (1:1)	5 & (1:2)	5 & (1:4)	5 & (1000)	5 & (10000)
MaxEnt	0.7295	0.8582	0.9077	0.9736	0.9949
BIOCLIM	0.863905	0.954142	0.953243	0.969348	0.987653
GLM	0.869822	0.931953	0.918552	0.911363	0.979544
RF	0.908284	0.954142	0.972851	0.944748	0.959225
SVM	0.91716	0.964497	0.966817	0.955877	0.993078

Table 5.1.2 Results of Test AUC

Based on the results, it shows that the number of pseudo absence data will affect the model performance. Overall, the results show that increasing the amount of pseudo absence data can improve and increase the model ability to learn and differentiate between presence and pseudo absence locations and data. For MaxEnt algorithm, the training AUC and testing AUC increases steadily when the amount of pseudo absence data increases. For Bioclim, the training AUC and testing AUC are relatively similar as the model algorithm is more focused on predicting with presence data rather than pseudo absence data. For the models with GLM, RF and SVM, these models training AUC and testing AUC have more optimal performance with the settings of 5-fold and (1:4) of pseudoabsence data. The test AUC for the models is optimized without having an inflated training AUC.

For the model settings of 5-fold with 1 to 1 ratio and 5-fold with 1 to 2 ratio would not be considered as a suitable setting to develop a good model for SDM. This is because there is not enough pseudoabsence data for the model to predict the occurrence. Since there are less pseudoabsence points, it is easier for the models to generalize and learn the patterns to make predictions. This scenario is like the model settings of 5-fold and 10000 pseudoabsence data. The model settings of 5-fold and 10000 pseudo absence data show the best AUC results compared to other methods. However, there is a risk of overfitting due to the large number of pseudo absence data. The more the pseudoabsence data, the easier for the models to identify and differentiate between presence and absence especially when the species presence data are more clumped at few certain locations. Therefore, a moderate number of pseudoabsence data such as using 1 to 4 ratio of pseudoabsence data or 1000 pseudoabsence data will be a better option overall for the models of SDM. It can improve the model performance to identify suitable and unsuitable habitat for the species and minimize overfitting. To identify which is the better model’s settings between 1 to 4 ratio pseudoabsence data and 1000 pseudoabsence data, a comparison of distribution maps will be done. By looking at the distribution map of the current occurrence of Malayan Partridge, it can provide a clearer view on which model settings would be more suitable.

5.2 Comparing Distribution Map

This section will focus on comparing the distribution map generated by the models under different settings of pseudoabsence data. The top image will show the model with 1000 pseudoabsence data and the bottom image will show the model with 1 to 4 ratio of pseudoabsence data. Figure 5.2.1 is a map distribution of Malayan Partridge in Malaysia based on the known occurrence data of the species. This figure can be a reference to understanding which area is suitable for Malayan Partridge. Based on Figure 5.2.1, the species are located at the area of Bukit Fraser, Pahang. The area is a montane forest environment with cooler temperatures and higher precipitation patterns compared to low land.

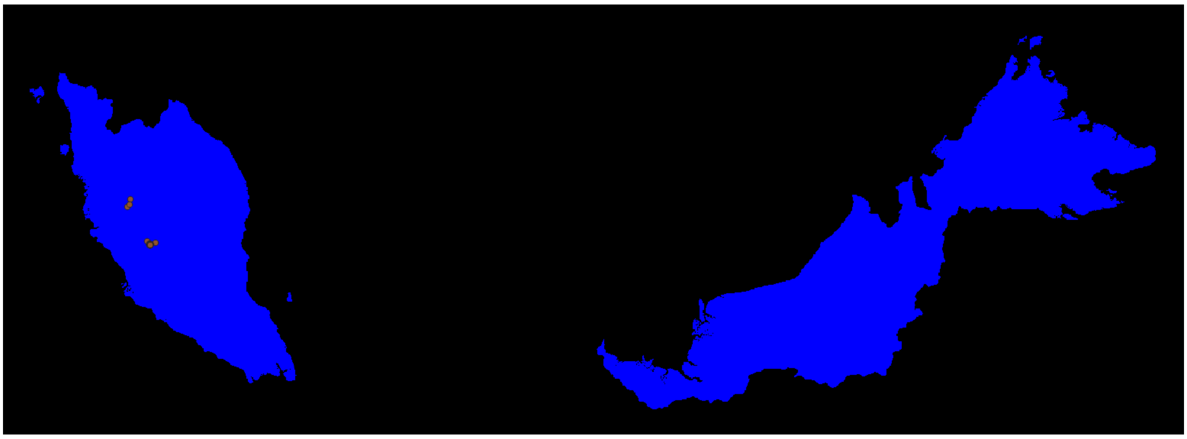


Figure 5.2.1 Occurrence Map of Malayan Partridge

5.2.1 MaxEnt Map Distribution

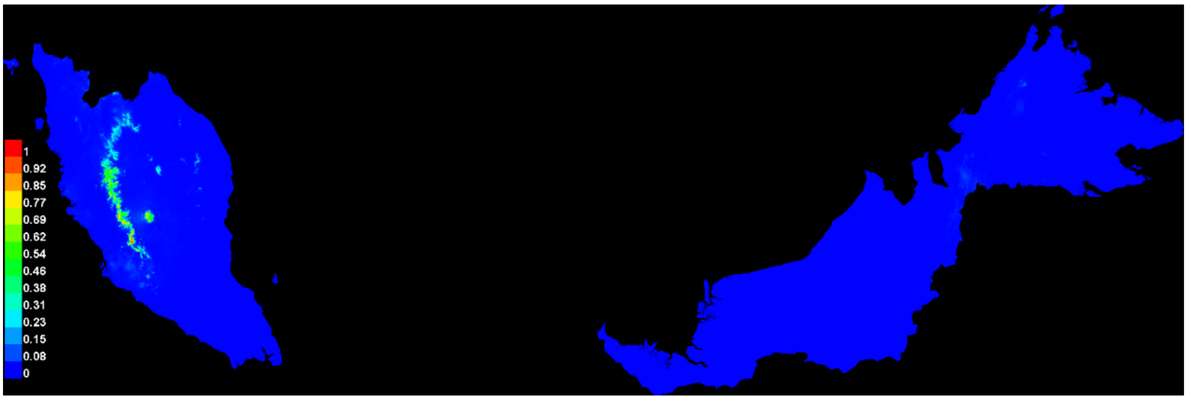


Figure 5.2.1.1 Prediction Map of MaxEnt with 1000 Pseudoabsence Data

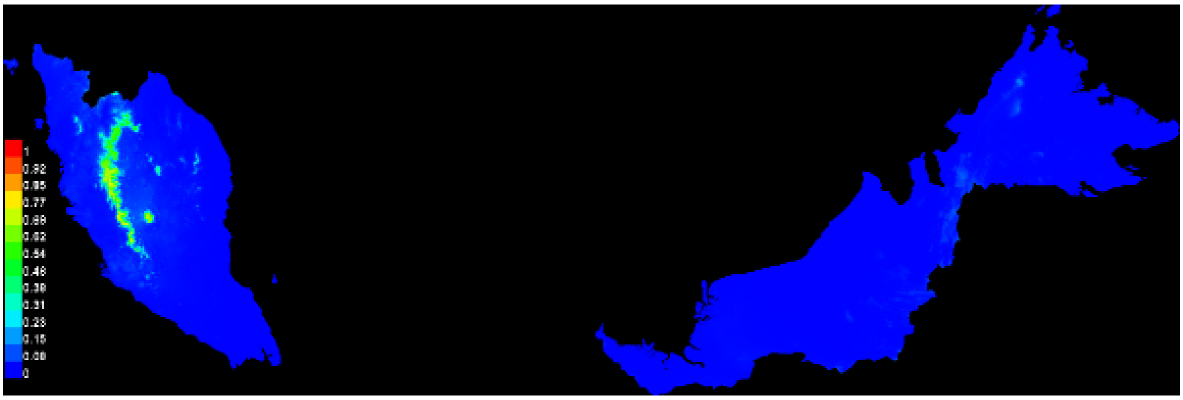


Figure 5.2.1.2 Prediction Map of MaxEnt with 1:4 ratio of Pseudoabsence Data

5.2.2 GLM Map Distribution

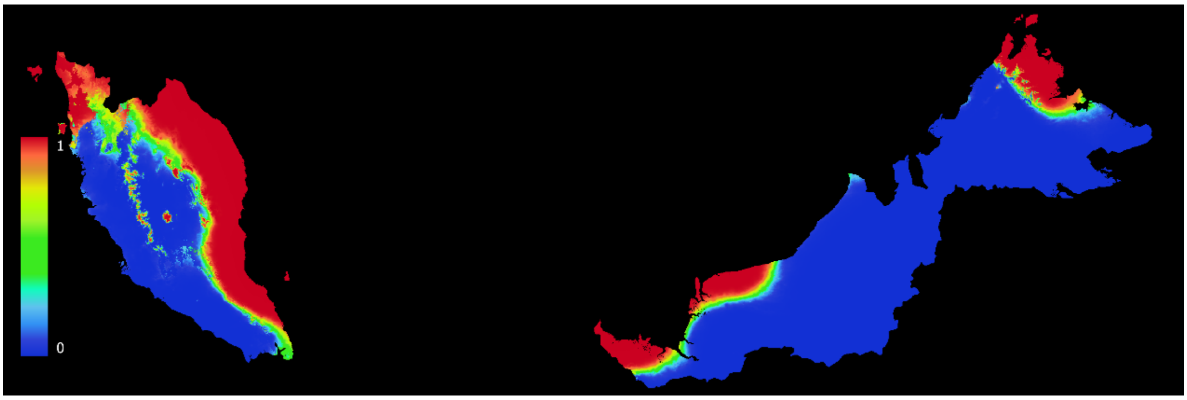


Figure 5.2.2.1 Prediction Map of GLM with 1000 Pseudoabsence Data

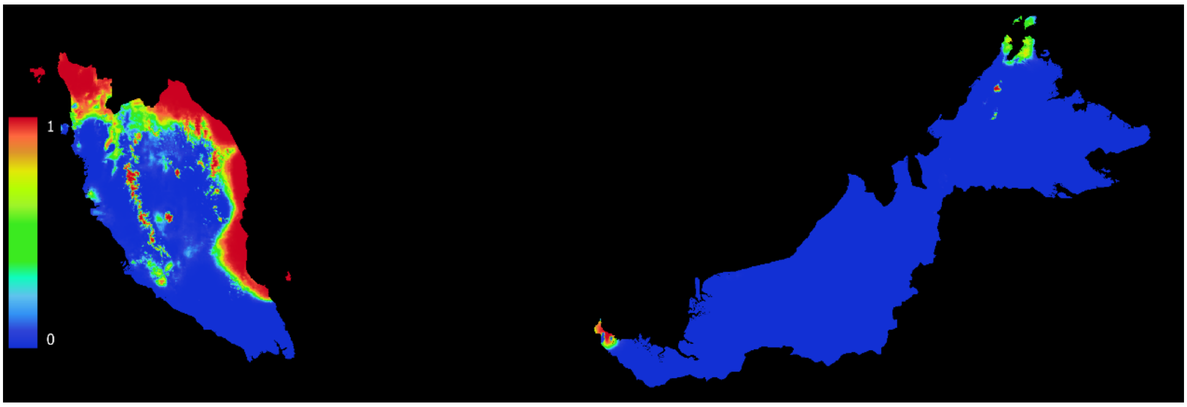


Figure 5.2.2.2 Prediction Map of GLM with 1:4 Pseudoabsence Data

5.2.3 RF Map Distribution

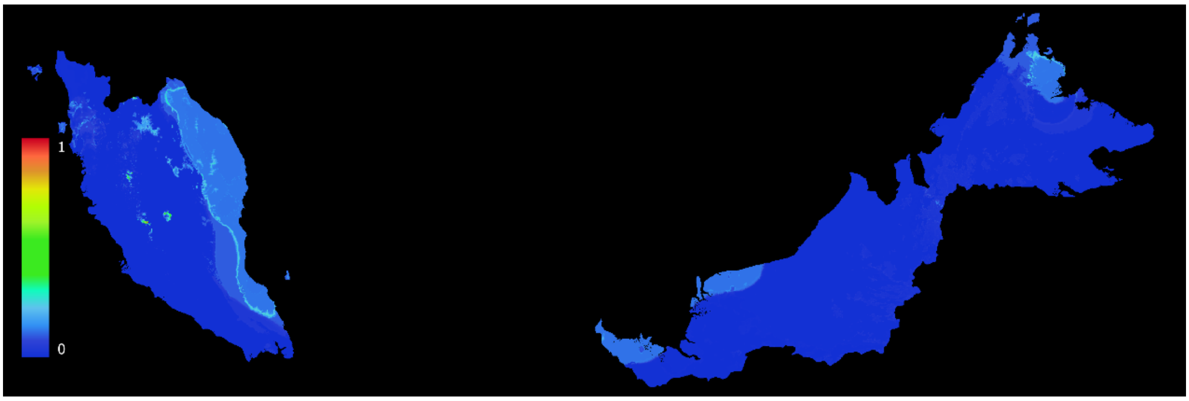


Figure 5.2.3.1 Prediction Map of RF with 1000 Pseudoabsence Data

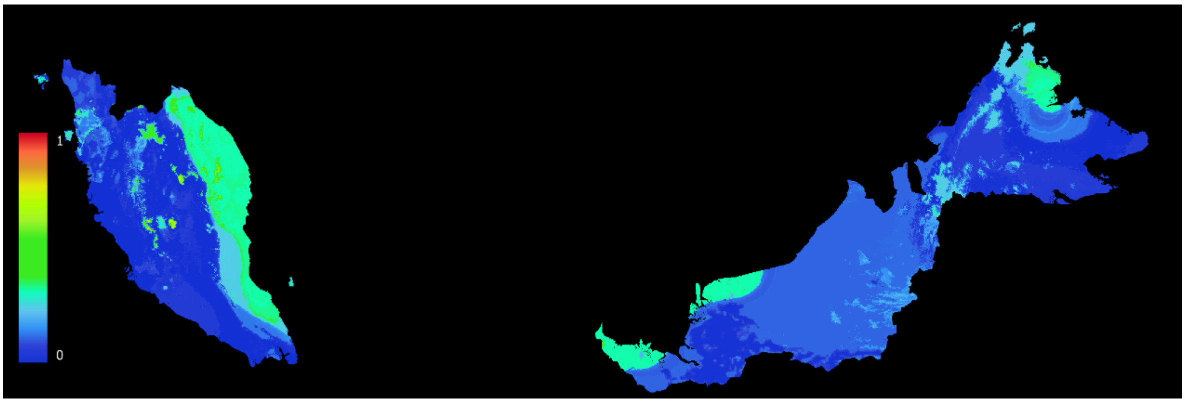


Figure 5.2.3.1 Prediction Map of RF with 1:4 Pseudoabsence Data

5.2.4 SVM Map Distribution

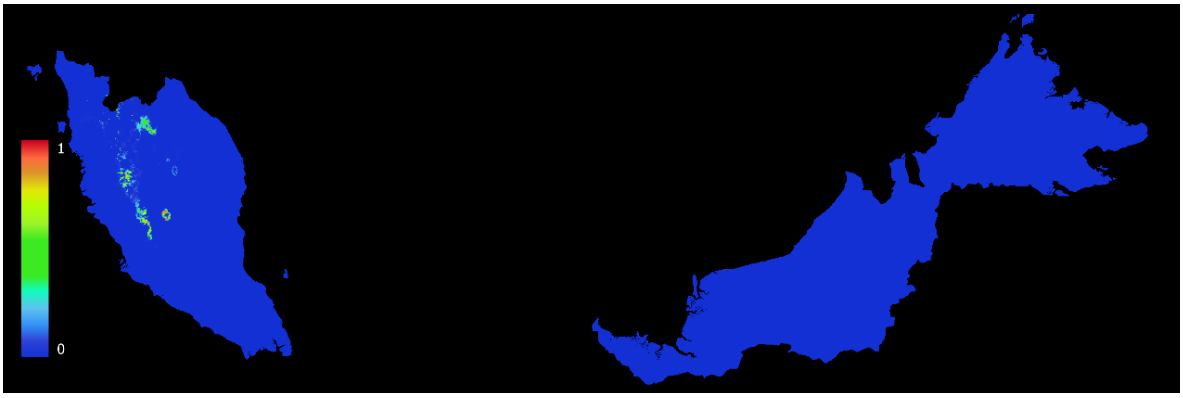


Figure 5.2.4.1 Prediction Map of SVM with 1000 Pseudoabsence Data

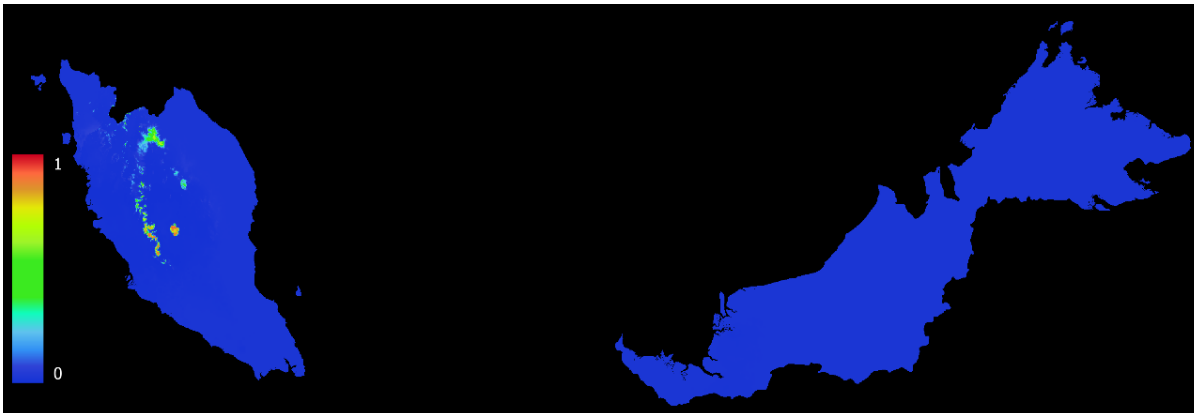


Figure 5.2.4.2 Prediction Map of SVM with 1:4 Pseudoabsence Data

5.2.5 Bioclim Map Distribution

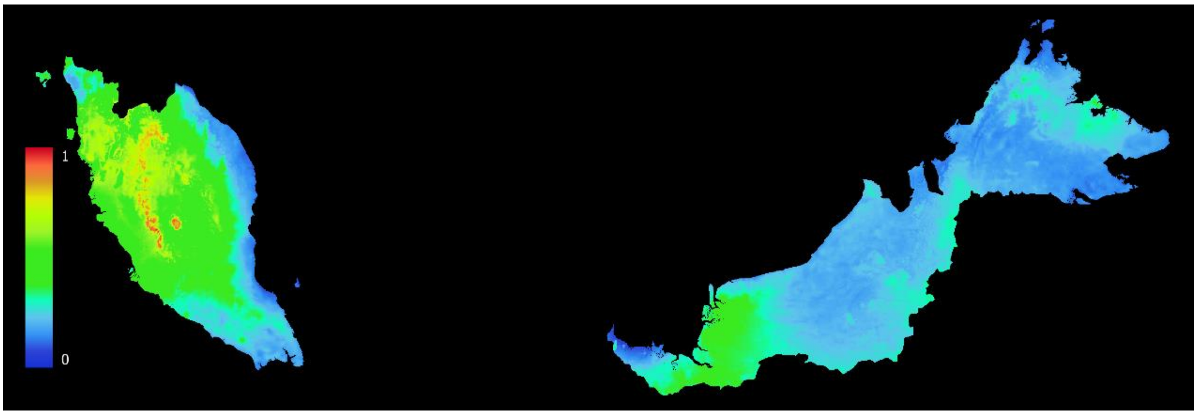


Figure 5.2.5.1 Prediction Map of Bioclim with 1000 Pseudoabsence Data

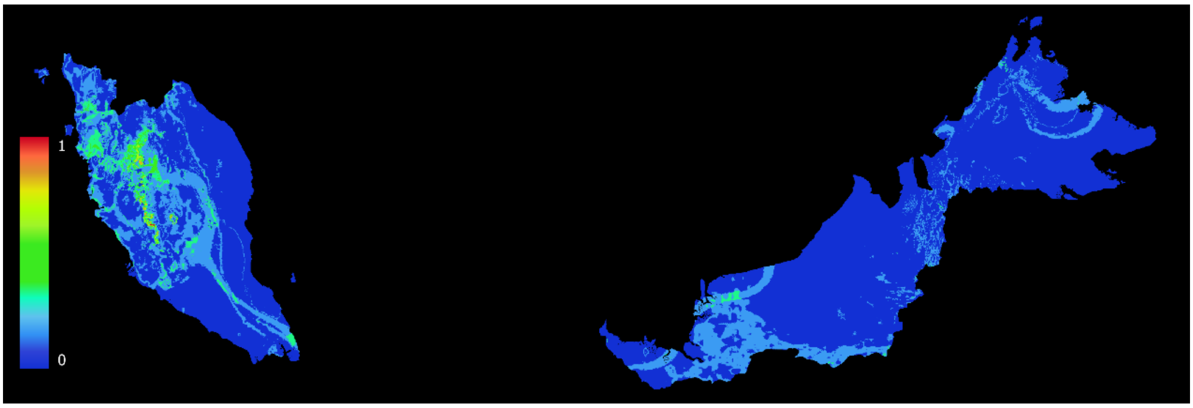


Figure 5.2.5.2 Prediction Map of Bioclim with 1:4 Pseudoabsence Data

5.2.6 Analyzing the Distribution Maps

The predicted distribution map for the species Malayan Partridge are generated with two different psuedoabsence settings, 1:4 ratio and 1000 points. The generated distribution map have a displayed a consistent patterns across the five different algorithms applied that are MaxEnt, GLM, RF, SVM and Bioclim, but they different settings shows some notable differences in the terms of habitat suitability. In overall, the psuedoabsence settings of 1:4 ratio shows a more broader and wider prediction. On the other hand, the setting of 1000 psuedoabsence data predicted a more restrictred prediction that focuses more on the main occurrence area of the species. This is trend is more obvious for the models of MaxEnt, RF, GLM and SVM. However, Bioclim shows an opposite trend, the distribution map generated with the settings of 1000 pseudoabsence data shows a broader prediction compared to the setting of 1:4 ratio of pseudabsence data. This is because Bioclim is a model that does not use pseudoabsence data like the 4 of the other models. Bioclim is an envelope model that will predict suitablity for a species when the environmental values fall within the range defined by presence points of the species. Therefore, there will be more points fall inside the envelope that lead to a broader prediction map when using 1000 pseudoabsence data.

In conclusion, the results highlight that with different number of pseudoabsence data, it can shift the prediction of the distribution between a broader and more restrictive prediction. Both of the settings can be useful in different application for Species Distribution Model. In terms of the setting of 1 to 4 ratio for psuedoabsence data, it will be more suitable when there is a need to explore unstudied areas for the species. In contrast, the setting with 1000 pseudoabsence data will be suitable if discovering the true occurrence of the species is more important.

Therefore, setting with 1000 psuedoabsence will be more suitable to predict the occurrence of Malayan Partridge as it is a species that is restricted to a montane forests area. Implementing a stricter prediction is better to reduce the risk of overpredicting the species distribution and provide a more realistic prediction for the occurrence of Malayan Partridge.

5.3 Selection of Best Models

Model	Train AUC	Test AUC
MaxEnt	0.9744	0.9736
Bioclim	0.956057	0.969348
GLM	0.964467	0.911363
RF	0.974209	0.944748
SVM	0.964094	0.955877

Table 5.3.1 Results of Models with 1000 Pseudoabsence Data

This section will analyze and select the best performed model in the pseudoabsence setting of 1000 pseudoabsence data. Based on the results produced, all 5 algorithms achieved and generated a very high predictive accuracy with the AUC values above 0.91. Among the 5 algorithms, MaxEnt has the best performance with train AUC of 0.9744 and test AUC of 0.9736. The value of train AUC and test AUC is nearly identical that proves the model generalize well. SVM is also another algorithm that performs well with just a slight difference between the train AUC and test AUC. Despite Bioclim being just an envelope model, it perform unexpectedly well to achieve a strong testing AUC value that suggest the ability to generalize. On the other hand, RF and GLM actually show signs of overfitting based on the result produced. Both of the models have a significant drop of test AUC from the train AUC. Therefore, the results shows that MaxEnt, Bioclim and SVM are great algorithm options for developing SDM while RF and GLM are not suitable in this senario. However, AUC values could not be the only consideration when choosing an algorithm for SDM. High train and test AUC does not mean the model can perform well in generating the right prediction map for the occurrence of Malayan Partridge.

Analyzing the prediction map for the occurrence of Malayan Partridge, Figure 5.2.2.1 and Figure 5.2.3.1 shows GLM and RF generated predictions that are not align with the known occurrence of Malayan Partridge. This is beacause RF model has overfitting that cause the predicted to be more fragmented, that means the prediction map will be showing many small scattered areas of prediction instead of a continuous prediction area. For GLM, it is a algorithm that focuses more on linear relationship, it might not able to process complex and nonlinear environmental responses of the species. Both of the models show signs of distribution in areas that are not suitable for Malayan Partridge, this is more obvious and visible as the prediction shows that the species may be appearing in area of Sabah, Sarawak and other non montane forest areas.

For Bioclim in Figure 5.2.5.1, it provided a good AUC result, however it did not show on the prediction map that Bioclim generated even though it predicted the montane forest area as ver suitable for Malayan Partridge. It shares the same trend with RF and GLM that predicts Sabah and Sarawak as potential distribution areas. This is because Bioclim is a very simple model that takes account only occurrence data to define the evironmental range. This leads the model to overpredict the areas that have similar environmental conditions.

MaxEnt and SVM are two of the best models in terms of predicting the current occurrence of Malayan Partridge. It correctly predicted the current occurrence of Malayan Partridge that located mainly at the area of montane forests. This is because MaxEnt and SVM are models that can handle complex and non-linear relationships very well that generates a smoother and accurate predictions for the species. Therefore, MaxEnt and SVM are two of the best models to predict the occurrence of Malayan Partridge based on the results of AUC and the generated prediction map for the occurrence of Malayan Partridge.

5.4 Prediction for Future Habitat Suitability

This section will show the results of the future habitat suitability map for Malayan Partridge from best models of MaxEnt and SVM. It will predict for the year 2021 to year 2040 and year 2041 to year 2050.

5.4.1 Future Habitat Suitability Map from MaxEnt

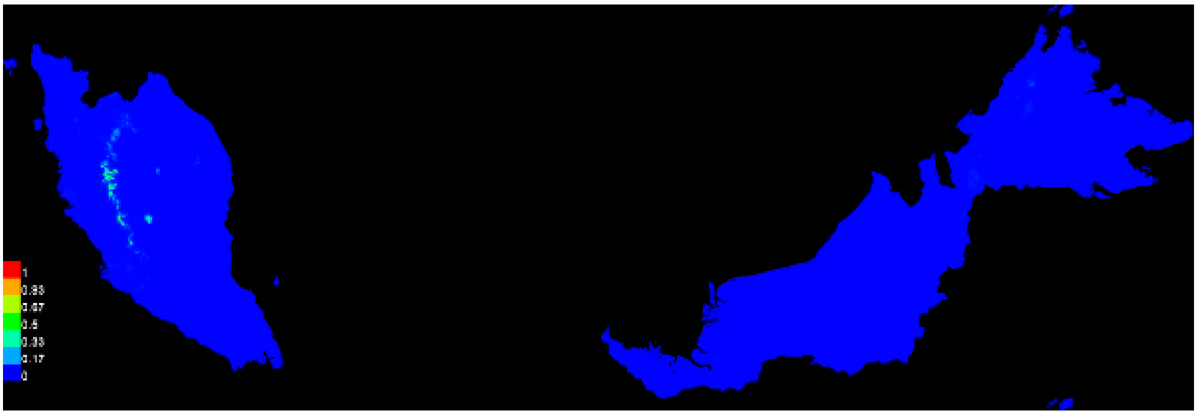


Figure 5.4.1.1 Future Habitat Suitability Map of MaxEnt (Year 2021 to 2040)

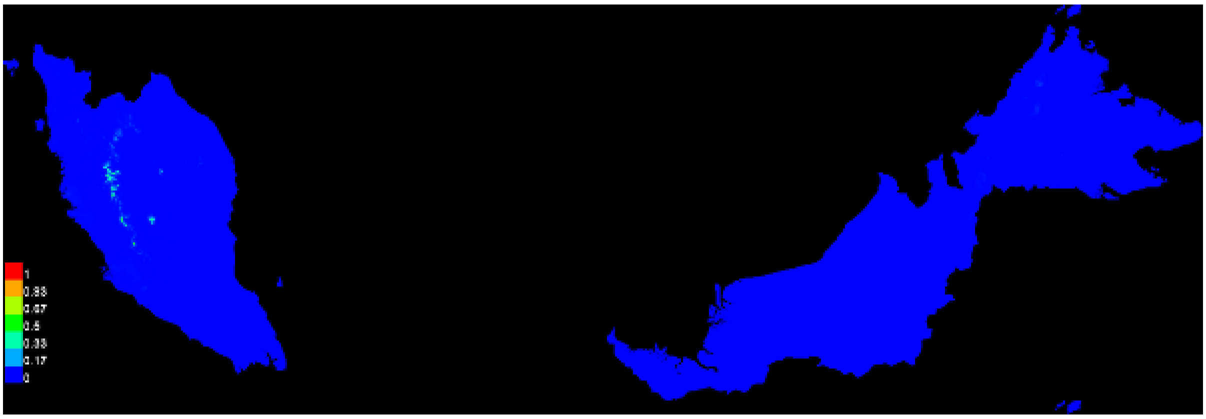


Figure 5.4.1.2 Future Habitat Suitability Map of MaxEnt (Year 2041 to 2060)

5.4.2 Future Habitat Suitability Map from SVM

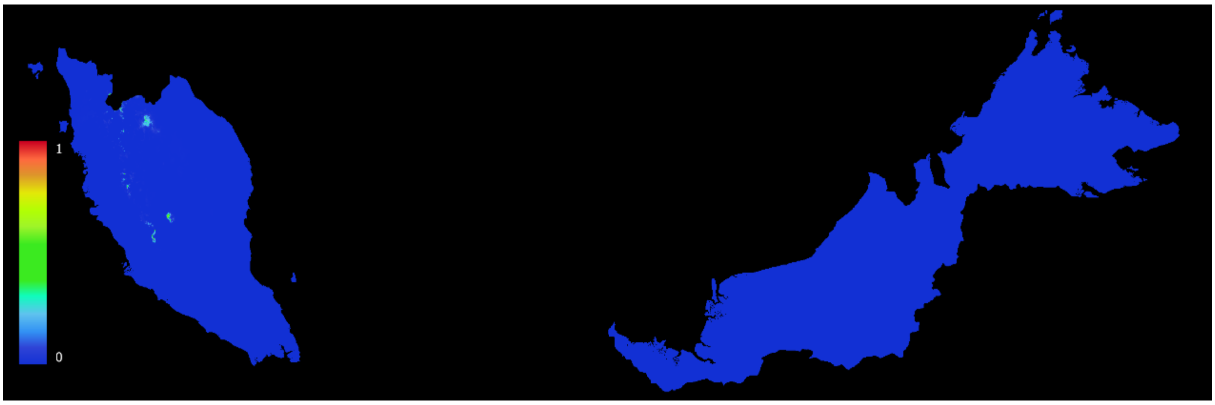


Figure 5.4.2.1 Future Habitat Suitability Map of SVM (Year 2021 to 2040)

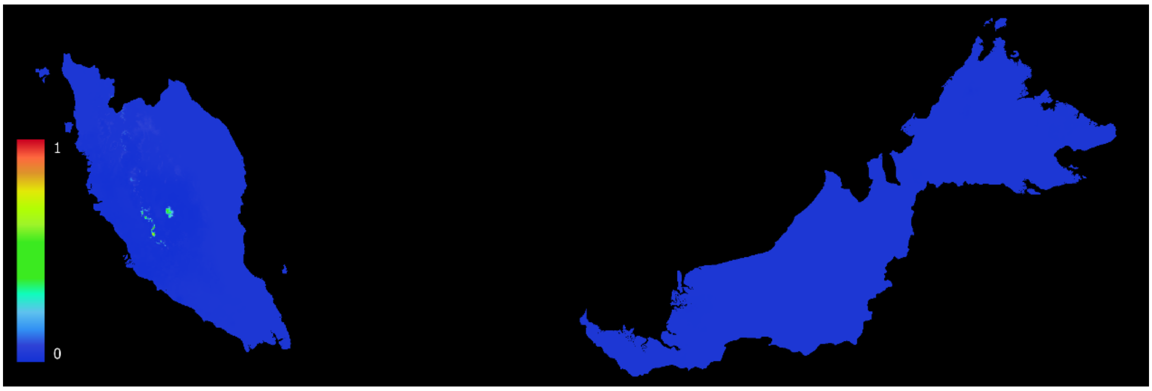


Figure 5.4.2.2 Future Habitat Suitability Map of SVM (Year 2041 to 2060)

5.4.3 Interpretation of Future Habitat Suitability Map

Figures in section 5.4.1 and section 5.4.2 shows the habitat suitability map for the future year of 2021 to 2040 and year 2041 to year 2041. All of the future habitat suitability map are generated using the projection of Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5) as mentioned in section 3.1.1. It is clear that comparing these future habitat suitability map with the current distribution map of Malayan Partridge, there is a huge drop of area of suitable habitat for the species as the year goes by.

This can be explained by the future projections of SSP2-4.5. In this projections, there will be an increase of socioeconomic development, population, economy and technology in Malaysia along with the whole world. With the increase of these activities, it will affect climate conditions such as the temperature and the precipitation patterns in the future.

Overall, the future trend for the species of Malayan Partridge does not look optimistic as there will be less suitable habitat for the species to thrive in. This will no doubt lead to the decline of the species in the future if there are no actions taken to preserve these environments.

5.5 Variable Importance

In the previous part of the variable selection, Figure 4.2.2 shows that bio1 and bio12 is the most important variable concluded with the most votes of three. Therefore, in this part, MaxEnt and SVM will be evaluated by predicting the occurrence map for Malayan Partridge with two of the most important variable. By doing so, it can identified whether that using just 2 bioclimatic variables are enough to make a map distribution result that is comparable to using 6 bioclimatic variables that are implemented in the previous models of MaxEnt and SVM

5.5.1 MaxEnt

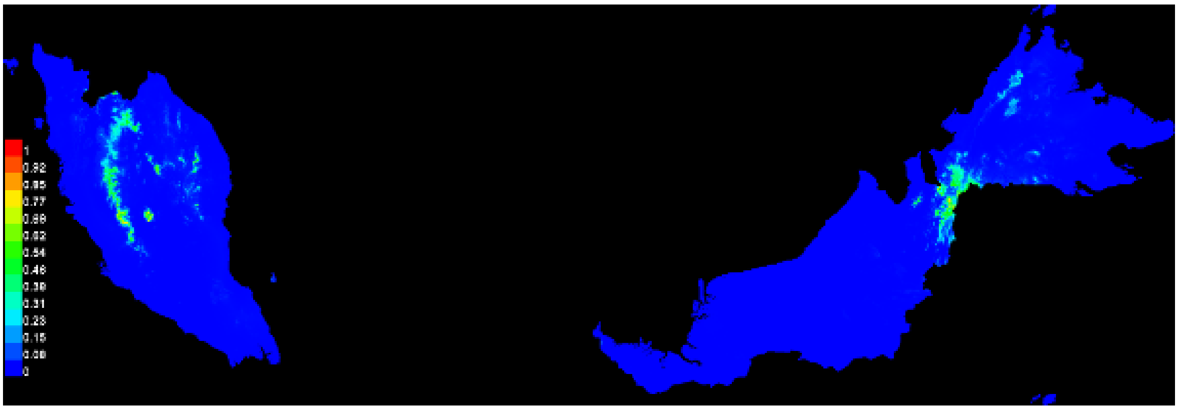
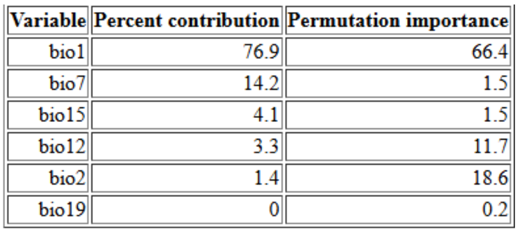


Figure 5.5.1.1 Prediction Map of MaxEnt with 2 variables

For the result of MaxEnt, it shows that using only two variables bio1 and bio12 is not recommended as it generated a distribution map that does not make ecological sense. Since the variable selection process uses RF, permutation importance and XGBoost, it uses a different interpretation to identify the most important variables. MaxEnt itself evaluates variables importance differently. RF and XGBoost identify variables that will affect their own predictive models, while MaxEnt uses permutation importance and percent contributions that take consideration of interactions and environmental variables differently.



Variable	Percent contribution	Permutation importance
bio1	76.9	66.4
bio7	14.2	1.5
bio15	4.1	1.5
bio12	3.3	11.7
bio2	1.4	18.6
bio19	0	0.2

Table 5.5.1.1 Percent Contribution and Permutation Importance of MaxEnt

The Table 5.5.1.1 shows the percent contribution and permutation importance generated by the MaxEnt model. Percent contribution refers to how much each variable contribution to the training of MaxEnt, while permutation importance refers to the how sensetive the model is when the bioclimate variable shuffles its value randomly. Based on the results, show bio1 is the most important bioclimate variable as it has a percent contribution of 76.9% and permutation importance of 66.4%. Even though bio12 and bio2 has a low percent contribution, but it has a moderate permutation importance of 11.7% and 18.6%, showing that the values of bio12 and bio2 is also important for the model prediction. Bio7 is also considered as important as it has 14.2% of contribution to the training of model. Bio15, and bio2 has low percent contribution and permutation importance, but these variables can to be kept as each variable can provide different minor information that will be needed for an accurate prediction. Bio9 has 0 percent contribution and 0.2 permutation importance for the MaxEnt model, but it can provide a little of useful information for the model.

5.5.2 SVM

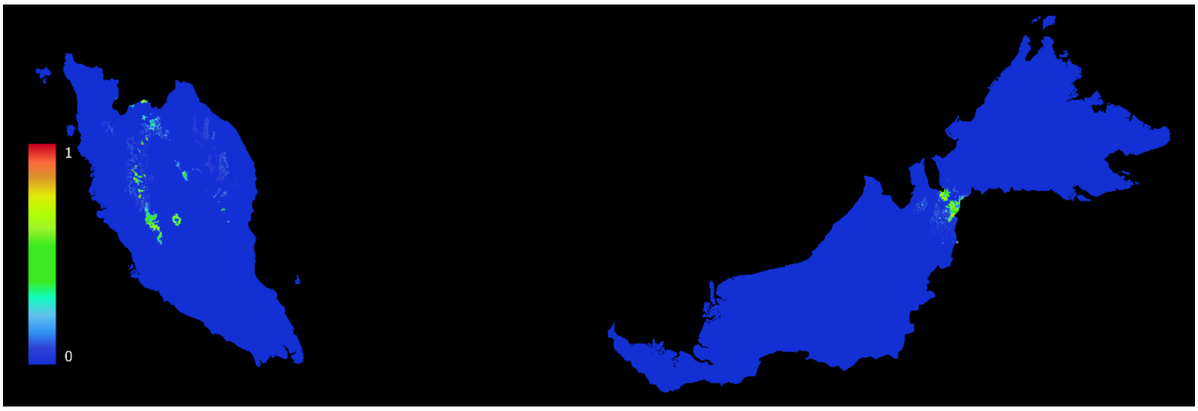


Figure 5.5.2.1 Prediction Map of SVM with 2 variables

Similarly, SVM does not provide an accurate prediction map for the occurrence of Malayan Partridge. The prediction map also shows signs of suitability in the area of Sabah. Even though SVM does not rank variable importance during training, it will use inputs together to separate the presence and absence points in the multidimensional space. Reducing 6 variables to 2 variables will reduce the dimensionality of the results boundary. This will lead the model to have less ability to capture and understand the full information of the speices. To identify what are the important variables for SVM, shapely value can be implemented. Shapely value is similar to the percent contribution and permutation importance from MaxEnt, it will show the variable importance of each variable towards the model.

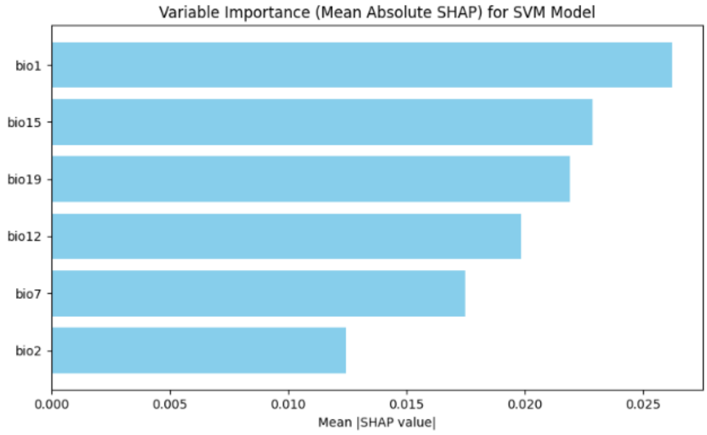


Figure 5.5.2.2 Mean SHAP for SVM

Figure 5.5.2.2 shows the variable importance of 6 of the input bioclimate variables to the model. The result shows that each bioclimate variables provide an importance for the model to produce a robust and reliable result. It provide the SHAP mean value as it will take the result of individual sample points and then average it. Bio1 is considered as the most important variable compare to other variable. However, the remaining variables also shows its importance towards the model performance. Therefore, these 6 variables will considered as the most important variables for the SVM model.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

The study aims to study and predict the occurrence of Malayan Partridge in which help to solves the current challenges of keep tracking of this species in Malaysia. Five different models of SDM MaxEnt, RF, SVM, GLM and Bioclim are being implemented with different pseudoabsence data settings. In the end, the research was able to identify the setting of 5-fold with 1000 pseudoabsence data is the most optimal setting for predicting the occurrence of Malayan Partridge in Malaysia. In this pseudoabsence data setting, the model of MaxEnt and SVM performs the best based on the train test AUC and the distribution map predicted by the models. Future habitat suitability maps were generated using Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5) with the best models and found a similar trend. The future projections show that there will be a massive decrease in suitable habitats for the species in Malaysia.

The study also shows that bio1 is the most important variable among the 6 variables that are chosen for the models to implement. This is proven by the results of MaxEnt and SVM that agreed bio1 is the most important variable in both models. For the 5 other variables, it also plays an important role in producing a reliable model for SDM. This is justified by testing the models to implement the two most important variables based on the variable selection. The result of both models shows distribution maps for Malayan Partridge that are inaccurate and do not make ecological sense. Therefore, bio1, bio2, bio7, bio12, bio15, bio19 are the 6 important variables that required to predict the occurrence of Malayan Partridge accurately in Malaysia. These variables are mainly focused on the temperature and the precipitation patterns referring to Table 4.5.1. These bioclimatic variables reflect the habitat of montane forest of the Malayan Partridge. This highlights the dependence of cool and humid climate of the habitat for Malayan Partridge to thrive in.

6.2 Future Work

In terms of future work, deep learning techniques can be considered as another option of exploration for Species Distribution Model. Deep learning is a technique that is common for other sectors and fields such as finance and business but not for application in SDM. This is because not every species has many presence points as deep learning requires large number of datasets to perform well. However, some deep learning techniques such as Few-Shot and One-Shot can be implemented for SDM. These methods will be useful when there is less knowledge and information about a specific species. For example, zero-shot is a great example that allows models to recognize species that they have never been trained on with related occurrence data. This is particularly useful when there is a need to predict a species that are rare. One example approach proposed to implement CLIP-driven zero-shot species recognition method to recognize species [23]. This can be a promising direction to develop a more robust SDM for the species with small occurrence data.

REFERENCES

[1] J. Brodie, E. Post, and W. F. Laurance, “Climate change and tropical biodiversity: a new focus,” *Trends in Ecology & Evolution*, vol. 27, no. 3, pp. 145–150, Mar. 2012, doi: https://doi.org/10.1016/j.tree.2011.09.008.

[2] M. C. K. Soh, N. S. Sodhi, and S. L. H. Lim, “High sensitivity of montane bird communities to habitat disturbance in Peninsular Malaysia,” Biological Conservation, vol. 129, no. 2, pp. 149–166, Apr. 2006, doi: https://doi.org/10.1016/j.biocon.2005.10.030.

[3] J. Elith and J. R. Leathwick, “Species Distribution Models: Ecological Explanation and Prediction Across Space and Time,” Annual Review of Ecology, Evolution, and Systematics, vol. 40, no. 1, pp. 677–697, Dec. 2023, doi: https://doi.org/10.1146/annurev.ecolsys.110308.120159

[4] D. L. Warren, N. J. Matzke, and T. L. Iglesias, “Evaluating presence-only species distribution models with discrimination accuracy is uninformative for many applications,” *Journal of Biogeography*, vol. 47, no. 1, pp. 167–180, Sep. 2019, doi: https://doi.org/10.1111/jbi.13705.

[5] J. Elith and J. Franklin, “Species Distribution Modeling ☆,” *Reference Module in Life Sciences*, 2017, doi: https://doi.org/10.1016/b978-0-12- 809633-8.02390-6.

[6] K. B. da Silva and A. Nedosyko, “Sea Anemones and Anemonefish: A Match Made in Heaven,” *The Cnidaria, Past, Present and Future*, pp. 425–438, 2016, doi: https://doi.org/10.1007/978-3-319-31305-4_27.

[7] S. I. Higgins, M. J. Larcombe, N. J. Beeton, T. Conradi, and H. Nottebrock, “Predictive ability of a process-based versus a correlative species distribution model,” *Ecology and Evolution*, vol. 10, no. 20, pp. 11043–11054, Oct. 2020, doi:

https://doi.org/10.1002/ece3.6712.

[8] N. C. Coops, R. H. Waring, and T. A. Schroeder, "Combining a generic process-based productivity model and a statistical classification method to predict the presence and absence of tree species in the Pacific Northwest, U.S.A.," *Ecological Modelling*, vol. 220, no. 15, pp. 1787– 1796, Aug. 2009, doi: https://doi.org/10.1016/j.ecolmodel.2009.04.029.

[9] C. F. Dormann et al., "Correlation and process in species distribution models: bridging a dichotomy," Journal of Biogeography, vol. 39, no. 12, pp. 2119–2131, Jan. 2012, doi: https://doi.org/10.1111/j.1365- 2699.2011.02659.x.

[10] C. F. Dormann et al., "Correlation and process in species distribution models: bridging a dichotomy," Journal of Biogeography, vol. 39, no. 12, pp. 2119–2131, Jan. 2012, doi: https://doi.org/10.1111/j.1365- 2699.2011.02659.x.

[11] A. Thapa et al., "Predicting the potential distribution of the endangered red panda across its entire range using MaxEnt modeling," Ecology and Evolution, vol. 8, no. 21, pp. 10542–10554, Oct. 2018, doi: https://doi.org/10.1002/ece3.4526.

[12] S. A. Reilly, "Forecasting the Spread and Invasive Potential of Apple Snails (Pomacea spp.) in Florida," *NSUWorks*, 2017. https://nsuworks.nova.edu/occ_stuetd/460/ (accessed April 05, 2025).

[13] S. Vallecillo, J. Maes, C. Polce, and C. Lavalle, "A habitat quality indicator for common birds in Europe based on species distribution models," Ecological Indicators, vol. 69, pp. 488–499, Oct. 2016, doi: https://doi.org/10.1016/j.ecolind.2016.05.008.

[14] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum entropy modeling of species geographic distributions," Ecological Modelling, vol. 190, no. 3–4, pp. 231–259, Jan. 2006, doi: https://doi.org/10.1016/j.ecolmodel.2005.03.026.

[15] C. Merow, M. J. Smith, and J. A. Silander, "A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter," *Ecography*, vol. 36, no. 10, pp. 1058–

1069, Jun. 2013, doi: https://doi.org/10.1111/j.1600-0587.2013.07872.x.

[16] X. LI and Y. WANG, “Applying various algorithms for species distribution modelling,” *Integrative Zoology*, vol. 8, no. 2, pp. 124–135, Jun. 2013, doi: https://doi.org/10.1111/1749-4877.12000.

[17] P. Mccullagh, “863-9 Advanced School and Conference on Statistics and Applied Probability in Life Sciences,” 2007.

[18] T. W. Yee and N. D. Mitchell, “Generalized additive models in plant ecology,” *Journal of Vegetation Science*, vol. 2, no. 5, pp. 587–602, Oct. 1991, doi: https://doi.org/10.2307/3236170.

[19] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/a:1010933404324.

[20] A. M. Prasad, L. R. Iverson, and A. Liaw, “Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction,” *Ecosystems*, vol. 9, no. 2, pp. 181–199, Mar. 2006, doi: https://doi.org/10.1007/s10021-005-0054-1.

[21] D. P. C. Peters, K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales, “Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology,” *Ecosphere*, vol. 5, no. 6, p. art67, Jun. 2014, doi: https://doi.org/10.1890/es13-00359.1.

[22] W. Rammer and R. Seidl, “Harnessing Deep Learning in Ecology: An Example Predicting Bark Beetle Outbreaks,” *Frontiers in Plant Science*, vol. 10, Oct. 2019, doi: https://doi.org/10.3389/fpls.2019.01327.

[23] L. Liu, B. Han, F. Chen, C. Mou, and F. Xu, “Utilizing Geographical Distribution Statistical Data to Improve Zero-Shot Species Recognition,” *Animals*, vol. 14, no. 12, pp. 1716–1716, Jun. 2024, doi: https://doi.org/10.3390/ani14121716.

[24] Y. Li *et al.*, “Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm,” *OpenReview*, 2022. https://openreview.net/forum?id=zq1iJkNk3uN (accessed Apr. 30, 2025).

[25] H. SÜEL, D. AKDEMİR, E. T. ERTUĞRUL, and S. ÖZDEMİR, “Determining Environmental Factors Affecting Bird Diversity,” *Kastamonu*

Üniversitesi Orman Fakültesi Dergisi, Dec. 2021, doi: https://doi.org/10.17475/kastorman.1049336.

[26] A. A. Shabani, L. C. McArthur, and M. Abdollahian, "Comparing different environmental variables in predictive models of bird distribution," *Russian Journal of Ecology*, vol. 40, no. 7, pp. 537–542, Nov. 2009, doi: https://doi.org/10.1134/s1067413609070133.

[27] B. O. Oindo, R. A. de By, and A. K. Skidmore, "Environmental factors influencing bird species diversity in Kenya," *African Journal of Ecology*, vol. 39, no. 3, pp. 295–302, Sep. 2001, doi: https://doi.org/10.1046/j.1365-2028.2001.00322.x.

[28] A. Ashoori, A. Kafash, H. Varasteh Moradi, M. Yousefi, H. Kamyab, N. Behdarvand & S. Mohammadi (2018) Habitat modeling of the common pheasant Phasianus colchicus (Galliformes: Phasianidae) in a highly modified landscape: application of species distribution models in the study of a poorly documented bird in Iran, The European Zoological Journal, 85:1, 372-380, DOI: 10.1080/24750263.2018.1510994

[29] T. Pegan, E. R. Gulson-Castillo, Alim Biun, J. I. Byington, and F. H. Sheldon, "An assessment of avifauna in a recovering lowland forest at Kinabalu National Park, Malaysian Borneo," *The Raffles Bulletin of Zoology*, vol. 66, pp. 110–131, Jan. 2018, Available: https://www.researchgate.net/publication/323392318_An_assessment_of_avifauna_in_a_recovering_lowland_forest_at_Kinabalu_National_Park_Malaysian_Borneo

[30] B. Li *et al.*, "Threatened birds face new distribution under future climate change on the Qinghai-Tibet Plateau (QTP)," *Ecological Indicators*, vol. 150, pp. 110217–110217, Jun. 2023, doi: https://doi.org/10.1016/j.ecolind.2023.110217.

[31] B. Zahoor, X. Liu, and M. Songer, "The impact of climate change on three indicator Galliformes species in the northern highlands of Pakistan," *Environmental Science and Pollution Research*, vol. 29, no. 36, pp. 54330–54347, Mar. 2022, doi: https://doi.org/10.1007/s11356-022-19631-y.

[32] B. Wang *et al.*, "Climate change affects Galliformes taxonomic, phylogenetic and functional diversity indexes, shifting conservation priority areas in

China," *Diversity and Distributions*, Dec. 2022, doi: https://doi.org/10.1111/ddi.13667.

[33] V. Moudrý *et al.*, "Optimising occurrence data in species distribution models: sample size, positional uncertainty, and sampling bias matter," *Ecography*, Aug. 2024, doi: https://doi.org/10.1111/ecog.07294.

[34] eBird, "eBird: An online database of bird distribution and abundance," Cornell Lab of Ornithology, Ithaca, New York, 2021. [Online]. Available: http://www.ebird.org. [Accessed: Mar. 2, 2025].

[35] WorldClim, "WorldClim version 2: Climate data," [Online]. Available: https://www.worldclim.org/data/worldclim21.html.

[36] Y. Fan, J. Dai, Y. Wei, and J. Liu, "Local Adaptation in Natural Populations of Toona ciliata var. pubescens Is Driven by Precipitation and Temperature: Evidence from Microsatellite Markers," *Forests*, vol. 14, no. 10, p. 1998, Oct. 2023, doi: https://doi.org/10.3390/f14101998.

[37] Z. Hausfather, "CMIP6: The next generation of climate models explained," *Carbon Brief*, Dec. 2, 2019. [Online]. Available: https://www.carbonbrief.org/cmip6-the-next-generation-of-climate-models-explained/

[38] J. N. Stokland, R. Halvorsen, and B. Støa, "Species distribution modelling—Effect of design and sample size of pseudo-absence observations," *Ecological Modelling*, vol. 222, no. 11, pp. 1800–1809, Jun. 2011, doi: https://doi.org/10.1016/j.ecolmodel.2011.02.025.

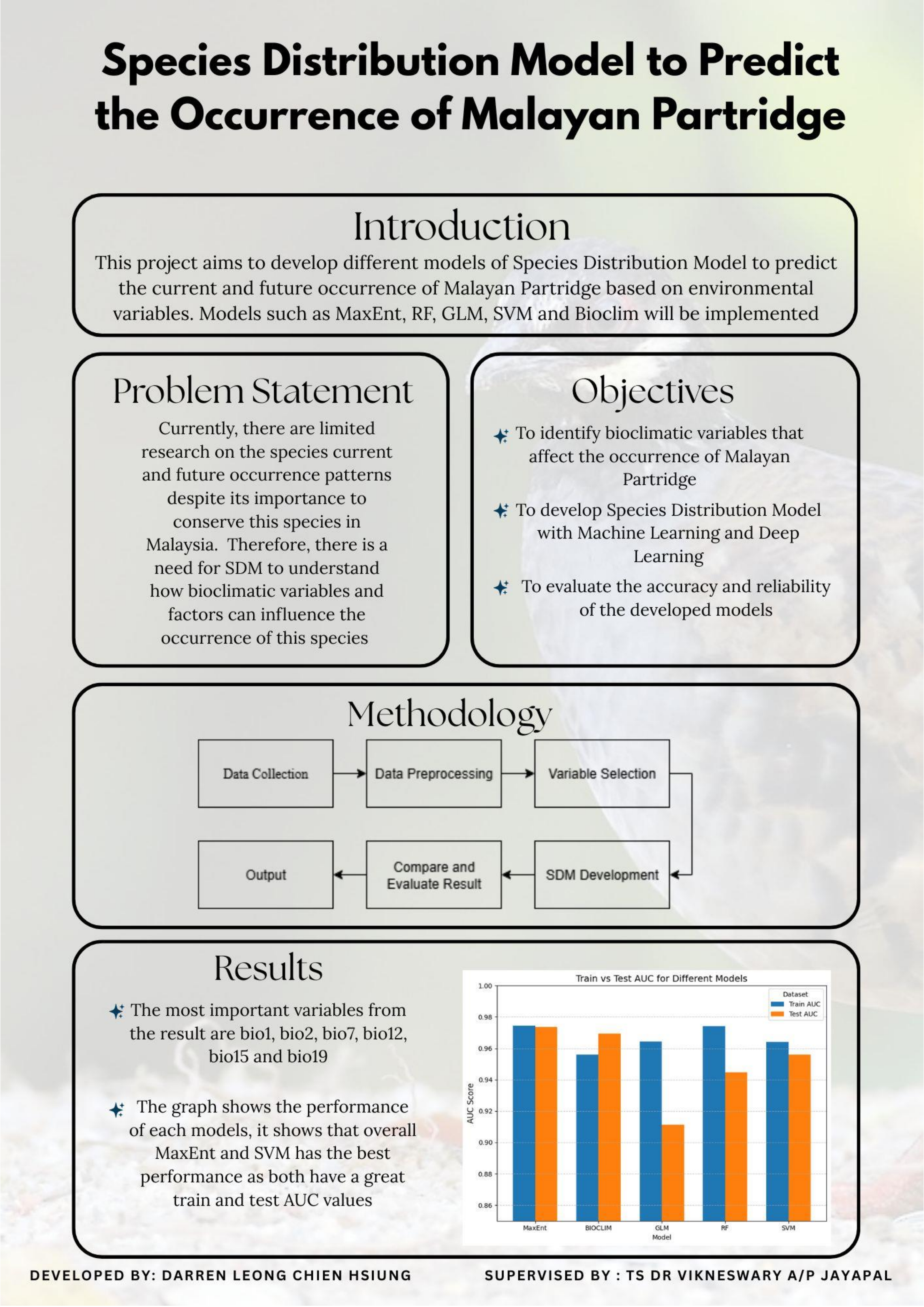
[39] P. Sedgwick, "Pearson's correlation coefficient," *BMJ*, vol. 345, pp. e4483–e4483, Jul. 2012, doi: https://doi.org/10.1136/bmj.e4483.

[40] Q. Guo, M. Kelly, and C. H. Graham, "Support vector machines for predicting distribution of Sudden Oak Death in California," *Ecological Modelling*, vol. 182, no. 1, pp. 75–90, Feb. 2005, doi: https://doi.org/10.1016/j.ecolmodel.2004.07.012.

[41] T. H. Booth, H. A. Nix, J. R. Busby, and M. F. Hutchinson, "bioclim: the first species distribution modelling package, its early applications and relevance to most currentMaxEntstudies," *Diversity and Distributions*, vol. 20, no. 1, pp. 1–9, Nov. 2013, doi: https://doi.org/10.1111/ddi.12144.

Appendix A

POSTER



Species Distribution Model to Predict the Occurrence of Malayan Partridge

Introduction

This project aims to develop different models of Species Distribution Model to predict the current and future occurrence of Malayan Partridge based on environmental variables. Models such as MaxEnt, RF, GLM, SVM and Bioclim will be implemented

Problem Statement

Currently, there are limited research on the species current and future occurrence patterns despite its importance to conserve this species in Malaysia. Therefore, there is a need for SDM to understand how bioclimatic variables and factors can influence the occurrence of this species

Objectives

- To identify bioclimatic variables that affect the occurrence of Malayan Partridge
- To develop Species Distribution Model with Machine Learning and Deep Learning
- To evaluate the accuracy and reliability of the developed models

Methodology

Data Collection → Data Preprocessing → Variable Selection → SDM Development → Compare and Evaluate Result → Output

Results

- The most important variables from the result are bio1, bio2, bio7, bio12, bio15 and bio19
- The graph shows the performance of each models, it shows that overall MaxEnt and SVM has the best performance as both have a great train and test AUC values

DEVELOPED BY: DARREN LEONG CHIEN HSIUNG

SUPERVISED BY : TS DR VIKNESWARY A/P JAYAPAL