ASSESSING BIODIVERSITY LOSS DUE TO ENVIRONMENTAL CHANGES USING ARTIFICIAL INTELLIGENCE TECHNIQUES

By

Chia Cheng Gun

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)

FEBRUARY 2025

COPYRIGHT STATEMENT

This Final Year Project report is submitted in partial fulfillment of the requirements for the degree of Bachelor of Computer Science (Honours) at Universiti Tunku Abdul Rahman (UTAR). This Final Year Project report represents the work of the author, except where due acknowledgment has been made in the text. No part of this Final Year Project report may be reproduced, stored, or transmitted in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author or UTAR, in accordance with UTAR's Intellectual Property Policy.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisor, Ts. Dr. Mogana a/p Vadiveloo, for providing me with this wonderful opportunity to engage in a species distribution modeling (SDM) project. This project marks a significant milestone in my academic journey and my aspirations to apply Artificial Intelligence in ecological modeling. I am deeply grateful for your guidance and support throughout this endeavor.

Finally, I am profoundly grateful to my parents and family for their unwavering love, support, and continuous encouragement, which have been my greatest source of strength during this journey.

ABSTRACT

This study explores the potential of artificial intelligence (AI) techniques to enhance species distribution modelling (SDM) for assessing biodiversity loss due to environmental change, focusing on Strigiformes (owls) in Malaysia, which remain understudied and vulnerable to climate threats. Traditional SDM methods often struggle to capture complex ecological interactions because they rely on linear assumptions. There is a lack of comprehensive studies focused on predicting the future distribution of these species under varying environmental scenarios in Malaysia.To address these issues, this study proposes the use of machine learning and deep learning models, specifically Random Forests (RF) and Multi-Layer Perceptrons (MLP), complemented by Explainable AI (XAI) techniques, to improve predictive accuracy, robustness, and interpretability of SDMs. The models were developed with eight key environmental variables which are annual mean temperature, mean diurnal range, isothermality, annual precipitation, precipitation of wettest month, primary forest, secondary forest and urban area cover for the genus Ketupa (a genus of Strigiformes) in Malaysia. Data splitting techniques, including random and spatial block were evaluated to address spatial autocorrelation and improve model generalization. Spatial block sampling demonstrated superior performance, with smaller performance gaps in Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision Recall curve (AUCPR) when tested on East Malaysia independent dataset, confirming its robustness for extrapolation. Environmental analysis identified urban area cover as the most influential predictor of habitat suitability, followed by annual precipitation. Response curve analysis revealed critical environmental thresholds that align with Ketupa's ecological preferences for tropical lowland and wetland habitats. Habitat suitability mapping under future climate and land-use scenarios indicates a potential loss of high-quality habitat and a flattening of suitability gradients.

Area of Study: Artificial Intelligence in Ecological Modelling

Keywords: Random Forest, Multi-Layer Perceptron, Species Distribution Modelling, Ketupa, Explainable AI (XAI)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
SDM	Species Distribution Modelling
MLP	Multi-Layer Perceptron
RF	Random Forest
SVM	Support Vector Machines
MAXENT	Maximum Entropy
GAM	Generalized Additive Models
MARS	Multivariate Adaptive Regression Splines
XAI	Explainable AI
LIME	Local Interpretable Model-agnostic Explanations
ANN	Artificial Neural Networks
DNN	Deep Neural Networks
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TSS	True Skill Statistic
CV	Coefficient of Variability
AUROC	Area Under the Receiver Operating Characteristic Curve
GBIF	Global Biodiversity Information Facility
IPCC	Intergovernmental Panel on Climate Change
SSP	Shared Socioeconomic Pathways
MAHAL	Mahalanobis Distance
GLM	Generalized Linear Models
CNN	Convolutional Neural Networks
CV	Coefficient of Variation
SSP	Shared Socioeconomic Pathways
LUH2	Land-Use Harmonization 2
SHAP	Shapley Additive Explanations
FS-SINR	Few-Shot Spatial Implicit Neural Representation

CHAPTER 1 INTRODUCTION

Environmental changes , such as climate change, has significantly impacted our world in recent decades, causing extensive disruptions to ecosystems and threatening biodiversity [1]. The Intergovernmental Panel on Climate Change (IPCC) reports that global temperatures have risen by approximately 1.0°C above pre-industrial levels, primarily due to human activities [2].

Environmental change has profoundly impacted various species in Malaysia, including Strigiformes (owls). Among the Strigiformes, the genus Ketupa is a powerful birds of prey that function as predators in freshwater ecosystems and serve as important indicators of healthy ecosystems [3–5]. In Malaysia, both Ketupa zeylonensis and Ketupa ketupu are classified as species of "Least Concern" by the IUCN [6]. However, Ketupa zeylonensis faces significant threats in regions such as Turkey and parts of Europe, where habitat loss and dam construction have critically impacted its population [4]. A 2015 study [7] revealed that hydroelectric reservoirs in Sarawak inundate habitats for 331 bird species, 164 mammal species, 2,100 tree species and 17,700 arthropod species, with 4–7 arthropod species extinctions predicted. While extinctions of birds, mammals, and trees are not anticipated, the ecological consequences remain substantial. Additionally, the Sarawak government recently announced the construction of dams on Sungai Gaat, Sungai Tutoh, and Sungai Belaga. This highlights the vulnerability of species under the global climate change and environmental change [8].

As global temperatures continue to rise, the habitats and survival of these species are increasingly threatened. Evidence suggests that for every degree Celsius increase in temperature, a significant number of species face heightened risk of extinction [9]. The shifting climate can lead to habitat loss, changes in food availability, and altered predator-prey dynamics, all of which can create cascading effects on species and the broader ecosystems they sustain.

In this context, species distribution modelling (SDM) has emerged as a critical tool for understanding how species respond to changing environmental conditions. SDMs allow researchers to predict how species distributions might shift in response to environmental change, helping to identify areas that are likely to remain suitable for species survival in the future [10, 11]. These models provide valuable insights that can

inform conservation strategies, such as the establishment of protected areas and the development of policies aimed at preserving biodiversity [11].

1.1 Problem Statement and Motivation

Species distribution modelling (SDM) is an established technique for predicting species distributions. However, conventional SDM models often have limitations in their ability to capture complex ecological relationships and interactions due to their reliance on linear assumptions and the assumption of independence among predictors [12]. In recent years, with the advancement of technology, artificial intelligence (AI) techniques have been introduced to enhance the accuracy and predictive power of SDMs, including machine learning models such as Random Forests (RFs), Support Vector Machines (SVM), and neural networks. They offer greater flexibility and can model complex non-linear relationships and interactions among environmental variables [13].

Strigiformes have shown significant reactions to environmental change, which is altering their habitats, food availability, and reproductive cycles, leading to shifts in population dynamics and species distribution [14]. Despite these emerging threats, there is a lack of comprehensive studies focused on predicting the future distribution of these species under varying environmental scenarios in Malaysia. This study aims to address this gap by focusing on Strigiformes, using AI-based species distribution modelling (SDM) techniques to predict how these species will be affected by ongoing and future environmental changes.

According to [15], although SDM have been widely applied across various regions, environmental conditions and ecological dynamics can differ significantly from one country to another. This makes it crucial for each country to conduct its own studies, even if similar research has been done elsewhere. There are substantial gaps in SDM applications across different biological groups and regions, particularly in highly biodiverse areas like many Asian territories, including Malaysia.

This motivation drives this study to explore SDM for species in Malaysia, focusing specifically on Strigiformes. Malaysia's unique climate and diverse ecosystems present distinct challenges and opportunities, underscoring the importance of localized research to account for the specific environmental variables and species interactions present in the region.

Moreover, accuracy and explainability are crucial in SDM, particularly when the findings are used to inform conservation strategies and policy decisions [16]. AI techniques, which are excel at handling multi-variate problems, offer significant advantages in this context [16–18]. By employing AI-based SDM, higher accuracy and better explainability can be achieved.

This study aims to leverage AI techniques to develop robust SDM models that can accurately predict the distributions of species within the Strigiformes under various environmental scenarios in Malaysia. Enhanced accuracy and explainability will empower decision-makers with reliable data, enabling them to make informed decisions to protect Malaysia's biodiversity effectively.

1.2 Objectives

The primary objective of this study is to develop a robust SDM framework that effectively addresses the limitations identified in current SDM research, particularly in the context of tropical regions like Malaysia. Specifically, this study will focus on modelling the distribution of Strigiformes (Ketupa) within Malaysia's unique environmental conditions. By doing so, it aims to fill the existing research gap and contribute to a more comprehensive understanding of how environmental change impacts species in underrepresented tropical regions.

To achieve the main objective, the study embarks the following sub-objectives:

1. To propose the development and comparison of Random Forest (RF) and Multi-Layer Perceptrons (MLP) for species distribution modelling.
   The objective is to identify the most suitable model by evaluating predictive performance under the specific challenges of the study, including a relatively small dataset size and Malaysia's unique tropical geographical conditions. This approach seeks to ensure accurate and robust predictions for species distributions under current and future environmental scenarios.
2. To enhance the interpretability of the predictive models.
   As models become more complex, they often become less interpretable. Therefore, this study aims to integrate Explainable AI (XAI) techniques, such as Shapley Additive Explanations (SHAP), to provide clear insights into how specific environmental variables influence species distributions in different regions of

Malaysia. By incorporating XAI, the study seeks not only to improve the model's predictive accuracy but also to ensure that the predictions are interpretable and actionable.

3. To evaluate the effectiveness of data splitting techniques, specifically random sampling and spatial block sampling techniques in species distribution modelling.
   Random sampling is widely used for its simplicity and ability to provide a randomized distribution of training and testing data, but it may lead to overestimated model performance due to spatial dependence in the data. In contrast, spatial block sampling reduces spatial autocorrelation by dividing the study area into distinct spatial blocks, ensuring that training and testing data are spatially independent. This study aims to compare these sampling techniques to determine which is better suited for modeling species distributions under diverse climatic scenarios.
4. To assess the impact of environmental change on species distributions.
   Probabilistic mapping is visualization tools used for capturing the likelihood of species presence and shifts in distribution under changing environmental conditions. This study will apply probabilistic mapping into species distribution models (SDMs) for evaluating both current distribution and forecast potential shifts of Strigiformes (Ketupa) in Malaysia. This integration will allow for a more accurate and actionable understanding of how species in tropical regions may respond to current and future environmental changes.

1.3 Project Scope and Direction

1. This research aims to develop a species distribution modelling (SDM) framework to assess the impact of environmental change on biodiversity in Malaysia, with a particular emphasis on Ketupa (a genus within Strigiformes). By utilizing occurrence data from the Global Biodiversity Information Facility [19],climatic variables such as temperature and precipitation from WorldClim [20], terrain attributes such as slope derived from digital elevation models [21] and land-use data from the Land-Use Harmonization 2 (LUH2) dataset [22]. This study will concentrate on the primary environmental factors influencing species distribution changes as a consequence of environmental change.
2. Moreover, by using the Python and R programming language, the project will implement various artificial intelligence techniques to develop a robust SDM framework.
3. This framework will not only forecast the future population distributions of Strigiformes (Ketupa) in Malaysia but also identify the contribution of environmental variables to these predictions.

1.4 Contributions

In this research work, the performance of various AI models shall be compared to identify the most effective approach for species distribution modelling (SDM). Additionally, the research will analyze the contribution of different environmental variables to the distributions of genus Ketupa, offering a detailed understanding of which environmental factors are most influential. The predictive results generated by the proposed models will support informed decision-making in conservation strategies, helping to mitigate the adverse effects of environmental change on biodiversity in Malaysia.

1.5 Report Organization

This report is organized into 6 chapters: Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 System Model, Chapter 4 System Implementation and Experiment Results, Chapter 5 Evaluation and Discussion, Chapter 6 Conclusion and Future Works. The first chapter detailed the introduction of this project which includes problem statement, project background and motivation, project scope, project objectives, project contribution and report organization. The second chapter is the literature review carried out on several existing studies in ecological modelling. The third chapter discusses the overall research methodology. The fourth chapter is regarding the details on how to implement the system pipeline. Furthermore, the fifth chapter reports the evaluation and discussion on the result of experiments. Lastly, the sixth chapter makes a conclusion and discussion on future work for this study.

CHAPTER 2 LITERATURE REVIEW

2.1 Previous Works

Statistical models have long been essential in species distribution modelling (SDM), offering a balance between simplicity and insight. Unlike more complex machine learning models, statistical approaches prioritize interpretability and explicitly define relationships between variables. These models typically assume a structured form, either linear or non-linear, and often incorporate domain-specific knowledge to guide the selection and transformation of predictor variables [23]. Statistical models are particularly valued for their ability to quantify the effects of individual environmental factors on species distributions, providing clear, interpretable outputs that can be directly linked to ecological theory. This makes them powerful tools for hypothesis testing and for understanding the underlying mechanisms driving species distributions, rather than just focusing on predictive accuracy.

2.1.1 Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions

The study [24] utilized presence-absence data for 15 freshwater fish species, which were modeled against 16 environmental predictors to evaluate the performance of five different model sets. It included individual Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS), both with and without interactions, as well as a multi-response MARS model.

GAM fitting in the study initially began with an attempt to use all predictor variables as smooth terms, followed by a backwards and forwards stepwise procedure to remove non-significant terms. However, due to the large dataset size, this approach proved slow and memory-intensive, making it difficult to compare the significance of linear versus smooth terms. As an alternative, the authors employed the BRUTO algorithm, which fits a GAM using an adaptive back-fitting procedure. This method reduced computation time and allowed for more rigorous performance assessment, ultimately showing that BRUTO/GAM models outperformed full stepwise GAM models.

On the other hand, MARS adjusts for fitting by using a process that involves fitting piecewise linear basis functions to model relationships between the response variable and the predictors. These basis functions are defined by selecting "knots" within the

range of each predictor variable, where the slope of the relationship can change. The selection of these knots and the corresponding basis functions is done automatically through a forward stepwise process, which aims to minimize the residual sum of squares. Once the model reaches a specified maximum size, a backward-pruning procedure is applied to remove basis functions that contribute the least to the model's fit, potentially eliminating entire predictors that do not significantly improve the model's performance.

The evaluation of the models focused on their ability to explain deviance and their discriminatory power by ROC area. According to Table 2.1, GAM models generally explained about 7% more deviance than non-interaction MARS models, suggesting a slight advantage for GAM in capturing complex species-environment relationships. Both GAM and MARS models performed similarly in terms of discriminatory power. This indicates that both methods are equally effective in distinguishing between sites where species are present or absent.

Table 2.1 Summary of GAM and MARS models [24]

Model	Deviance explained	Variables retained	ROC^{train}	ROC^{boot}
GAM individual	1505	13.2	0.863	0.847 (0.013)
MARS individual—non-interaction	1409	9.4	0.853	0.839 (0.016)
MARS multiresponse—non-interaction	1410	12	0.854	0.842 (0.016)
MARS individual with interactions	1541	9.7	0.861	0.838 (0.024)
MARS multiresponse with interactions	1473	10	0.859	0.845 (0.023)

Table values indicate: the average amount of deviance explained; the average number of predictor variables retained in the final models; area under the receiver operator characteristic curve statistics (ROC) averaged across 15 species and calculated using the training data (ROC^{train}); and ROC scores calculated using bootstrap re-sampling (ROC^{boot}) to assess performance when predicting to independent sites, with standard errors shown in brackets.

Both GAM and MARS models are capable of evaluating the contribution of individual environmental predictors to species distributions. In GAM, this is achieved through the flexibility of the smoothing functions applied to the predictor variables. The significance of each predictor can be assessed by observing changes in deviance when the predictor is added or removed from the model, as well as by examining the smooth curves that depict the relationship between the predictor and species presence.

In contrast, MARS evaluates the contribution of predictors by fitting piecewise linear basis functions that define the relationship between the response variable and the predictors. These basis functions enable MARS to model interactions and non-linear relationships in a more segmented and interpretable manner. During the model-building process, MARS selects the most relevant predictors by retaining only those basis

functions that significantly reduce the residual sum of squares, while less important predictors are pruned during the backward elimination process. Table 2.2 identified a small set of dominant predictors, such as stream flow (SegFlow), summer air temperature (SegJanT), distance to the coast (DSDist), and catchment slope (USSlope), which were consistently significant across both GAM and MARS models, though their relative importance varied slightly between the two methods.

Table 2.2 Summary of contributions of predictors to models [24]

	GAM		MARS individual		MARS multiresponse		Average	
	Count	Δ-dev.	Count	Δ-dev.	Count	Δ-dev.	Δ-dev.	Rank
DSDist	15	139.8	14	177.9	15	136.7	151.4	1
SegJanT	15	106.2	14	140.8	15	126.4	124.5	2
SegFlow	14	66.2	13	108.2	15	69.5	81.3	3
USSlope	15	67.9	13	84.7	15	77.0	76.5	4
USRainDays	15	47.9	12	63.9	15	43.3	51.7	5
DSAveSlope	10	29.3	10	36.6	15	31.3	32.4	6
DSMaxSlope	15	31.2	11	30.2	15	30.5	30.6	7
SegTSeas	13	29.7	13	33.4	15	27.2	30.1	8
SegShade	14	28.0	9	29.8	15	25.1	27.6	9
USIndigForest	13	15.2	9	13.4	15	16.1	14.9	10
USPhos	9	14.1	4	8.1	15	14.6	12.3	11
USLake	8	10.6	2	4.6	15	11.3	8.9	12
USCalc	9	14.2	3	3.5	0	0.0	5.9	13
USHard	12	14.9	4	2.5	0	0.0	5.8	14
SegSlope	10	7.6	5	7.1	0	0.0	4.9	15
USPeat	11	7.2	5	3.0	0	0.0	3.4	16

Table entries indicate both the number of models for which each variable was retained as a significant predictor (Count), and the mean change in residual deviance when dropping that variable from final models (Δ-dev.). The two right-hand columns indicate changes in deviance averaged across all three modelling techniques, and their ranking, based on this average. Assessment of the contribution of environmental variables to MARS models fitted using interactions was not attempted.

2.1.2 Advantages and Disadvantages of Statistical Model

Overall, the advantages of conventional models in species distribution, such as statistical models, have been highlighted in [24]. These models offer significant benefits due to their ease of interpretation and implementation. They are straightforward to understand, making it easier for researchers and decision-makers to interpret the relationships between species occurrences and environmental variables. Additionally, statistical models are well-suited for smaller datasets, as they do not require the extensive data that more complex models might need to perform effectively. Their simplicity also facilitates quick implementation, allowing for more immediate application in ecological studies and environmental management.

However, according to [25], statistical models have notable limitations, including inflexibility and a lack of interaction handling. These models often assume linear or simple non-linear relationships, which can oversimplify the true nature of ecological

interactions. This rigidity makes it difficult for statistical models to accurately capture intricate, non-linear patterns or the interactions between multiple environmental variables. As a result, they may fail to account for the complexity inherent in ecological systems, leading to less accurate predictions and potentially misleading conclusions when the underlying relationships are more complex than the model structure allows.

2.2 Related Works

Recent advancements in species distribution modelling (SDM) have increasingly relied on machine learning and deep learning techniques, which offer substantial improvements over conventional statistical models [13]. Machine learning methods such as Random Forest (RF), Support Vector Machines (SVM), and Maximum Entropy (MAXENT) provide greater flexibility and accuracy by effectively capturing non-linear relationships and interactions among environmental variables. Deep learning approaches, including Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs), further enhance predictive capabilities by leveraging neural networks to model high-dimensional data. They excel at handling large datasets and uncovering complex patterns that traditional models often miss. As a result, these advanced techniques have become essential in SDM, providing more reliable predictions and deeper insights into species-environment relationships.

2.2.1 The predictive performance and stability of six species distribution models

According to [26], the study assesses the predictive performance and stability of six widely used SDMs: BIOCLIM, DOMAIN, Mahalanobis distance (MAHAL), RF, MAXENT, and SVM. The focus is on evaluating how well these models predict the potential distribution areas for five common tree species in China.

The study uses presence-only distribution data for one coniferous species, Pinus massoniana, and four broad-leaf species, Betula platyphylla, Quercus wutaishanica, Quercus mongolica, and Quercus variabilis. These species were selected as test subjects from the Eco-Environmental Sciences Research Center and Ecosystems and Ecosystem Service Zoning in China. The data were rasterized at a spatial resolution of five arc-minutes.

The study uses 26 ecological-environmental variables, which include 19 bio-climatic factors, three human disturbance factors, and three soil factors. The data for the bio-

climatic factors were extracted from the Global Climate Data, representing the period 1950-2000. Human disturbance factors were obtained from the Center for International Earth Science Information Network, and soil factors were sourced from the Atlas of the Biosphere.

In order to avoid overfitting and improve the model performance, the study calculates Pearson's correlation coefficients between pairs of variables. After this preprocessing step, 13 final environmental variables are selected for modelling which are shown in Table 2.3. For SDMs requiring presence/absence data, 500 pseudo-absence points are randomly generated across China, excluding known presence points. The dataset is split into a training set (80%) and a test set (20%). This process is repeated 100 times to evaluate model stability.

Table 2.3 26 environment variables [26]

Variable	Symbol
Annual mean temperature (°C)[a,b]	Bio1
Mean diurnal range (Mean of monthly (max temp - min temp)) (°C)[a,b]	Bio2
Isothermality (×100)[b]	Bio3
Temperature seasonality (standard deviation×100) (°C)[a,b]	Bio4
Max temperature of warmest month (°C)[b]	Bio5
Min temperature of coldest month (°C)[b]	Bio6
Temperature annual range (°C)[b]	Bio7
Mean temperature of wettest quarter (°C)[a,b]	Bio8
Mean temperature of driest quarter (°C)[b]	Bio9
Mean temperature of warmest quarter (°C)[b]	Bio10
Mean temperature of coldest of quarter (°C)[b]	Bio11
Annual precipitation (mm)[a,b]	Bio12
Precipitation of wettest month (mm)[b]	Bio13
Precipitation of driest month (mm)[b]	Bio14
Precipitation seasonality (coefficient of variation) (mm)[a,b]	Bio15
Precipitation of wettest quarter (mm)[b]	Bio16
Precipitation of driest quarter (mm)[b]	Bio17
Precipitation of warmest quarter (mm)[b]	Bio18
Precipitation of coldest quarter (mm)[a,b]	Bio19
Human footprint[a,c]	HF
Human influence index[c]	HII
Human population density in year 2000 (persons/km^2)[a,c]	HPD
Soil organic carbon density (kg/m^2 at 1 m depth)[a,d]	SOC
Soil pH value[a,d]	SPH
Soil moisture index[a,d]	SMI
Altitude (m)[a,b]	ALT

[a]Variables used in modeling.
[b]See http://www.worldclim.org/.
[c]See http://sedac.ciesin.columbia.edu/.
[d]See http://www.sage.wisc.edu/atlas/maps.php.
Human footprint (HF) is based on the premise that the impact of human influence varies by biogeography and HF expresses as a percentage the relative human influence in every biome on the land's surface.
Human influence index (HII) is a measure showing direct human influence on ecosystems using eight measures of human presence (population density/km^2, score of railroads, score of major roads, score of navigable rivers, score of coastlines, score of nighttime stable lights values, urban polygons, and land cover categories).
Soil moisture index (SMI) reflects the ability of soil to supply moisture to plants and SMI can identify a quick onset of drought by demonstrating the observed dryness of a soil relative to the plant's ability to extract water as scaled over the range from field capacity to wilting point.
doi:10.1371/journal.pone.0112764.t001

Table 2.4 provides a concise comparison of six SDMs used in the study. It summarizes each model's methodology, key advantages, and disadvantages.

Table 2.4 Comparison of Six SDMs

Model	Description	Advantages	Disadvantages
BIOCLIM	Uses percentile distribution of climatic variables to assess suitability	• Simple to understand and implement	• Ignores variable interactions
		• Handles climatic suitability well	• May oversimplify niches
DOMAIN	Uses Gower distance to assign habitat suitability based on proximity to known occurrences	• Considers environmental similarity	• Threshold sensitivity
		• Flexible threshold setting	• Limited handling of complex interactions
MAHAL	Uses Mahalanobis distance to rank sites based on environmental correlations	• Accounts for variable correlations	• Risk of overestimating suitability
		• Effective area identification	• Careful interpretation needed
RF	Ensemble of decision trees using random subsets of data	• Handles complex interactions	• Computationally intensive
		• Robust to overfitting	• Complex interpretation
MAXENT	Maximum entropy modelling assuming uniform species spread	• Performs well with limited data	• Sensitive to regularization
		• Produces continuous suitability maps	• Potential overfitting without tuning
SVM	Finds optimal hyperplane, handles non-linear relationships via kernel	• Effective in high-dimensional spaces	• Computationally demanding
		• Robust with proper regularization	• Sensitive to kernel and parameters

Each model was implemented using the same 13 environmental variables, following the standard procedures specific to each modelling technique. The predictive performance was evaluated using the Kappa statistic and area under the curve (AUC) values, which are widely accepted metrics for assessing model accuracy.

In Table 2.5, the models MAHAL, RF, MAXENT, and SVM consistently outperformed BIOCLIM and DOMAIN in both mean AUC and Kappa values. BIOCLIM and DOMAIN also had significantly higher standard deviations, indicating more variability and less consistent predictions, whereas MAHAL, RF, MAXENT, and SVM showed lower standard deviations, reflecting greater consistency. Figure 2.1 further illustrates that BIOCLIM and DOMAIN had higher coefficients of variability (CV) for AUC and Kappa values, indicating greater variation in performance compared to the more consistent MAHAL, RF, MAXENT, and SVM models.

Table 2.5 Mean Value and Confidence Interval of AUC and Kappa [20]

	AUC (Mean±SD)	Kappa (Mean±SD)	Confidence interval of AUC (99% confidence level)	Confidence interval of Kappa (99% confidence level)
BIOCLIM	0.945±0.019 b	0.850±0.037 b	0.940−0.950	0.840−0.859
DOMAIN	0.956±0.014 b	0.829±0.039 b	0.953−0.960	0.819−0.839
MAHAL	0.971±0.012 a	0.887±0.033 a	0.968−0.974	0.879−0.895
RF	0.976±0.010 a	0.902±0.030 a	0.973−0.978	0.894−0.910
MAXENT	0.975±0.010 a	0.889±0.031 a	0.972−0.977	0.881−0.897
SVM	0.970±0.012 a	0.891±0.031 a	0.967−0.973	0.883−0.899

Means with different letters differ significantly among the six SDMs (BIOCLIM, DOMAIN, MAHAL, RF, MAXENT, and SVM). SD is the abbreviation for standard deviation.
doi:10.1371/journal.pone.0112764.t002

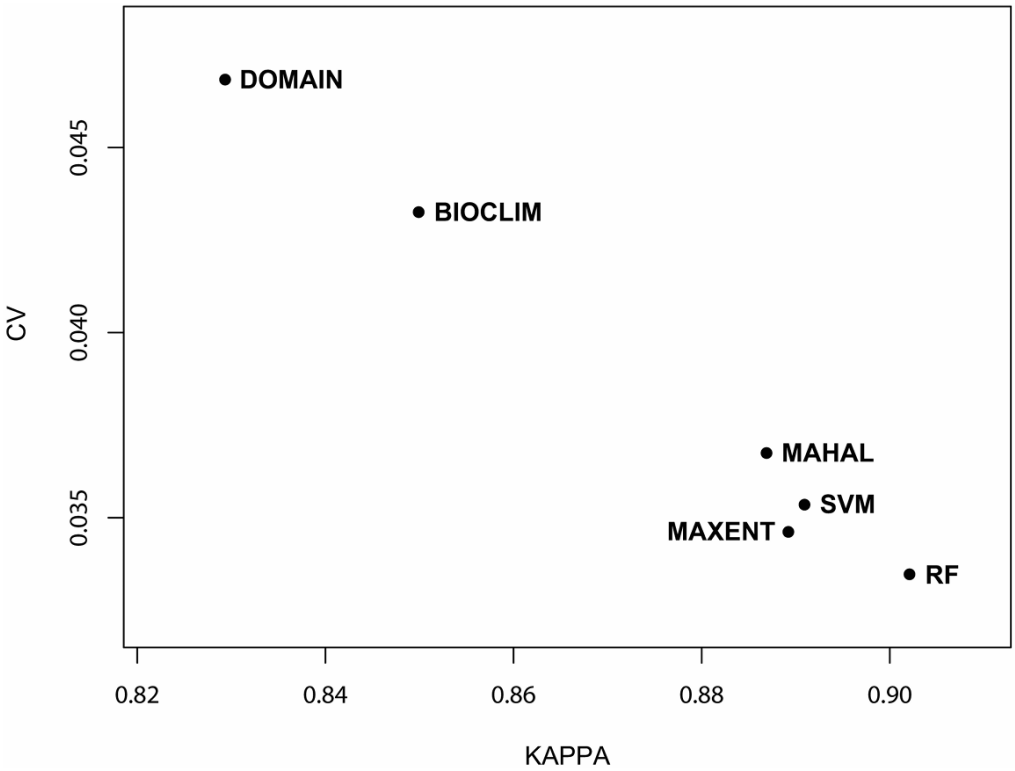


Figure 2.1 The variable coefficient (CV) of Kappa for six SDMs [26]

The study's key contribution is its comprehensive evaluation, showing that advanced machine learning models like RF and MAXENT offer superior predictive performance and stability, making them ideal for modelling complex ecological systems. On the other hand, the study also highlights the value of simpler models like BIOCLIM and DOMAIN. While they are less accurate, they are highly interpretable and easy to use, making them valuable for applications where simplicity and ease of implementation are important. Overall, the study underscores the importance of selecting SDMs based on the specific needs of the research, balancing the trade-offs between model complexity, accuracy, and interpretability.

2.2.2 Exploring the potential of neural networks for species distribution modeling

The study [27] explores the use of neural networks, specifically Multi-Layer Perceptrons (MLPs) in SDM. Traditionally, SDMs have relied on statistical models like Generalized Linear Models (GLMs), Generalized Additive Models (GAMs), and machine learning methods such as RF and MAXENT. However, with advancements in deep learning, there is growing interest in exploring neural networks as a more flexible and potentially more powerful approach. This study compares the performance of

MLPs with these established methods, focusing on both single-species and multi-species modelling to assess whether neural networks can deliver comparable or superior results in predicting species distributions.

The dataset used in this study includes occurrence data for 225 species across six geographically diverse regions: Australian Wet Tropic (AWT), Canada (CAN), New South Wales (NSW), New Zealand (NZ), South America (SA), and Switzerland (SWI). Each region has 11 to 13 environmental covariates, including climatic and pedological variables, essential for modelling species distributions. The environmental covariates, which include both continuous and categorical variables, undergo normalization to ensure that the inputs to the neural network are on a similar scale. Categorical variables, where present, are one-hot encoded before being input into the MLP models.

This study focuses on applying MLPs in SDM, designed for both single-species and multi-species modelling. The architecture consists of multiple layers, each including a fully connected layer, followed by batch normalization, the sigmoid linear unit (SiLU) activation function, and dropout for regularization. For the single-species model, the output layer contains a single neuron, while the multi-species model expands the output layer to accommodate predictions for multiple species simultaneously. This multi-species approach is particularly advantageous for species with limited occurrence data, as it allows the model to leverage co-occurring environmental patterns from related species. The architecture of the MLP models is shown in Figure 2.2.

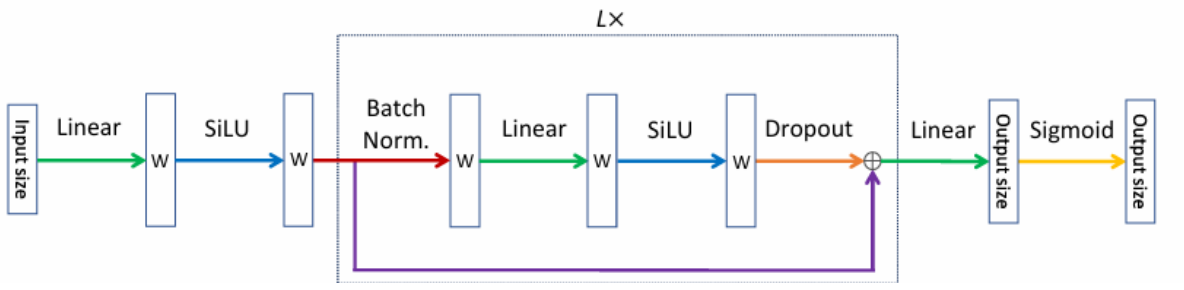


Figure 2.2 Architecture Diagram of MLP model [27]

The hyperparameters of the MLP models were fine-tuned using the Optuna library, which automates the search process and identifies the best-performing configurations. For each model, the authors conducted 50 iterations of the hyperparameter search, adjusting parameters such as the number of layers, width of the MLP, learning rate,

weight decay, and dropout rate. As referred in Table 2.6, the best hyperparameters were selected based on performance on a validation set.

Table 2.6 Best hyperparameters for multi-species model [27]

	Range	AWT	CAN	NSW	NZ	SA	SWI
#Layers	{2,. . . ,6}	4	5	4	5	6	5
MLP width	{256,. . . ,2048}	1327	285	456	465	1140	420
Learning rate	[1e-5, 1e-2]	4e-3	1e-4	2e-5	2e-3	2e-5	4e-4
Weight decay	[1e-5, 1e-2]	3e-3	8e-5	2e-3	9e-4	8e-5	9e-3
Dropout	[0.0, 0.3]	0.17	0.018	0.15	0.28	0.04	0.018

The evaluation of the MLP models was conducted using the Area Under the Receiver Operating Characteristic curve (AUROC). Table 2.7 compares the MLP models with established SDM methods. The authors found that the MLP models achieved comparable AUROC scores with state-of-the-art SDM methods, indicating that neural networks can perform at a level similar to or slightly better than traditional methods, particularly when trained on multiple species simultaneously.

Table 2.7 Mean AUROC for MLP and state-of-the-art models [27]

	AWT	CAN	NSW	NZ	SA	SWI
MaxEnt	**0.686**	0.584	0.713	0.738	0.804	0.809
XGBoost	0.653	0.568	0.706	0.720	0.788	0.815
Random Forest	0.675	0.572	0.718	0.746	**0.813**	**0.818**
Ensemble	0.683	0.580	**0.723**	**0.749**	0.806	0.812
Single-species MLP	0.666	0.589	0.688	0.715	0.799	0.808
Multi-species MLP	0.617	**0.605**	0.708	0.714	0.803	0.815

This study contributes to the growing body of research on the application of neural networks to SDM, demonstrating that MLPs can achieve competitive performance compared to state-of-the-art methods. It also highlights the potential benefits of multi-species modelling, particularly for species with limited occurrence data, suggesting a new direction for improving SDM accuracy.

2.2.3 Effects of sample size and network depth on a deep learning approach to species distribution modeling

According to [28], it explores the application of deep learning, particularly Artificial Neural Networks (ANNs), to the task of SDM. This study aims to understand how variations in sample size and the depth of neural network architectures impact the performance of SDMs. With the increasing use of deep learning in ecological modelling, this study provides valuable insights into the conditions under which deep neural networks (DNNs) can be effectively utilized, compared to traditional methods like RF.

Occurrence data for this study were sourced from the National Aquatic Monitoring Center, focusing on freshwater macroinvertebrates collected across various sites in the western United States. The environmental predictors were derived from the StreamCat dataset, which provides a comprehensive set of 242 variables characterizing geoclimatic conditions across millions of stream segments. Table 2.8 presents ten key variables chosen for their significance in macroinvertebrate ecology.

Table 2.8 Variables included as predictors in SDM [28]

Predictor	Description
CatAreaSqKm	NHDPlus[a] catchment area (km^2)
HydrlCondCat	Mean catchment hydraulic conductivity of surface lithology (μm/s)
Mean_MSST	Predicted mean summer water temperature (C) for each segment averaged over years 2008, 2009, 2013, and 2014
Precip8110Cat	Catchment-scale PRISM[b] normal mean precipitation (mm) for 1981–2010
Tmax8110Cat	Catchment-scale PRISM normal maximum air temperature (C) for 1981–2010
Tmean8110Cat	Catchment-scale PRISM normal mean air temperature (C) for 1981–2010
Tmin8110Cat	Catchment-scale PRISM normal minimum air temperature (C) for 1981–2010
ElevCat	Mean elevation of the catchment (m)
BFICat	Catchment-scale base flow index describing the ratio of baseflow to total flow (%)
RunoffCat	Mean catchment runoff (mm)

[a] National Hydrography Dataset Plus.
[b] Parameter elevation Regression on Independent Slopes Model.

In data preprocessing, the environmental variables were standardized to ensure comparability. The dataset was then split into training, validation, and test sets, with stratification to maintain proportional species representation. As to assess the impact of sample size on model performance, three data subsets (100, 1,000, and 10,000 sites) were created for each genus, with each subset split into 70% training and 30% validation sets.

This study used the Adam optimizer, known for its robustness and efficiency, with a default learning rate of 0.001. The rectified linear unit (ReLU) activation function was applied across all hidden layers to enable faster learning and avoid issues like vanishing gradients. Batch normalization was used throughout, but dropout was excluded as it did

not improve performance in preliminary tests. The binary cross-entropy loss function was used, appropriate for the binary classification tasks, and the batch size was set to 50 for all models. Early stopping was implemented to determine the optimal number of training epochs, stopping when no improvement in validation loss was observed for 10 consecutive epochs. This approach enhanced model performance and reduced optimization time compared to traditional grid search methods.

For comparison, RF models were also developed using the randomForest package in R, with 500 trees and three variables randomly selected at each node, providing a strong baseline against which to evaluate the neural networks.

The study employed a random grid search strategy to optimize the number of nodes in each hidden layer. In order to manage the computational complexity, probabilistic reduction techniques were applied, narrowing the hyperparameter space. This approach involved randomly sampling a small fraction of possible node configurations and further reducing the search space by eliminating configurations that showed negative correlations with model performance in the initial runs. Table 2.9 summarizes the tested hyperparameter combinations for each neural network architecture after applying these reduction techniques.

Table 2.9 Optimization strategy [28]

Network depth	Possible combinations	Reduction fraction	Combinations tested
1 layer	20	1	20
2 layer	400	1	400
3 layer	8000	0.8	6400
4 layer	160,000	0.04	6400
5 layer	3,200,000	0.002	6400
6 layer	64,000,000	0.0001	6400

The primary performance metric used was the True Skill Statistic (TSS), which balances sensitivity and specificity, making it particularly suitable for evaluating models dealing with imbalanced datasets like species distribution data. TSS was calculated for both training and validation sets to monitor overfitting and generalization performance.

The results in Figure 2.3 showed that as the sample size increased from 100 to 10,000 sites, validation set performance improved significantly, indicating better generalization with larger datasets. Deeper networks performed better on larger datasets but offered no consistent advantage on smaller ones, suggesting that added complexity benefits only when sufficient data is available.

RF models performed comparably or slightly better than neural networks, especially on smaller datasets. This highlights the robustness of RF as a modelling technique in ecological studies, where data scarcity is often a challenge.

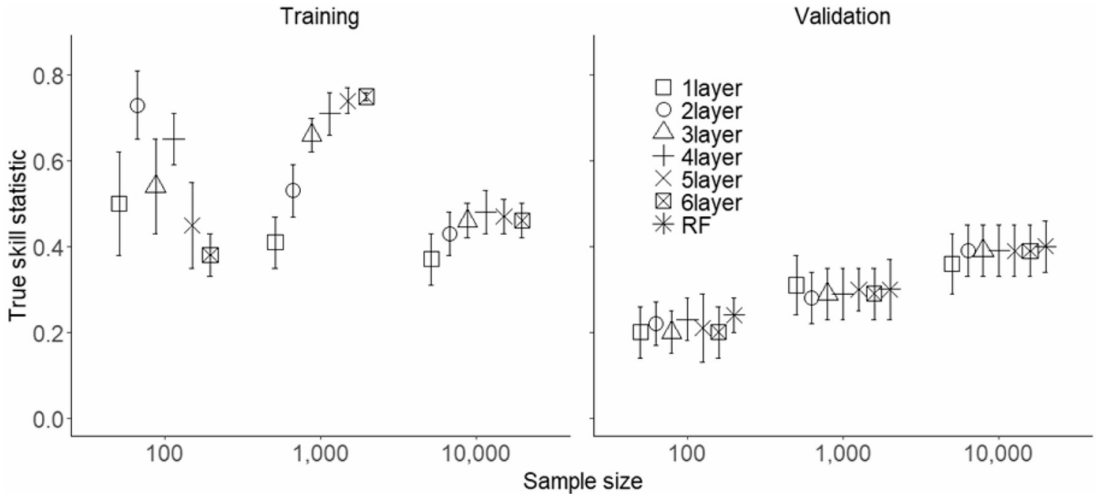


Figure 2.3 Performance metrics of models [28]

This study systematically compares the impact of network depth and sample size on the performance of DNNs, providing valuable insights into when and how deeper networks might be beneficial. The research also validates the performance of DNNs, showing that they can outperform shallow neural networks when trained on larger datasets, although they struggle with smaller datasets due to overfitting. In addition, the study offers a comparative analysis with RF models, demonstrating that RF often perform comparably or even better than DNNs, particularly with smaller datasets. This finding is significant as it challenges the assumption that deeper and more complex models always lead to better performance.

2.2.4 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure

In most cases, ecological data have inherent dependency structures such as spatial, temporal, hierarchical or phylogenetic correlations. These dependencies often violate the independence assumptions required by traditional validation methods. Therefore, random sampling as one of these traditional methods, which usually leads to overfitting and underestimation of prediction errors. To address the negative effects of these dependencies, this paper [29] introduces spatial block sampling and demonstrates that it is a more robust cross-validation strategy. This paper demonstrates that spatial block sampling produces error estimates that are closer to the true value, reduces overfitting, and improves model evaluation for predicting new regions or environmental conditions compared to traditional validation methods.

Block cross-validation has great potential in biogeographic research, but it is underestimated. In addition, this paper points out that many studies aim to predict new time periods or geographic regions but fail to describe the motivation for their validation methods, which is a concern. Through simulations and case studies, this paper shows that block cross-validation is valuable for extrapolated predictions and predictions within the same time period and region. Four scenarios are explored in this paper: spatial blocking, stratified group blocking, phylogenetic blocking and predictive spatial blocking. The results confirm that random cross-validation underestimates errors, even when models are designed to account for dependencies. In addition, block cross-validation provides more reliable error estimates that closely reflect the true values.

The Figure 2.4 compares cross-validation strategies for spatially structured data using Root-mean-square deviation (RMSE) to assess model accuracy. The black vertical line indicates the ideal RMSE (true error) for independent data. The overall bias of the re-substitution and random sampling to the left, showing a significant underestimation of the error due to spatial autocorrelation, which produces overly optimistic results. The block cross-validation shows that medium-sized blocks (20×20) provide the most accurate RMSE estimates by balancing independence and data availability, while small blocks (10×10) underestimate the error and large blocks (25×50) slightly overestimate the error. The buffer leave-one-out method (LOO) suggests that larger buffer radii can be effective in reducing spatial dependence and closely aligning with the ideal RMSE.

Overall, blocks cross-validation and buffered LOOs outperform random sampling and provide more realistic and reliable error estimates for spatially structured data.

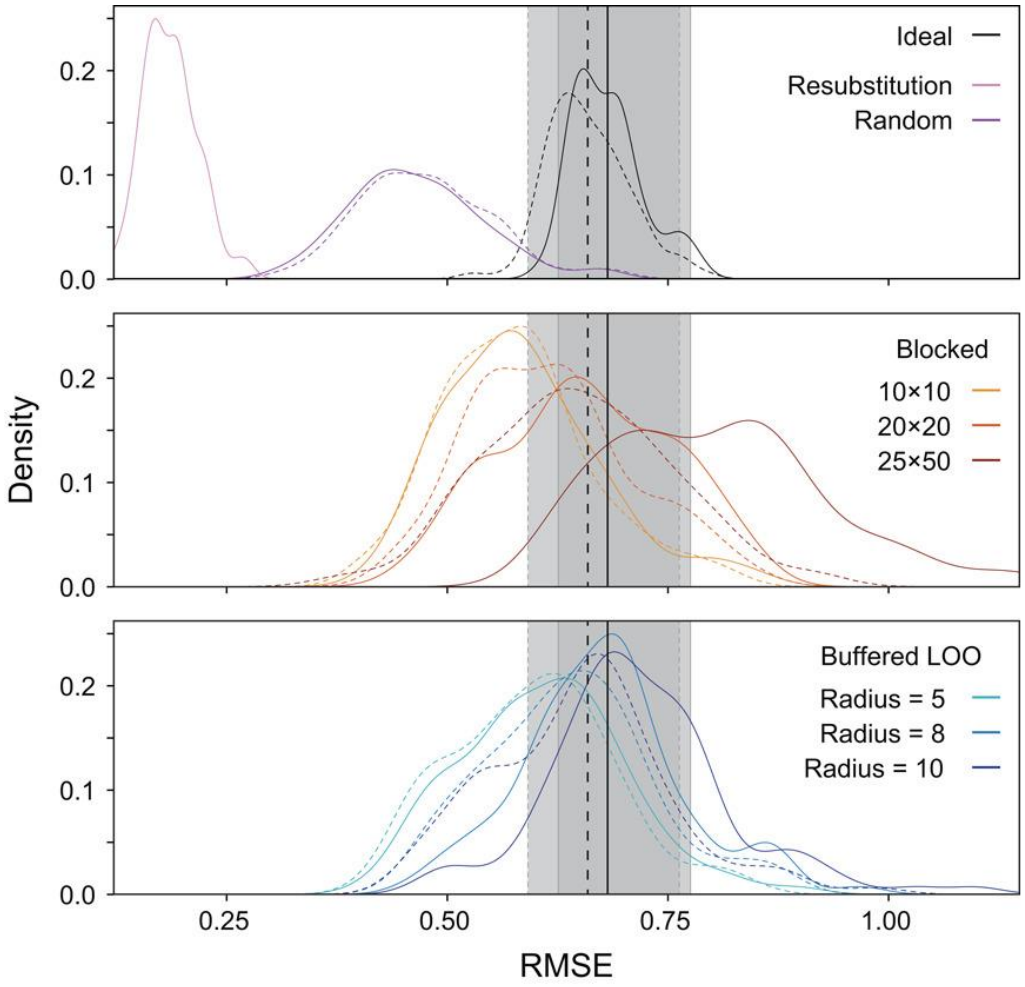


Figure 2.4 RMSE for Different Sampling Methods Across 100 Simulations [29]

The paper conclude that random sampling is often used and is an excellent method in general cross-validation. When random sampling is used in ecological data, where it is clearly unable to address the dependency structure of ecological data. Because of the random nature, it usually includes data points from nearby locations in the training and test sets, leading to overly optimistic error estimates. In addition, random sampling can lead to overfitting as the model absorbs local patterns, reducing its ability to generalise to new, unseen data.

In contrast, spatial block sampling divides the data into geographically distinct blocks, ensuring spatial independence between the training and test datasets. This approach reduces overfitting by isolating spatial structure and preventing the model from learning

local dependencies. It also provides more reliable error estimates that better reflect the model's true performance on independent data.

2.2.5 Limitations of Related Works

Most SDM studies focus on temperate regions in the United States and Europe, with significantly less research dedicated to the distinctly different tropical climates in Asia [15]. Malaysia, in Southeast Asia, has a tropical rainforest climate with consistently high temperatures, humidity, and significant rainfall year-round, sharply contrasting with the temperate climates of the US, EU, and China [30–33]. These conditions pose distinct challenges for SDM in Malaysia, requiring tailored approaches to accurately predict species distributions.

Overfitting is a common challenge in machine learning, especially as models become more complex [34]. While traditional machine learning methods typically focus on single-species predictions, deep neural networks (DNNs) like multi-layer perceptrons (MLPs) can capture intricate relationships in both single-species and multi-species scenarios. However, this strength can also be a drawback. Ecological datasets are usually small [28], which exacerbates the risk of overfitting in complex models like MLPs, causing them to learn noise rather than meaningful patterns. This challenge underscores the importance of careful model selection and regularization to ensure robust, generalizable predictions in ecological research.

A significant gap in the reviewed papers is the lack of discussion on how environmental variables contribute to species distribution in a specific region. It is important to understand the influence of environmental factors such as temperature and precipitation in a specific region, especially for the prediction of species' responses to climate change [35]. Without this focus, the models may fail to capture the full complexity of species-environment interactions, leading to less accurate predictions and potentially overlooking critical factors that drive species distribution.

2.3 Proposed Solutions

As to address the limitations identified in current SDM research, this study will focus on modelling the distribution of Strigiformes in Malaysia. By doing so, this study aims to fill the existing research gap and contribute to a more comprehensive understanding of how environmental change impacts species in tropical regions, particularly within Malaysia's unique climatic conditions, which are underrepresented in current SDM research.

To evaluate the most suitable data splitting method for environmental condition predictions, this study implements both random sampling and spatial block sampling, following the methodology discussed in [29]. The results will reveal which sampling method better generalizes across spatially independent regions, guiding the approach to future environmental predictions. The identified sampling method will be applied in subsequent analyses or further studies to enhance predictive accuracy and reliability

Based on the reviewed papers in Chapter 2, Random Forest (RF) models demonstrated strong performance across various species and sample sizes, while Multi-Layer Perceptrons (MLPs) are particularly effective with large datasets and multi-species distributions. However, MLPs are prone to overfitting, particularly when dealing with smaller datasets. Therefore, this study focuses on the development and comparison of RF and MLP models to determine the most suitable approach for the specific context of a relatively small dataset size and Malaysia’s unique tropical geographical conditions. The comparison aims to evaluate predictive performance, robustness, and generalization ability under current and future environmental scenarios.

In many existing studies, the role of environmental variables in influencing species distribution is often underexplored. Simultaneously, models used in SDM face challenges in balancing accuracy, complexity, and interpretability. As models become more sophisticated, like MLPs and RFs, they tend to become increasingly complex and harder to interpret. As to address this, this study will integrate Explainable AI (XAI) techniques, such as Shapley Additive Explanations (SHAP)to provide insights into how environmental factors influence predictions in specific regions in Malaysia. By applying XAI, the aim is to improve the proposed model’s accuracy while ensuring that predictions are both interpretable and actionable.

In order to have a more comprehensive understanding of the impact of environmental change on Strigiformes in Malaysia, this study will integrate probabilistic mapping into species distribution models. The probabilistic maps will be carried out for both current and future environmental scenarios, allowing the assessment current suitability of habitats and forecasting potential shifts.

CHAPTER 3 SYSTEM MODEL

3.1 Research Methodology

The methodology for this study is designed to develop and evaluate predictive models for species distribution under current and future environmental scenarios. The approach involves the development of machine learning and deep learning models, including Random Forest (RF), Multi-Layer Perceptrons (MLP).

The Figure 3.1 below outlines the overview of research framework of this study, which consists of three main phases. The Data Preprocessing phase (red section) includes collecting presence and environmental data, removing duplicates, generating pseudo-absence data, resampling environmental variables, and conducting Pearson correlation analysis to produce a finalized occurrence dataset. The Data Splitting Method Evaluation phase (yellow section) will compare random sampling and spatial block sampling methods by developing and evaluating models to determine the better splitting strategy. Next, the Model Development and Evaluation phase (blue section) applies the selected sampling method to the whole Malaysia dataset for model development and evaluation. Finally, the Environmental Impact and Habitat Analysis phase (green section) used the trained models to conduct environmental impact assessments and create habitat suitability maps for current and future conditions.

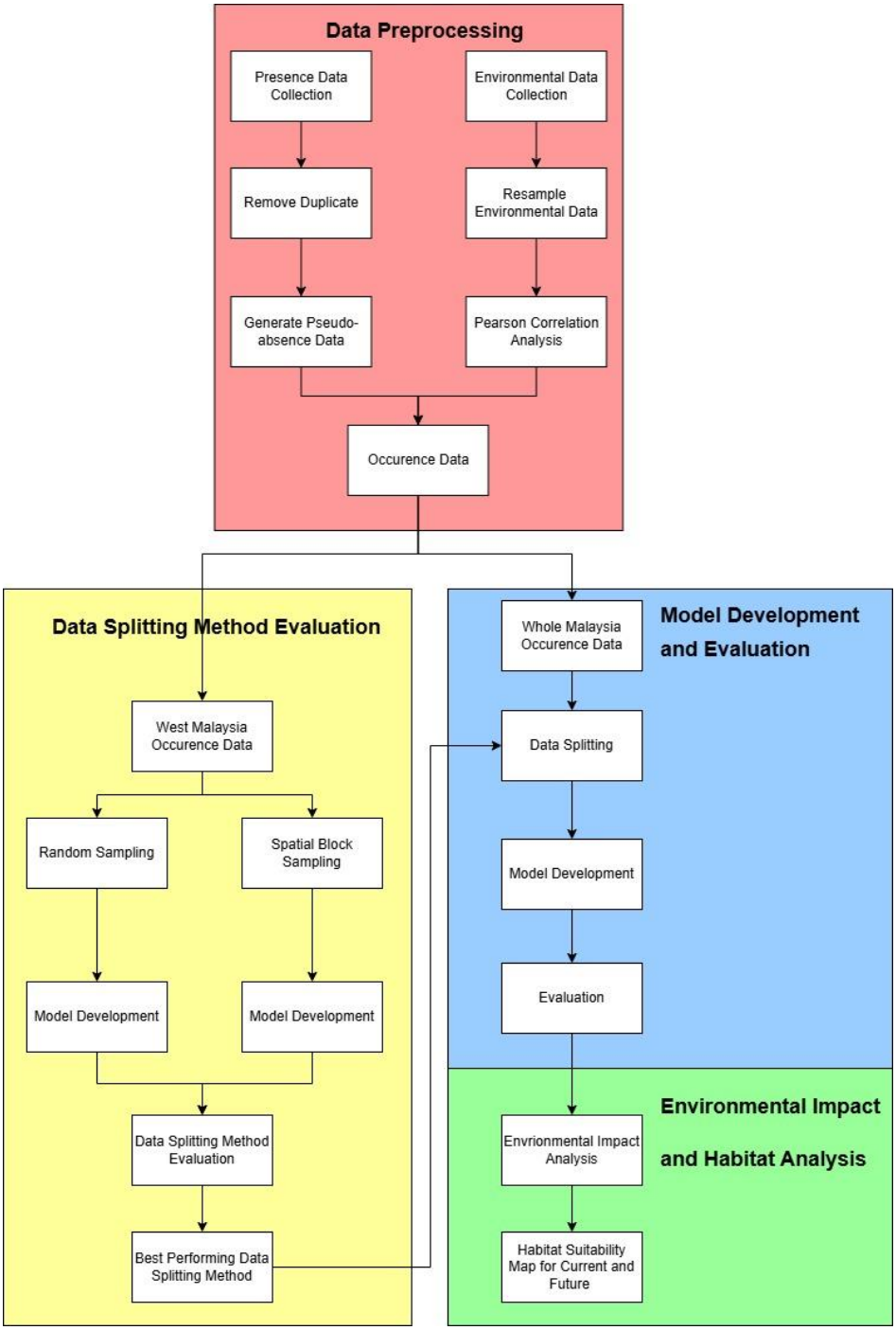


Figure 3.1 Overview of Research Framework

3.1.1 Data Preprocessing

The first phase of the project focuses on data preprocessing, which includes the collection, cleaning and preparation of presence and environmental data. It begins with the collection of presence data from biodiversity databases from Global Biodiversity Information Facility (GBIF) [19], and the acquisition of 19 environmental variables from sources WorldClim [20]. In addition, elevation and slope data were retrieved from National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM) Digital Elevation 30m dataset [21], while land cover variables, including primary forest, secondary forest, and urban areas, were obtained from the Land-Use Harmonization (LUH2) dataset [22].

A. Presence Data

The study area includes both West and East Malaysia, focusing on modelling the distribution of Ketupa (a genus of Strigiformes). For this research, presence-only data shall be obtained from the GBIF, concentrating on records within Malaysia's geographical boundaries [19]. The specific species for study are listed in Table 3.1.

Table 3.1 Selected species under Ketupa

Order	**Genus**	**Species**
Strigiformes	**Ketupa**	Ketupa ketupu
		Ketupa zeylonensis
		Ketupa sumatrana
		Ketupa coromanda

B. Environmental Data

Furthermore, the 19 bioclimatic variables sourced from WorldClim [20], along with additional terrain and land-use variables are listed in Table 3.2. The total 24 envrionmental variables have been demonstrated to play a critical role in species distribution modelling, making them highly suitable for this study [26, 36].

Table 3.2 24 Environmental variables

Variable	**Description**
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (×100)
BIO4	Temperature Seasonality (standard deviation ×100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5 - BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter
Elevation	Elevation above sea level (meters) from NASA SRTM
Slope	Terrain slope derived from NASA SRTM
Primf_median	Proportion of grid cell covered by forested primary land (LUH2 primf)
Secdf_median	Proportion of grid cell covered by potentially forested secondary land (LUH2 secdf)
Urban_median	Proportion of grid cell covered by urban areas (LUH2 urban)

C. Remove Duplicates

To ensure data integrity and usability, duplicate presence records are identified and removed. The spatial resolution for analysis is set at 1 kilometre (km) to ensure consistency with environmental variables and to account for species-level habitat precision. Within each 1 km grid, only one presence record is retained, while others are treated as duplicates to avoid overrepresentation of data in densely sampled areas [37]. The environmental variables are similarly resampled to match this 1 km spatial resolution, ensuring compatibility between the environmental predictors and occurrence data.

C. Pseudo-absence Generation

To prepare presence-only data for use in machine learning models like RF and MLP, pseudo-absence data are required to be generated. According to the suggestion in [38], random selection of pseudo absence with minimum distance away from presence data is a robust method for such models. Based on the recommendation, the number of pseudo-absences should match the number of presence data. Additionally, the generation process should be repeated 5 times to improve reliability.

There is no specific research on the home range of genus Ketupa is currently available. To address this limitation, a study [39] of related species within same order, Strigiformes was used as a reference. The barn owl, which also belongs to Strigiformes order, has been found that its home range size is approximately 10 km. Therefore, this study will set the minimum distance of pseudo-absence generation as 10 km. This approach will eventually increase the likelihood that pseudo-absences represent true absences, improving the robustness and reliability of models.

D. Pearson Correlation Analysis

The Pearson correlation coefficient is a widely used statistical measure that quantifies the linear relationship between two variables, providing a value between -1 and 1, where values closer to -1 or 1 indicate a stronger linear relationship [40]. The threshold of 0.7 is commonly used in various studies, as it effectively minimizes the risk of multicollinearity, thereby enhancing the accuracy and reliability of SDM projections [41]. After this, data standardization will be applied to ensure all variables are on a

comparable scale, which is essential for optimizing the performance of deep learning models like MLP.

The formula of Pearson Coefficient is shown in equation 3.1:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad (3.1)$$

where:

- Covariance(X, Y) = $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
- Standard Deviation(X) = $\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Standard Deviation(Y) = $\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

3.1.2 Data Splitting Method Evaluation

The second phase of the project evaluates different data splitting methods to identify the most suitable approach for species distribution modeling (SDM) in Malaysia. The occurrence data is divided into West Malaysia (used for model training and testing) and East Malaysia (used as an independent evaluation dataset). Two sampling strategies are employed: random sampling and spatial block sampling.

As discussed in Chapter 2, the paper [29] demonstrates dependency structures such as spatial, temporal, or phylogenetic correlations in ecological data often violate the independence assumptions of traditional validation methods, leading to overestimated model performance. Spatial block sampling is an effective method for ensuring spatial independence between training and test datasets.

For this study, random sampling randomly selects data points for training and testing, while spatial block sampling divides the dataset into geographically distinct blocks. Both models are evaluated using the independent data from East Malaysia and the West Malaysia test set.

Based on previous research [29], models trained with random sampling are expected to perform better on the West Malaysia test set but may struggle with extrapolation to the

East Malaysia dataset. Conversely, spatial block sampling models, designed for better generalization across regions, are expected to outperform random sampling in the East Malaysia evaluation. This analysis will determine which sampling method is more robust and will be adopted for subsequent steps, including model development, environmental data impact analysis and habitat suitability map.

3.1.3 Model Development and Evaluation

This phase includes applying the best sampling method identified in the second phase, followed by model development and evaluation. The whole Malaysia occurrence dataset is used for model training and evaluation.

Subsequently, Random Forest (RF) and Multi-Layer Perceptron (MLP) are developed to capture species-environment relationships effectively.

A. Random Forest (RF) Model Architecture

This research develops RF models [42] across both datasets and genera under study. Each RF model is configured with 500 decision trees. At each decision node within the trees, three variables are randomly selected for consideration when determining the best split. This approach helps in generating a diverse set of trees, reducing the risk of overfitting [28].

B. Multi-Layer Perceptron (MLP) Model Architecture

The MLP model [28] is employed to capture the complex, non-linear relationships between environmental variables and species distributions. The input layer processes environmental variables after Pearson correlation [40], and the output layer consists of a single node with a sigmoid activation function. The MLP architecture consists of an input layer, three hidden layers and a single node output layer.

Input Layer:

The input layer processes the standardized environmental variables after Pearson correlation [40].

Hidden Layers:

This study uses three hidden layers to balance the model complexity and dataset size (approximately 1000 training records), following suggestions from reviewed paper

[28]. The Rectified Linear Unit (ReLU) activation function is used on all hidden layers to enhance learning efficiency by addressing the vanishing gradient problem.

Output Layer:

The output layer comprises a single neuron with a sigmoid activation function, predicting the probability of species presence.

Optimization and Loss Function:

The Adam optimizer is used for model training [28]. The binary cross-entropy loss function is utilized to minimize the classification error between predicted and true labels.

Training Strategy:

The MLP models underwent hyperparameter tuning using a custom randomized search procedure, primarily focusing on the number of neurons in each hidden layer, learning rate, dropout rate, batch size, and number of epochs. The specific search space and implementation details are provided in Chapter 4.

C. Evaluation

The Area Under the Receiver Operating Characteristic Curve (AUROC) [43] and the Area Under the Precision-Recall Curve (AUCPR) [44] are metrics used for evaluating binary classification models, offering complementary insights into model performance.

AUROC is derived from the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various classification thresholds.

TPR, also known as sensitivity or recall, is the proportion of actual positive instances correctly identified by the model as shown in equation 3.2:

$$TPR = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{3.2}$$

FPR represents the proportion of actual negative instances incorrectly identified as positive as presented in equation 3.3:

$$FPR = \frac{False\ Positives\ (FP)}{True\ Negatives\ (TN) + False\ Positives\ (FP)} \tag{3.3}$$

By evaluating TPR as given in equation 3.2 and FPR as given in equation 3.3 across all possible thresholds, the ROC curve visualizes the trade-off between true positives and false positives. AUROC, as the area under this curve, provides a comprehensive evaluation of the model's discriminative power, particularly valuable in scenarios with class imbalance, ensuring that the model's performance is not biased toward the majority class.

AUCPR is calculated by integrating the Precision-Recall (PR) curve, which plots Precision against Recall. The curve represents how these two metrics trade off as the classification threshold changes. The formula of Precision and Recall are shown in equation 3.4 and equation 3.5.

Precision:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{3.4}$$

Recall:

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \tag{3.5}$$

By evaluating Precision as given in equation 3.4 and Recall as shown in equation 3.5 across all possible thresholds, the Precision-Recall curve visualizes the trade-off between the accuracy of positive predictions and the model's ability to identify all actual positives. AUCPR quantifies this trade-off, providing a focused evaluation of the model's performance on the positive class.

AUROC evaluates a model's ability to distinguish between positive and negative classes by incorporating all classes from confusion matrix as illustrated in Figure 3.2. In contrast, AUCPR is calculated based on Precision and Recall, which do not account for true negatives class. In this study, the datasets from GBIF [19] include only presence-only data, while absence data is artificially generated as pseudo-absence data, resulting in an imbalanced dataset.

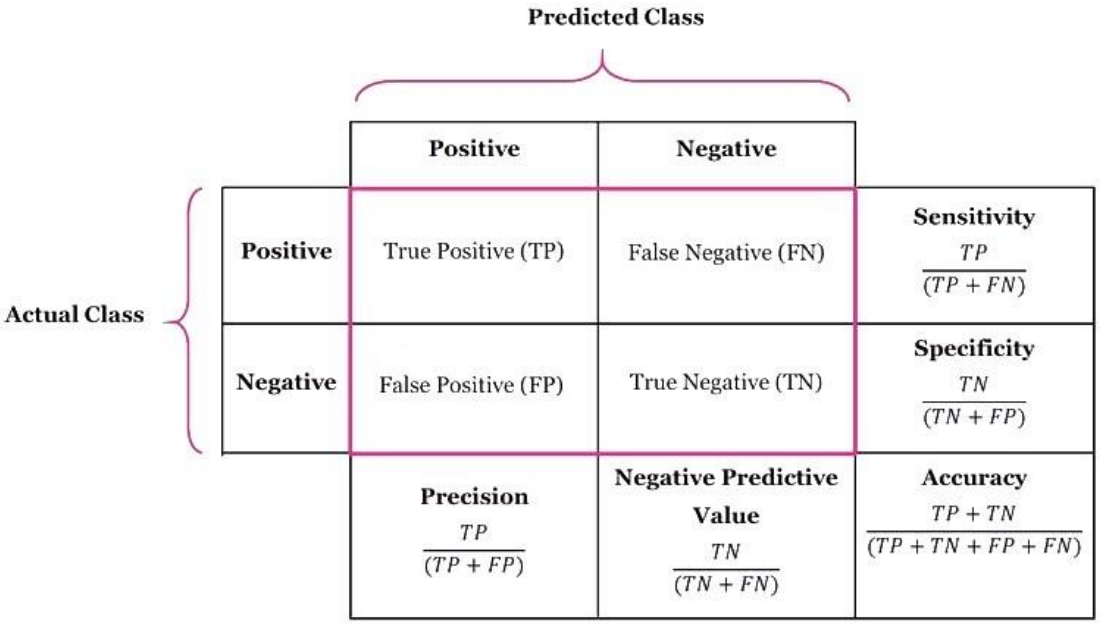


Figure 3.2 Confusion Matrix

In ecological modelling, datasets often consist of presence-only data, and even when true absence data is available, presence data typically represents a rare event within the dataset. According to [45], AUROC often inflates performance when dealing with imbalanced datasets,, either due to a majority of true negatives or because true negatives are represented by pseudo-absence data. The study suggests using AUCPR as a complementary metric for assessing performance of model because it ignores true negatives (including pseudo-absence data) and focuses solely on the model's ability to predict presence data. By using both AUROC and AUCPR, it ensures a more accurate and reliable evaluation of model performance

3.1.4 Environmental Impact and Habitat Analysis

Finally, the environmental impact and habitat analysis phase focuses on evaluating the influence of environmental variables on species distribution using advanced interpretability techniques. This phase aims to provide insights into the relationships between predictor variables and species occurrence

The analysis employs three key methods: Mean Decrease in Impurity (MDI) [42], Shapley values (SHAP) [46], and response curves.

A. Mean Decrease in Impurity

Mean Decrease in Impurity (MDI) [42] ranks environmental variables based on their contribution to reducing impurity in decision trees, providing an overall measure of variable importance within the Random Forest (RF) model. Impurity is measured using Gini Impurity, which quantifies the likelihood of incorrectly classifying a randomly chosen sample if it were randomly labeled according to the distribution of the labels in a given node.

The formula for Gini Impurity is given in equation 3.6 below:

$$Gini = 1 - {p_1}^2 - {p_2}^2 \tag{3.6}$$

where:

- p_1 : The proportion of samples in presence class.
- p_2 : The proportion of samples in absence class.

The Gini Impurity Reduction is calculated as presented in equation 3.7:

$$Gini\ Reduction = Gini_{\ Parent} - \left(\frac{n_{left}}{n_{parent}} \cdot Gini_{\ left} + \frac{n_{right}}{n_{parent}} \cdot Gini_{\ right}\right) \tag{3.7}$$

where:

- $Gini_{parent}$: Gini Impurity of the parent node.
- $Gini_{left}$, $Gini_{right}$: Gini impurities of the left and right child nodes after split.
- n_{parent}, n_{left}, n_{right}: Number of samples in parent, left child and right child nodes.

MDI measures the overall importance of a feature by summing the Gini impurity reductions for that feature across all nodes and averaging it across all trees in a RF.

It is important to note that MDI method can be only applied to tree-based models. Since MDI relies on the hierarchical structure of decision trees to calculate the feature

importance, it is not applicable to non-tree models like MLP. Therefore, in this study, MDI analysis is conducted exclusively on the RF models.

B. Shapley Values

Shapley values [46] offer an interpretable approach by quantifying each variable's contribution to individual predictions, capturing both localized effects and variable interactions. Shapley values are derived from cooperative game theory and provide a fair allocation of contribution among variables.

The formula of Shapley Values is shown in equation 3.8:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!} \cdot [f(S \cup \{i\}) - f(S)] \tag{3.8}$$

where:

- The set of all environmental variables ($\{x_1, x_2, \ldots, x_n\}$)
- S: A subset of variables excluding
- $|S|$: The number of variables in subset
- $f(S)$: The prediction of the model using only the variables in S (other variables are marginalized or ignored).
- $f(S \cup \{i\}) - f(S)$: The marginal contribution of variable x_i when added to subset S.
- $\frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!}$: A weighting factor that ensures all subsets are weighted fairly based on their size.

Shapley values provide a robust understanding of variable importance by accounting for interactions and dependencies among predictors.

C. Response Curves

Response curves are generated to illustrate how each environmental variable influences the predicted probability of species presence [47]. To create these curves, the target variable is systematically varied across its range while all other variables are held

constant at their mean values. This approach isolates the effect of the target variable, allowing the model's response to changes in that variable to be assessed independently. These curves provide a visual understanding of species responses to environmental conditions, highlighting critical thresholds and ranges essential for determining habitat suitability.

By combining MDI, Shapley values, and response curves, it ensures a comprehensive understanding of the environmental drivers of species distribution, contributing to the creation of accurate habitat suitability models and maps.

D. Habitat Suitability Map

The trained models generate spatially explicit probability layers that depict the present-day likelihood of genus Ketupa occurrence as well as projections for 2061–2080 under two Coupled Model Intercomparison Project Phase 6 (CMIP-6) climate pathways: Shared Socioeconomic Pathway 2–4.5 (SSP2-4.5), representing a moderate-emissions scenario, and Shared Socioeconomic Pathway 5–8.5 (SSP5-8.5), representing a high-emissions trajectory.

To move beyond simple visual comparison, the continuous habitat suitability probabilities are classified into four discrete classes as shown in Table 3.3. The analysis focuses on net area changes within each suitability category to assess spatial and ecological shifts between current and future scenarios [48].

The classification thresholds are summarized in Table 3.3.

Table 3.3 Habitat Suitability Classification

Suitability Class	**Probability Range**
High	0.75 – 1.00
Moderate	0.50 – 0.75
Poor	0.25 – 0.50
Unsuitable	0.00 – 0.25

3.2 System Requirements

3.2.1 Hardware

Table 3.4 Specifications of Hardware

Description	**Specifications**
Model	Acer Nitro 5 AN515 Gaming Laptop
Processor	Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz 2.40 GHz
Operating System	Windows 10
Graphic	NVIDIA GeForce GTX 1650
Memory	12.0 GB RAM 2667MHz DDR4
Storage	512GB PCIe® 3.0 NVMe™ M.2 SSD

3.2.2 Software Requirement

This research utilizes Python as the primary programming language, with Jupyter Notebook for coding and workflow documentation. Google Earth Engine (GEE) is employed for preprocessing environmental data, generating geospatial layers. Machine learning and deep learning models are developed by Scikit-learn, while TensorFlow is planned for training deep learning models such as Multi-Layer Perceptrons. Matplotlib and Seaborn are used for data visualization, while GeoPandas is essential for managing geospatial data. Additionally, R programming is used specifically for spatial block analysis, taking advantage of its specialized ecological and spatial statistics packages.

CHAPTER 4 IMPLEMENTATION & EXPERIMENT RESULTS

4.1 Data Preprocessing Implementation

4.1.1 Presence Data Cleaning

First, after loading the genus Ketupa data from the dataset, the rows with missing values were removed using the .dropna() method. The cleaned data was then converted into a GeoPandas DataFrame to prepare it for use in Google Earth Engine. Figure 4.1 below illustrates the data size and structure of the DataFrame.

```
df = df.dropna()
print("len ketupa:", len(df))
gdf = gpd.GeoDataFrame(
    df,
    geometry=gpd.points_from_xy(df.decimalLongitude,
                                df.decimalLatitude),
    crs="EPSG:4326"
)[["genus", "geometry"]]
print(gdf.info())
```

```
len ketupa: 5575
<class 'geopandas.geodataframe.GeoDataFrame'>
Index: 5575 entries, 4 to 17646
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   genus     5575 non-null   object
 1   geometry  5575 non-null   geometry
dtypes: geometry(1), object(1)
memory usage: 130.7+ KB
None
```

Figure 4.1 Showing data size and structure

4.1.2 Selecting Area of Interest

After that, the next important step was to define the area of interest (AOI) for this study. The focus of the study is Malaysia, as shown in Figure 4.2.

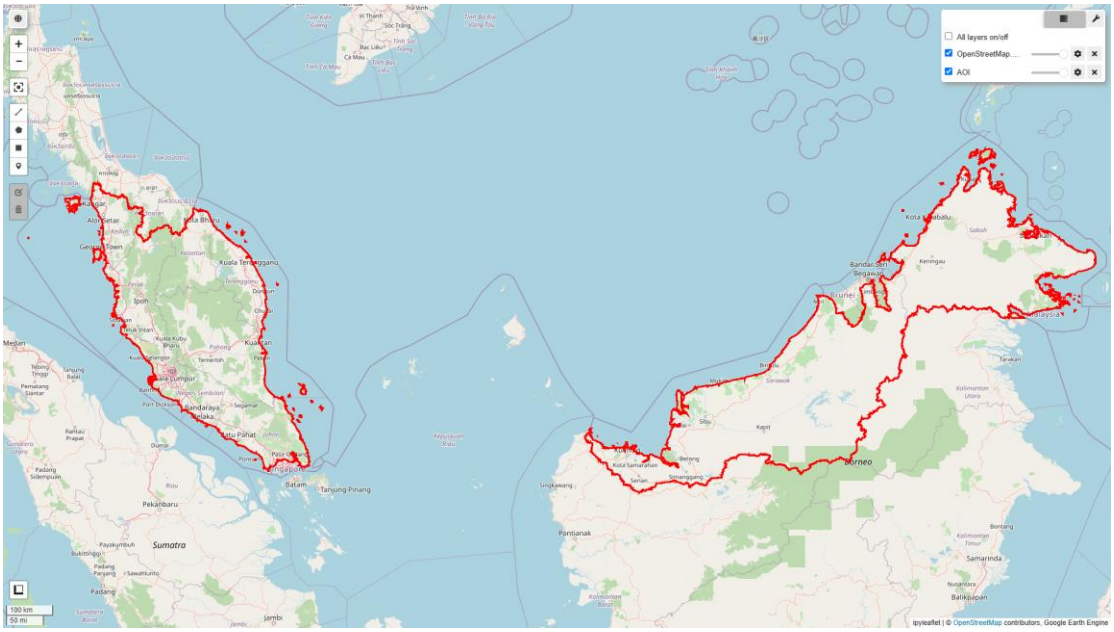


Figure 4.2 Area of Interest of Study (Highlighted with a Red Boundary)

4.1.3 Remove Duplicates

The duplicate and closely located points were removed, with the grain size set to 1 km (1000 metres) [37], ensuring that only data within Malaysia was retained. The original genus Ketupa dataset size of 5575 points was reduced to a final dataset size of 749 points. This significant reduction indicates that many points were either duplicates or located too close to one another, which could increase spatial autocorrelation. The data points have been plotted on a map, as shown in Figure 4.4, where the blue points represent the original data, and the red points represent the final dataset.

As mentioned previously, the genus Ketupa includes several species. After presence data preprocessing, which involved duplicate removal and spatial filtering, the final dataset retained four Ketupa species. The distribution of presence points among the species is shown in Figure 4.5: Ketupa ketupu with 477 records, Ketupa sumatrana with 242 records, Ketupa coromanda with 25 records, and Ketupa zeylonensis with 5 records.

```python
# Spatial resolution setting (meters)
grain_size = 1000
def remove_duplicates(data, grain_size):
    # Select one occurrence record per pixel at the chosen spatial resolution
    random_raster = ee.Image.random().reproject("EPSG:4326", None, grain_size)
    rand_point_vals = random_raster.sampleRegions(
        collection=ee.FeatureCollection(data), geometries=True
    )
    return rand_point_vals.distinct("random")


data = remove_duplicates(data_raw, grain_size)

#filter data within Malaysia
data = data.filterBounds(aoi)
```

Figure 4.3 Implementation of Removing Duplicates

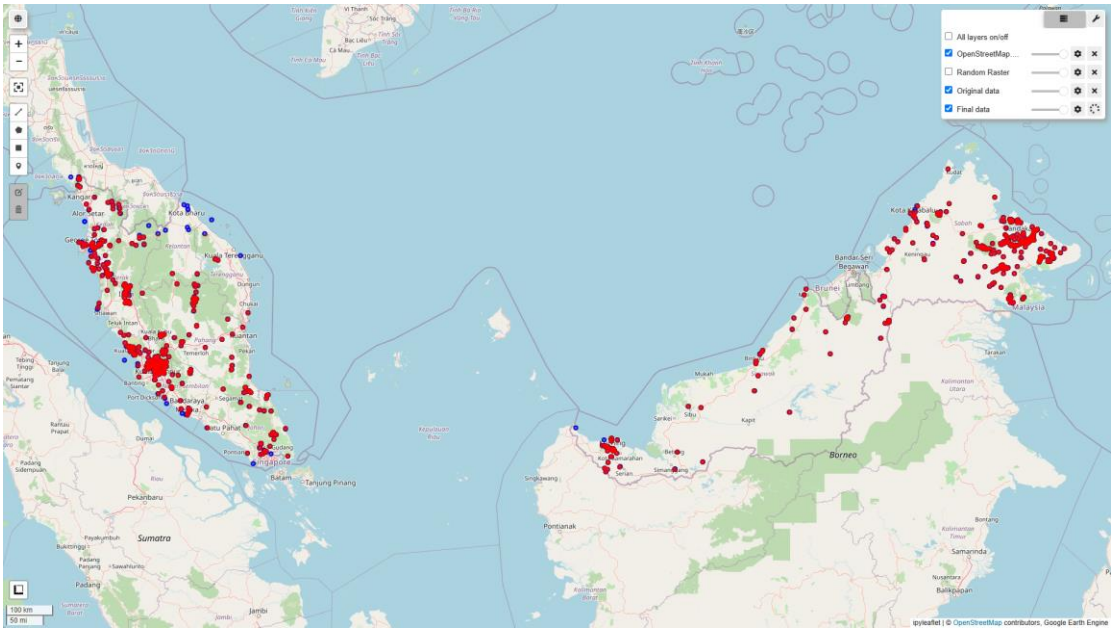


Figure 4.4 Plotting Data Points on Google Map

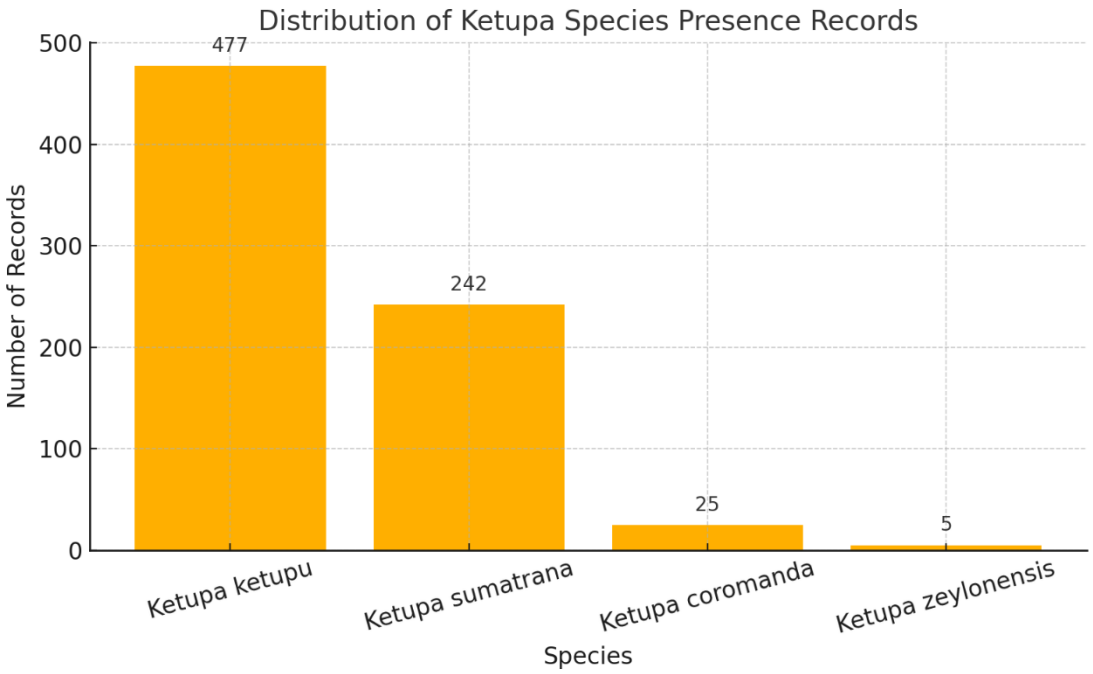


Figure 4.5 Distribution of Ketupa Species Presence Records

4.1.4 Pearson Correlation Analysis

After loading the datasets from WorldClim [20], NASA SRTM Digital Elevation [21] and Land-Use Harmonization [22], Pearson Correlation Analysis [40] is performed to address potential multicollinearity. A total of 10,000 random points within the AOI are selected and paired with all 24 environmental variables. A Pearson correlation matrix is then generated for these variables using the data from the 10,000 points.

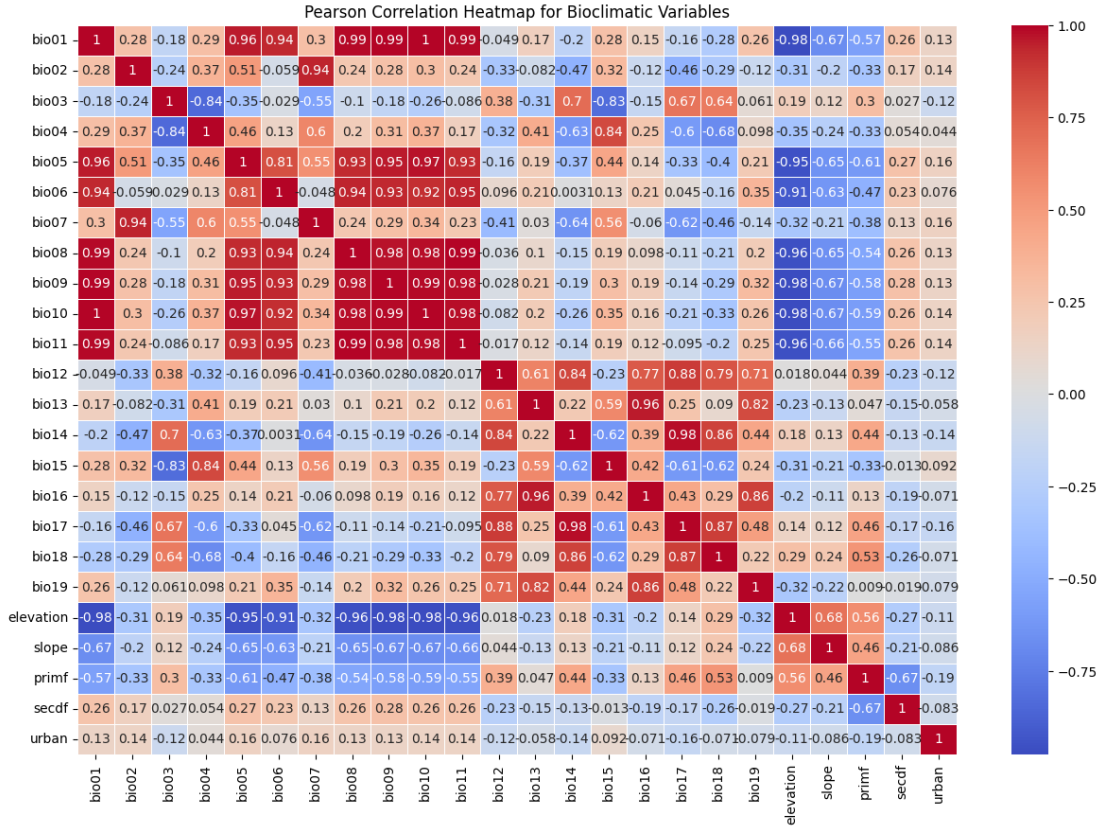


Figure 4.6 Pearson Correlation Matrix

The correlation matrix has been generated, and as discussed in Chapter 3.1.1 Data Preprocessing section, a threshold of 0.7 is applied. This threshold is commonly used in various studies as it effectively reduces the risk of multicollinearity [41]. After removing highly correlated variables based on this criterion, the remaining variables are bio01 (annual mean temperature), bio02 (mean diurnal range), bio03 (isothermality), bio12 (annual precipitation), bio13 (precipitation of wettest month), slope (Terrain Slope), primf (Primary Forest Cover Fraction), secdf (Secondary Forest Cover Fraction), and urban (Urban Area). The updated correlation matrix is presented

in

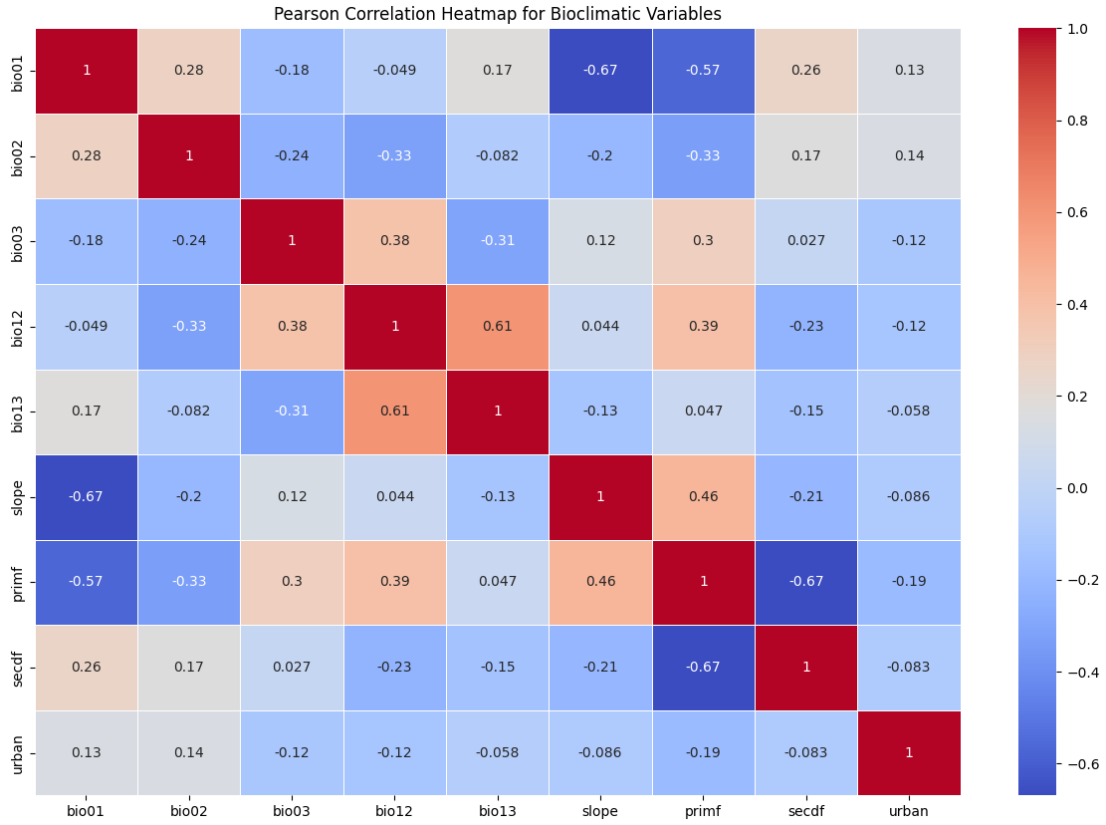

Pearson Correlation Heatmap for Bioclimatic Variables
bio01
bio02
bio03
bio12
bio13
slope
primf
secdf
urban
1 0.28 -0.18 -0.049 0.17 -0.67 -0.57 0.26 0.13
0.28 1 -0.24 -0.33 -0.082 -0.2 -0.33 0.17 0.14
-0.18 -0.24 1 0.38 -0.31 0.12 0.3 0.027 -0.12
-0.049 -0.33 0.38 1 0.61 0.044 0.39 -0.23 -0.12
0.17 -0.082 -0.31 0.61 1 -0.13 0.047 -0.15 -0.058
-0.67 -0.2 0.12 0.044 -0.13 1 0.46 -0.21 -0.086
-0.57 -0.33 0.3 0.39 0.047 0.46 1 -0.67 -0.19
0.26 0.17 0.027 -0.23 -0.15 -0.21 -0.67 1 -0.083
0.13 0.14 -0.12 -0.12 -0.058 -0.086 -0.19 -0.083 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 4.7.

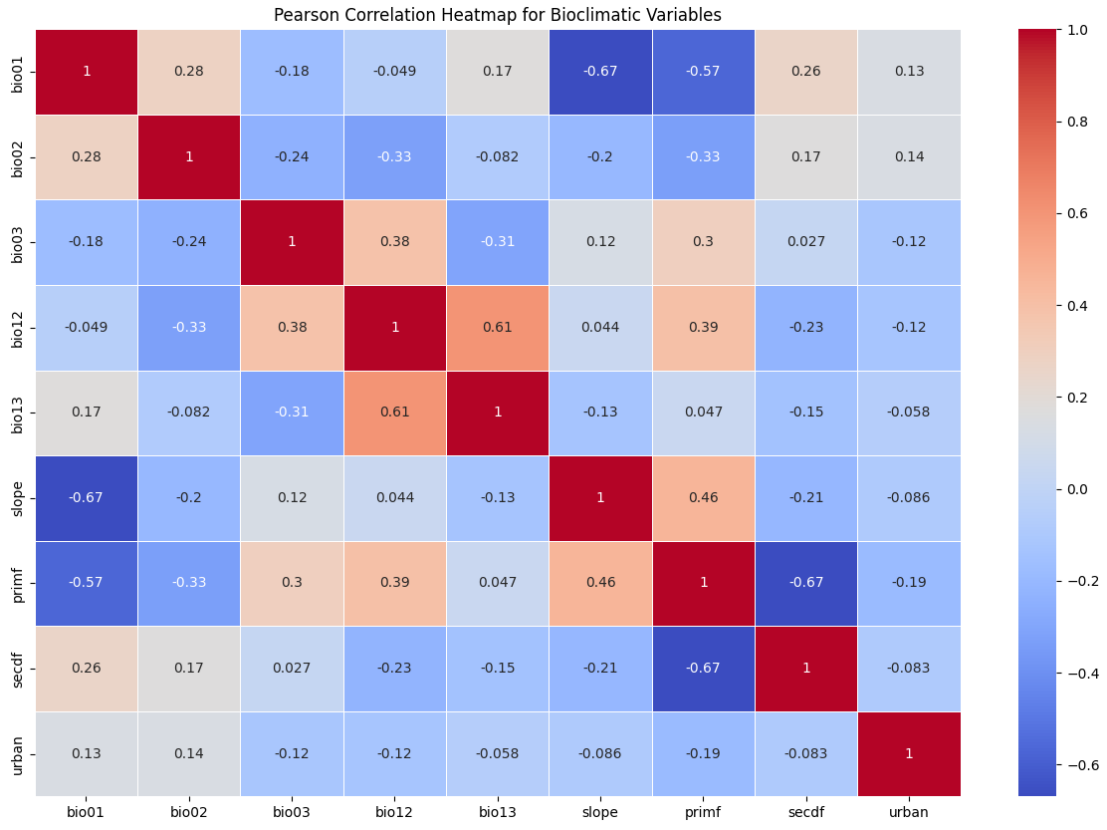


Figure 4.7 Updated Correlation Matrix

4.1.5 Initial Feature Importance Analysis

The initial set of predictors was determined based on Pearson correlation analysis, as summarized in Table 4.1. To further assess the true contribution of each predictor to species distribution modelling, an initial feature importance evaluation was conducted. Three complementary methods were used to examine variable importance: Response Curves, Shapley values (SHAP), and performance metrics.

Table 4.1 Predictors after Pearson Correlation Analysis

	Variable	**Description**
1	BIO1	Annual Mean Temperature
2	BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
3	BIO3	Isothermality (BIO2/BIO7) (×100)
4	BIO12	Annual Precipitation
5	BIO13	Precipitation of Wettest Month
6	Slope	Terrain slope derived from NASA SRTM
7	Primf_median	Proportion of grid cell covered by forested primary land (LUH2)
8	Secf_median	Proportion of grid cell covered by potentially forested secondary land (LUH2)
9	Urban_median	Proportion of grid cell covered by urban areas (LUH2)

Random Forest models were developed using these nine predictors, with each model configured with 500 decision trees. At each decision node, three variables were randomly selected to determine the best split, promoting diversity among the trees.

Analysis of variable importance revealed that the slope predictor displayed inconsistent response curve patterns across all cross-validation folds in Figure 4.8, suggesting weak and unstable relationships with species occurrence. Furthermore, slope showed the lowest mean Shapley value impact among all predictors, indicating minimal contribution to model predictions, as illustrated in Figure 4.9.

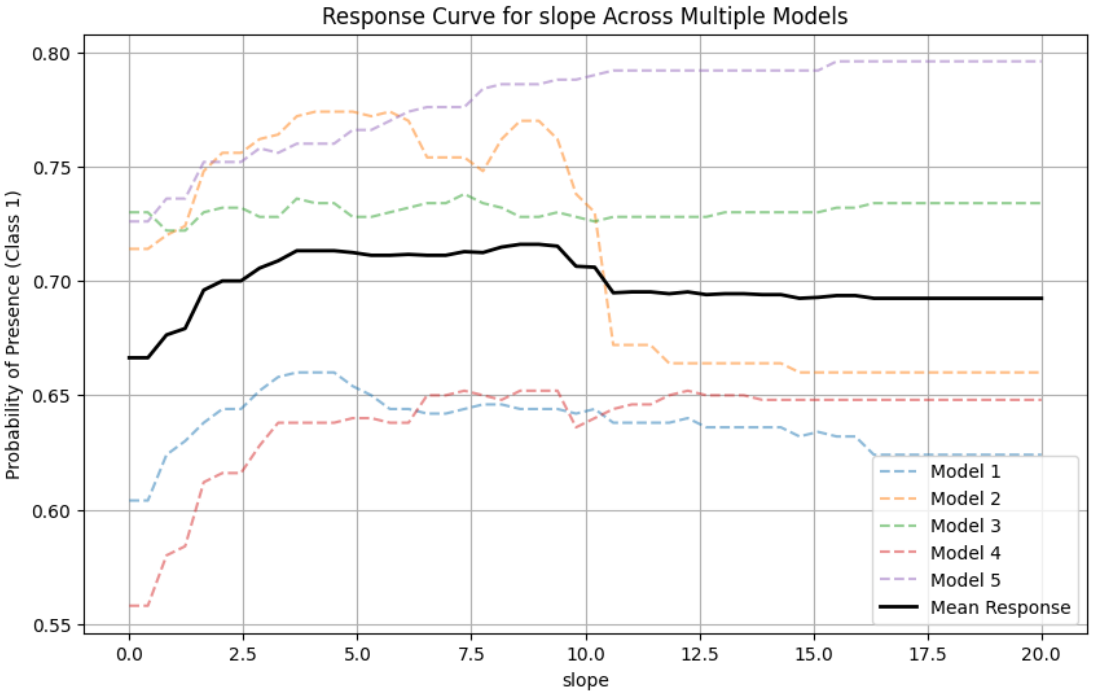


Figure 4.8 Response Curve for Slope using Random Forest

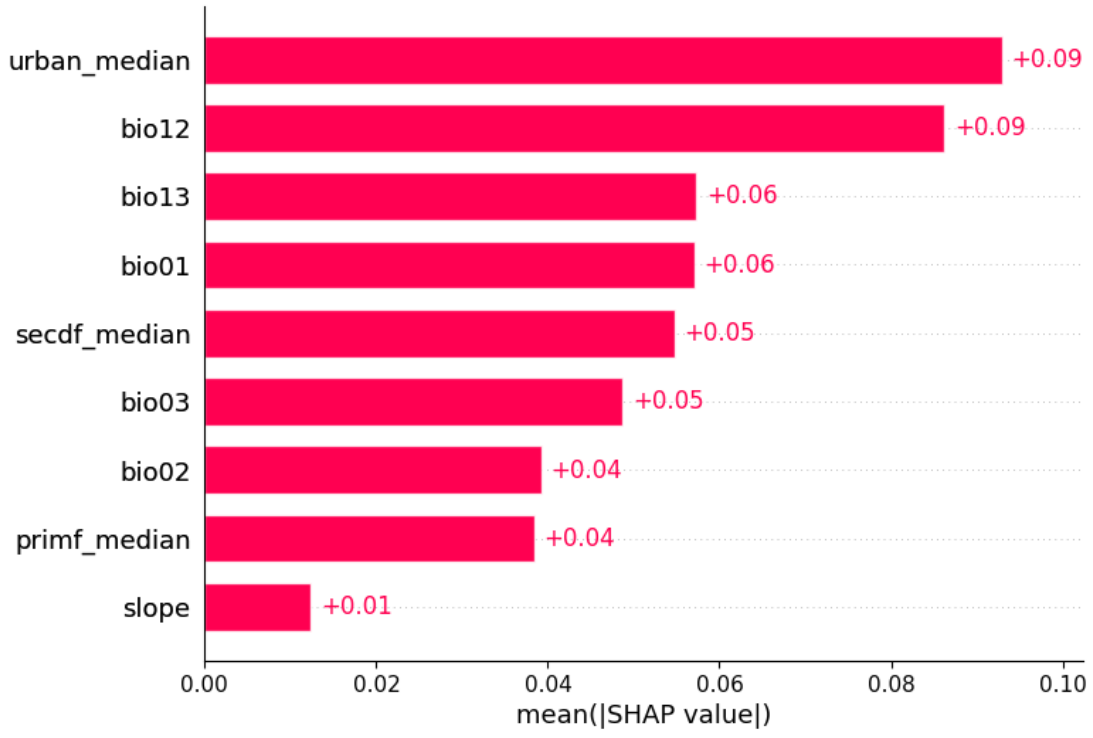


Figure 4.9 Mean Shapley Values of Random Forest

To further validate this observation, the slope variable was removed, and the model was retrained using the remaining eight predictors: bio01, bio02, bio03, bio12, bio13, urban_median, primf_median, and secdf_median.

Comparison of the final results demonstrated that removing the slope variable did not lead to any significant decline in model performance, as illustrated in Figure 4.10. These outcomes confirmed that the slope variable did not meaningfully contribute to the predictive power of the model, and its exclusion served to simplify the model without compromising accuracy or robustness.

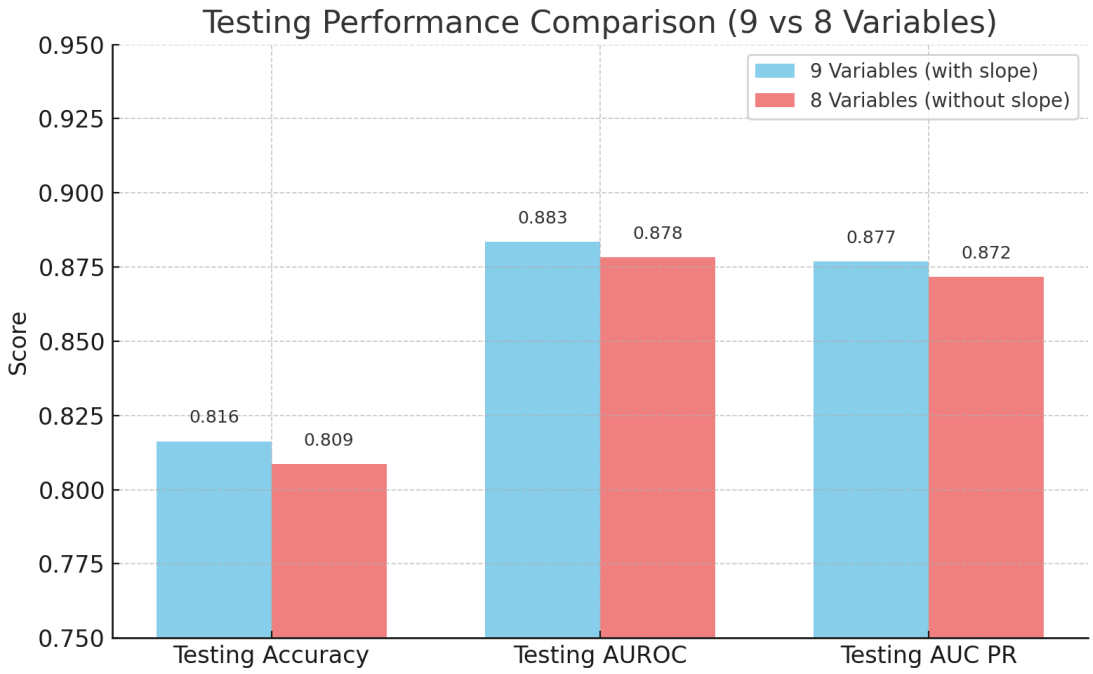


Figure 4.10 Performance Comparison between 9 variables and 8 variables

4.1.6 Pseudo-absence Generation

The GBIF datasets [19] include only presence-only data. Before model development, pseudo-absence data needs to be generated to act as the negative class. As discussed in Chapter 3.2.1 Data Preprocessing section, buffer is set at 10 km (10,000 m), meaning pseudo-absence points are located at least 10 km away from any presence point for genus Ketupa. As shown in Figure 4.12, the black-colored regions indicate the possible areas for pseudo-absence generation.

```
# Create a buffer around presence locations
buffer = 10000  # Distance in meters
buffer_presence = data.geometry().buffer(buffer)

presence_mask = data.reduceToImage(properties=['random'],
reducer=ee.Reducer.first()
).reproject('EPSG:4326', None,
            grain_size).mask().neq(1).selfMask()

# Mask the area outside the buffer and intersect it with the AOI
outside_buffer = aoi.difference(buffer_presence)
area_for_pa = presence_mask.clip(outside_buffer)
```

Figure 4.11 Implementation of Generating Area for Pseudo-Absence

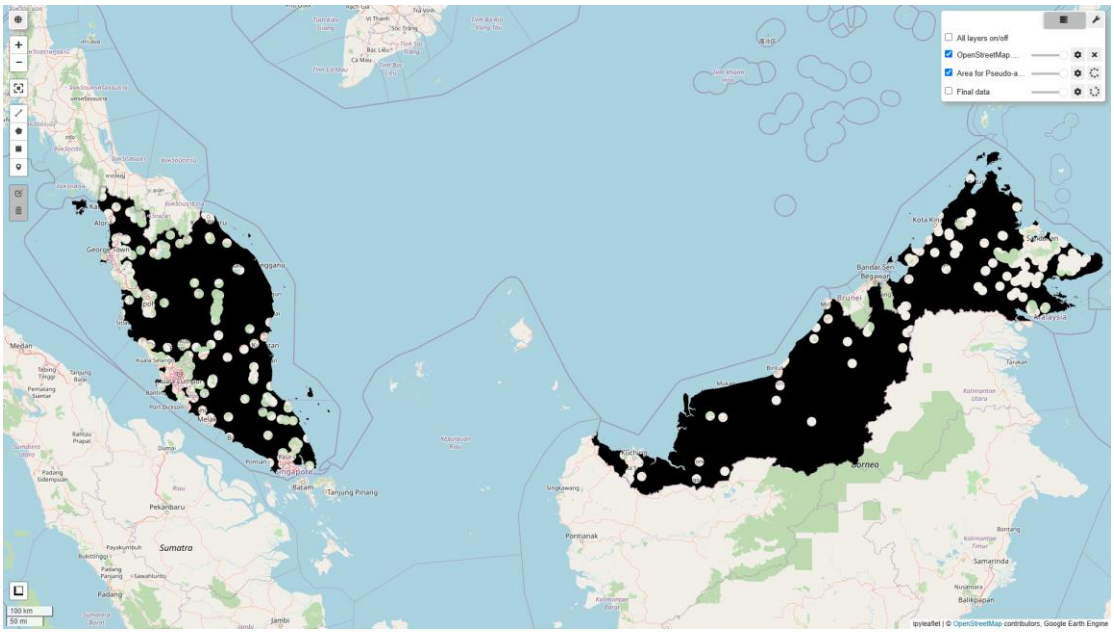


Figure 4.12 Area for Pseudo-Absence Generation

4.2 Data Splitting Method Evaluation Implementation

In this phase, the target is to determine the better sampling method between random sampling and spatial block sampling. The occurrence data is divided into West Malaysia (used for model training and testing) and East Malaysia (used as an independent evaluation dataset). Both sampling methods adopt a 7:3 proportion for splitting the data [49].

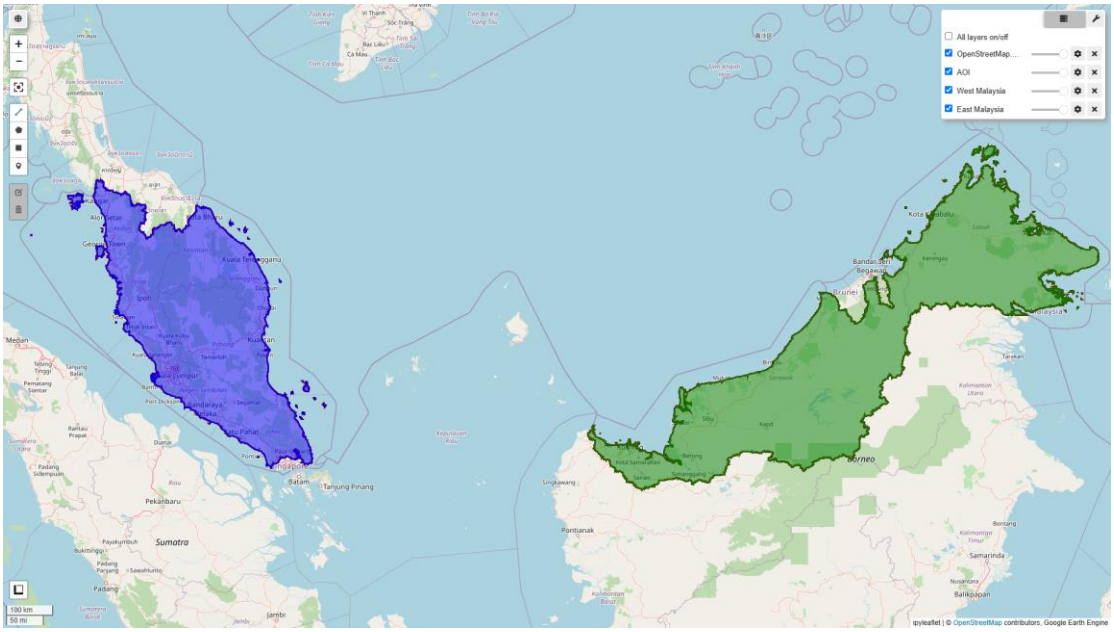


Figure 4.13 West Malaysia(Blue-colored Region) and East Malaysia(Green-colored Region)

4.2.1 Random Sampling

Random Sampling implementation is shown in Figure 4.14, it uses randomColumn method to split the data into training and testing sets based on a specified threshold (e.g., 0.7 for a 70:30 split). While it does not guarantee an exact 70:30 proportion, the split is very close, especially for large datasets.

```
def random_sdm(x, split):
    seed = ee.Number(x)

    presence_points = data.randomColumn(seed=seed).sort("random")
    presence_points = presence_points.map(lambda feature: feature.set("PresAbs", 1))
    tr_presence_points = presence_points.filter(ee.Filter.lt('random', split))
    te_presence_points = presence_points.filter(ee.Filter.gte('random', split))

    # Pseudo-absence
```

Figure 4.14 Implementation of Random Sampling

4.2.2 Spatial Block Sampling

To prepare the spatial block sampling for this study, a water mask was created to exclude water bodies. Each grid cell covers an area of 25 km². Multiple grids were generated to cover West Malaysia. Similar to the random sampling method,

randomColumn method was used for splitting, but in this case, it splits the grids into training and testing sets, rather than data points.

```
# Making grid for sampling method
watermask = terrain.select('elevation').gt(0)
Scale = 25000
grid = watermask.reduceRegions(
    collection=west_malaysia.coveringGrid(scale=Scale, proj='EPSG:4326'),
    reducer=ee.Reducer.mean()).filter(ee.Filter.neq('mean', None))
```

```
def grid_sdm(x, split):
    seed = ee.Number(x)

    # Random block division for training and validation
    rand_blk = ee.FeatureCollection(grid).randomColumn(seed=seed).sort("random")
    training_grid = rand_blk.filter(ee.Filter.lt("random", split))  # Grid for training
    testing_grid = rand_blk.filter(ee.Filter.gte("random", split))  # Grid for testing
```

Figure 4.15 Implementation of Spatial Block Sampling

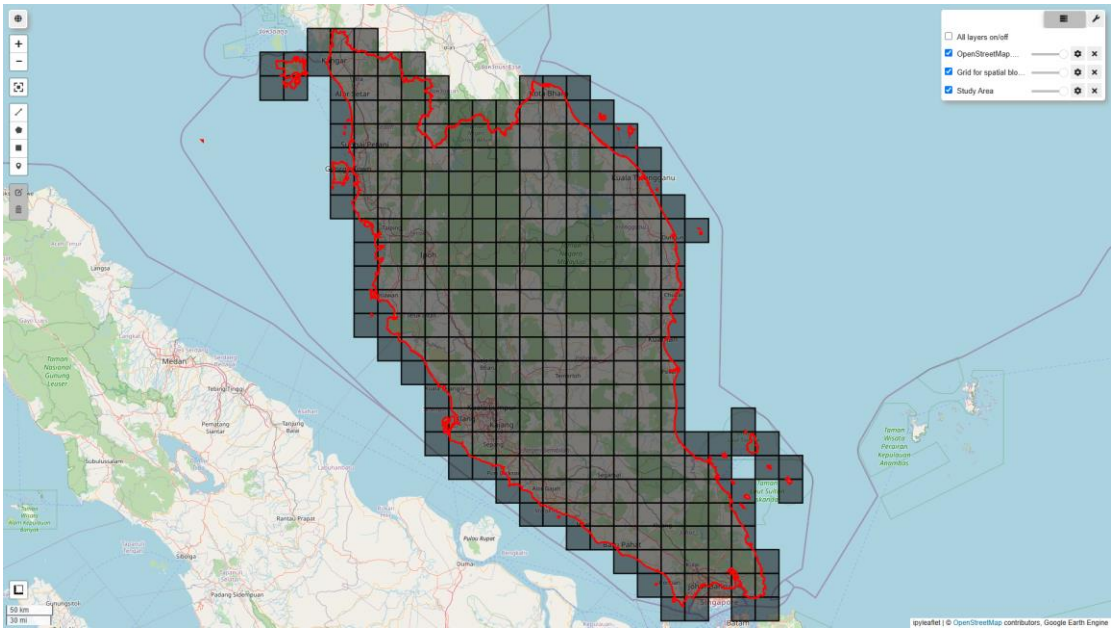


Figure 4.16 Spatial Grid Layout Covering West Malaysia (Black Grids)

4.2.3 Data Splitting Method Evaluation

To determine the best-performing data splitting method, the AUROC and AUCPR performance of each sampling method is compared using the RF model, as explained in Chapter 4.3.1, Model Architecture Implementation.

For AUROC performance:

1. **Random Sampling Method:**
    - As shown in Table 4.1, this method performs significantly better on the West Malaysia test set compared to the East Malaysia independent dataset.
    - As illustrated in Figure 4.17, it experiences a 41.14% performance decrease (AUROC) when evaluated on independent data (East Malaysia), confirming that models trained with random sampling struggle with extrapolation across different regions.
2. **Spatial Block Sampling Method:**
    - Shows a smaller performance gap between the West Malaysia test set and the East Malaysia evaluation dataset.
    - As shown in Figure 4.17, it experiences a 27.50% performance decrease in AUROC, indicating better generalization compared to the random sampling model.

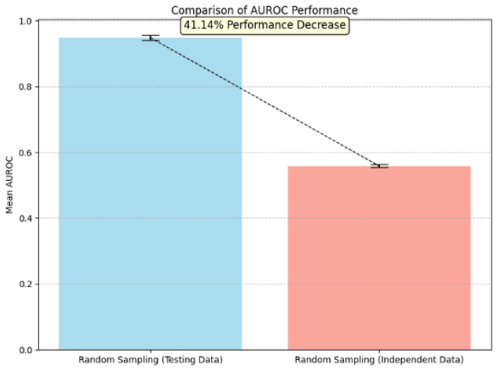
For AUCPR performance:

1. **Random Sampling Method:**
    - As illustrated in Figure 4.18, it performs well on the West Malaysia test set but demonstrates a significant 42.45% performance decrease when evaluated on East Malaysia.
    - This indicates poor generalization to unseen regions, highlighting the limitations of random sampling for cross-region predictions.
2. **Spatial Block Sampling Method:**

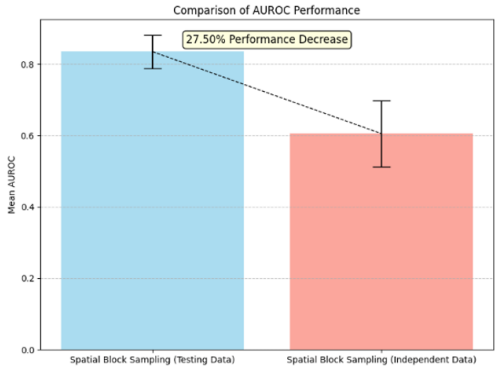
- As illustrated in Figure 4.18, it shows better generalization compared to the random sampling method, with a smaller 30.49% performance decrease when moving from West Malaysia (test set) to East Malaysia (independent evaluation dataset).
- The smaller performance gap reflects the improved ability of the spatial block sampling method to handle geographic variability and extrapolation.

The Table 4.2 summarizes the evaluation results of Random Sampling and Spatial Block Sampling methods based on AUROC and AUCPR metrics, including an assessment of standard deviation values. Models trained using Random Sampling performed well on the random sampling test set (AUROC: 0.9483 ± 0.0077, AUCPR: 0.9598 ± 0.0049), indicating high accuracy with minimal variability. However, on the independent set, they showed a significant performance drop (AUROC: 0.5582 ± 0.1310, AUCPR: 0.5524 ± 0.0975), with high standard deviations reflecting inconsistent generalization. In contrast, Spatial Block Sampling models displayed more stable performance on the independent set (AUROC: 0.6050 ± 0.0925, AUCPR: 0.5941 ± 0.0796) and slightly lower performance on the random sampling set (AUROC: 0.8345 ± 0.0462, AUCPR: 0.8547 ± 0.0434).

The results align with the findings of paper [29], which highlights that random sampling struggles to handle spatial autocorrelation, leading to an overestimation of the model's performance. In contrast, spatial block sampling helps to mitigate this issue by addressing spatial dependency. This study confirms that spatial block sampling is the more reliable method, as it provides better generalization and is less affected by performance drops when applied to new regions. This is particularly beneficial since this study aims to create habitat suitability maps under future environmental conditions, which inherently involve extrapolation.

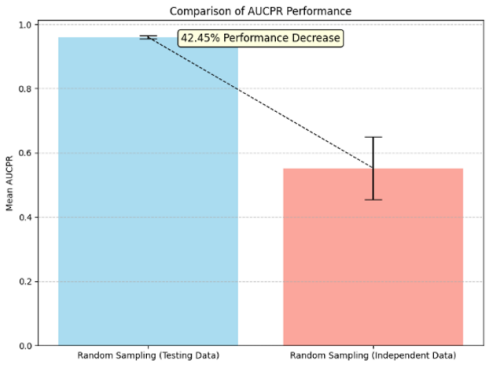


Random Sampling Method

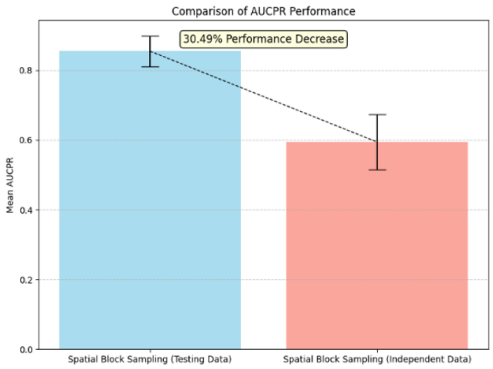


Spatial Block Sampling Method

Figure 4.17 AUROC Performance Comparison Between Different Data Splitting Methods



Random Sampling Method



Spatial Block Sampling Method

Figure 4.18 AUCPR Performance Comparison Between Different Data Splitting Methods

Table 4.2 Performance of Different Sampling Methods

	Evaluation Set	**AUROC (Mean±StdDev)**	**AUCPR (Mean±StdDev)**
Random Sampling	Random Sampling Set	0.9483 ± 0.0077	0.9598 ± 0.0049
Random Sampling	Independent Set	0.5582 ± 0.1310	0.5524 ± 0.0975
Spatial Block Sampling	Random Sampling Set	0.8345 ± 0.0462	0.8547 ± 0.0434
Spatial Block Sampling	Independent Set	0.6050 ± 0.0925	0.5941 ± 0.0796

Spatial Block Sampling has been shown to mitigate the bias introduced by spatial autocorrelation (SAC). However, according to literature review paper [29], there is no a standard rule of measurement of the optimal block size. To determine an appropriate block size for the occurrence data, the SAC range was quantified using cv_spatial_autocor() function form the blockCV package [50]. This function estimates the spatial autocorrelation structure by generating semivariograms of model residuals across spatially structured data and calculating the empirical range where autocorrelation becomes negligible. An example of a semivariogram generated from one of the splits is illustrated in Figure 4.19, highlighting the spatial dependence pattern observed in the data.

All data splits were evaluated using this function, returning empirical SAC ranges between 68.8 km and 77.5 km, with a mean range of 68.5 km, as presented in Table 4.3. Given that the initial experimental design considered a 25 km block size, with 25 km as the incremental unit, this study adopted a 75 km block size. The 75 km block width was chosen because it exceeds the mean SAC range, while 50 km would have fallen below it. This ensures that training and testing data points are sufficiently separated to strongly reduce the effects of SAC, thus improving model robustness and generalization.

Table 4.3 Spatial Autocorrelation Range Results

Dataset	SAC Range (meters)
Dataset 1	68808.45
Dataset 2	69193.98
Dataset 3	62311.07
Dataset 4	64517.68
Dataset 5	77465.46
Mean	68459.33

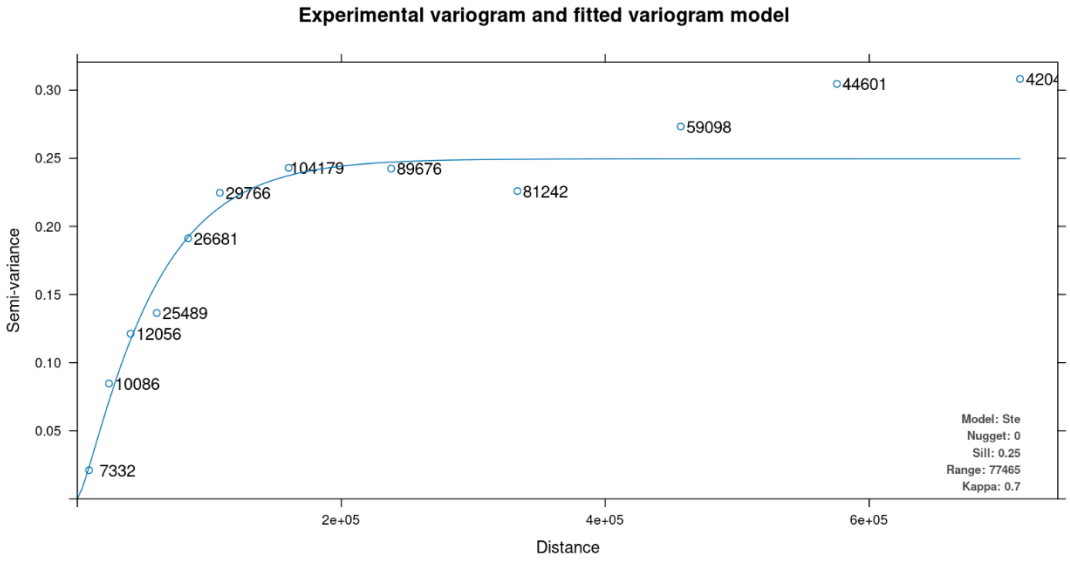


Figure 4.19 Semivariogram of Occurrence Data

4.3 Model Development and Evaluation Implementation

4.3.1 Model Architecture Implementation

A. Random Forest Models

In this study, the Random Forest (RF) models were developed using the RandomForestClassifier from the Scikit-learn library [51] in Python. The classifier was configured with 500 decision trees and the number of features considered at each split was set to 3.

Based on the methodology outlined in Chapter 3.1.1 pseudo-absence generation section, 5 RF models with same hyperparameters were developed to ensure the reliability of pseudo-absence generation [38]. For each iteration, a new set of pseudo-absence points was generated while maintaining the same presence points. Spatial block sampling was then applied to each dataset to ensure spatial independence between training and testing sets. This iterative process enhances the robustness of the model by accounting for variability in pseudo-absence generation and ensuring consistent performance across different datasets.

```
rf_model = RandomForestClassifier(
    n_estimators=500,
    random_state=42,
    max_features=3
    )
rf_model.fit(X_train, y_train)
```

Figure 4.20 Implementation of Random Forest Architecture

B. Multi-Layer Perceptron (MLP) Models

Following the RF development, a suite of five MLP models was trained on exactly the same presence and pseudo-absence datasets that had been generated with spatial block splitting.

To optimize the MLP architecture and training configurations, a custom randomized search procedure was implemented. The search space included:

- Number of neurons in each hidden layer (n1, n2, n3): Random integers between 32 and 128
- Learning rate (lr): Continuous values between 10^{-4} and 10^{-3} (log-uniform distribution)
- Dropout rate (dropout): Continuous values between 0.0 and 0.3
- Batch size: 16, 32, or 64
- Number of training epochs: 10, 20, or 30

A total of 100 random trials were conducted (n_iter = 100). For each trial, a model was trained and evaluated across all spatial folds, and the mean AUROC across folds was used as the selection criterion.

The best performing hyperparameter configuration identified from the random search is as follows:

- First hidden layer: 84 neurons
- Second hidden layer: 81 neurons
- Third hidden layer: 42 neurons
- Learning rate: 0.00053
- Dropout rate: 0.168
- Batch size: 16
- Epochs: 30

This optimized MLP configuration was subsequently used for final model training and evaluation. The implementation details of the custom randomized search procedure for MLP hyperparameter optimization are illustrated in Figure 4.21.

```python
def sample_params(space: dict) -> dict:
    params = {}
    for key, val in space.items():
        if isinstance(val, tuple):
            if key.startswith("n"):
                params[key] = random.randint(val[0], val[1])
            elif key == "lr":
                params[key] = 10 ** random.uniform(np.log10(val[0]), np.log10(val[1]))
            else:  # dropout
                params[key] = random.uniform(val[0], val[1])
        else:
            params[key] = random.choice(val)
    return params
```

```python
def make_mlp(input_dim: int, p: dict) -> tf.keras.Model:
    """Build an MLP model dynamically."""
    tf.keras.backend.clear_session()
    m = models.Sequential([
        layers.Input(shape=(input_dim,)),
        layers.Dense(p["n1"], activation="relu"),
        layers.Dropout(p["dropout"]),
        layers.Dense(p["n2"], activation="relu"),
        layers.Dropout(p["dropout"]),
        layers.Dense(p["n3"], activation="relu"),
        layers.Dropout(p["dropout"]),
        layers.Dense(1, activation="sigmoid"),
    ])
    m.compile(
        optimizer=tf.keras.optimizers.Adam(learning_rate=p["lr"]),
        loss="binary_crossentropy",
        metrics=[tf.keras.metrics.AUC(name='auroc')]
    )
    return m
```

```python
# --- Build & train model ------------------------------------------
model = make_mlp(X_train.shape[1], params)
model.fit(
    X_train, y_train,
    epochs=params["epochs"],
    batch_size=params["batch_size"],
    verbose=0
)

# --- Evaluate -----------------------------------------------------
y_prob = model.predict(X_test, verbose=0).ravel()
fold_aurocs.append(roc_auc_score(y_test, y_prob))
```

Figure 4.21 Implementation of Multi-layer Perceptron Architecture

4.3.2 Evaluation of AUROC and AUCPR

After training the RF models and MLP models, it was evaluated using the spatial block sampling test set to assess its performance in terms of AUROC and AUCPR. As illustrated in Figure 4.22, the evaluation metrics were computed using the scikit-learn library to ensure accurate measurement of AUROC and AUCPR values.

```
from sklearn.metrics import roc_auc_score, precision_recall_curve, auc
```

```
# Create a DataFrame
df = pd.DataFrame(data)

y_true = df['PresAbs']  # True labels (0 or 1)
y_scores = df['classification']  # Predicted probabilities

# AUC-ROC
auc_roc = roc_auc_score(y_true, y_scores)
# Precision-Recall Curve
precision, recall, _ = precision_recall_curve(y_true, y_scores)

# AUC-PR
auc_pr = auc(recall, precision)
return (auc_roc, auc_pr)
```

Figure 4.22 Implementation of AUROC and AUCPR

4.4 Environmental Impact and Habitat Analysis

4.4.1 Gini Impurity Analysis

Gini Impurity Reduction analysis was applied to evaluate the contribution of each environmental variable to reducing classification uncertainty within the Random Forest model [42, 46]. This method identifies the most influential predictors by measuring how much each variable improves the model's ability to classify data points at decision tree splits. A total of eight environmental variables were included in the analysis: bio01 (Annual Mean Temperature), bio02 (Mean Diurnal Range), bio03 (Isothermality), bio12 (Annual Precipitation), bio13 (Precipitation of Wettest Month), primf_median (Primary Forest Cover), secf_median (Secondary Forest Cover), and urban_median (Urban Area Cover).

The results in Figure 4.23 suggest that precipitation-related variables (bio12 and bio13) are the most influential factors in predicting species distribution of Ketupa,

underscoring the importance of water availability and seasonal rainfall patterns for habitat suitability. Temperature-related variables (bio01, bio02, and bio03) are relatively less influential compared to precipitation.

The chart in Figure 4.24 illustrates the Coefficient of Variation (CV) for the selected bioclimatic variables, highlighting differences in their variability across the study area. A total of 5000 points were randomly selected from Malaysia. Precipitation-related variables, such as bio13 (precipitation of wettest month) with the highest CV (26.99%) and bio12 (annual precipitation) with a CV of 22.72%, show significantly greater variability compared to temperature-related variables like bio03 (isothermality) with the lowest CV (4.51%). This suggests that precipitation variables may have a more dynamic influence on Ketupa distributions, while temperature variables tend to exhibit more stable spatial patterns, which explains why bio03 has the lowest variable importance.

This observation also aligns with Malaysia’s climatic patterns, as described in [52]. Malaysia's equatorial location ensures uniform temperatures with annual variations below 2°C, except during cold surges affecting the east coast, where variations remain below 3°C. Additionally, precipitation follows typical tropical region patterns but is influenced by seasonal wind flow and local topography. This further supports the observed variability in precipitation-related variables, such as bio12 and bio13, compared to the more stable temperature-related variables.

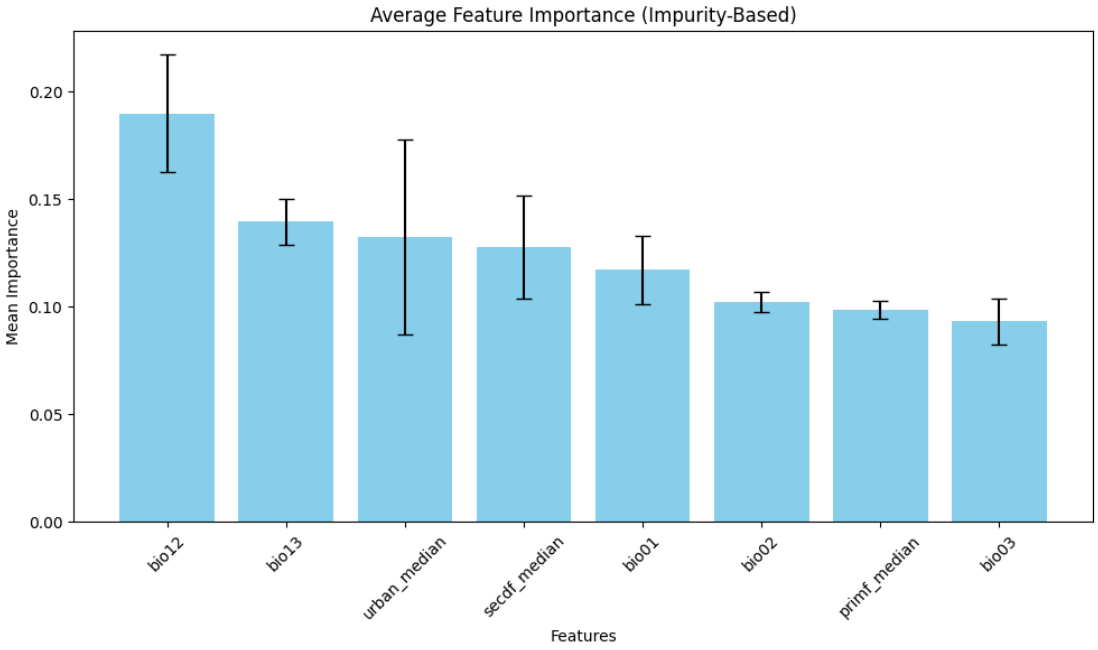


Figure 4.23 Mean Decrease in Impurity of Environmental Variables

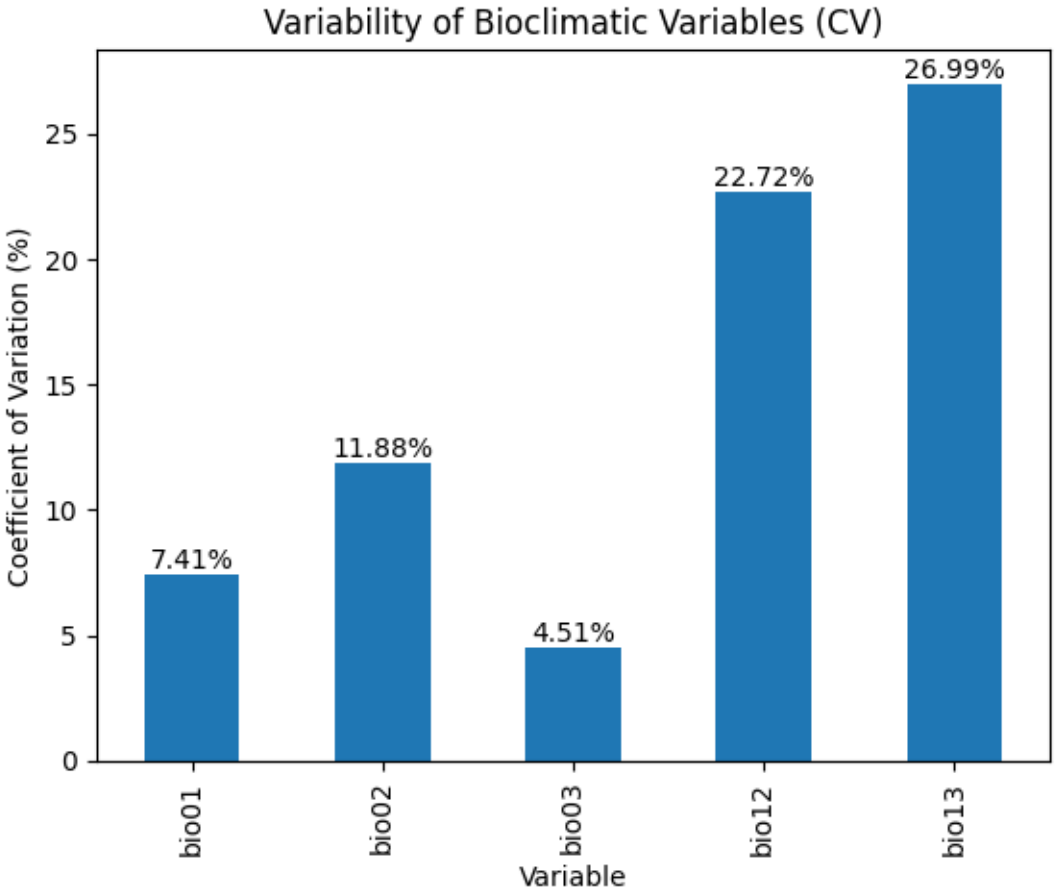


Figure 4.24 Coefficient of Variation for 5 Environmental Variables

4.4.2 Shapley Values Implementation

To implement Shapley Values analysis, this study use Shapley Additive Explanation (SHAP) library from [53], which provides a unified framework for interpreting the contributions of each feature to individual model predictions. Shapley Values assign each feature an importance value based on its marginal contribution across all possible feature combinations.

In this study, separate Shapley Values analyses were conducted for both Random Forest (RF) and Multi-Layer Perceptron (MLP) models. For RF models, the TreeExplainer was used for optimization of tree-based algorithms. For MLP models, a model-agnostic Explainer was applied. Bar plots were generated to visualize the average absolute Shapley values for each feature, allowing identification of the most significant predictors in species distribution modelling.

```
for model, X_train in zip(rf_models, rf_X_train):
    explainer = shap.TreeExplainer(model)
    shap_values = explainer(X_train)
    class_1_shap_values = shap_values.values[:, :, 1]

    shap_values.values = np.array(class_1_shap_values)
    shap.plots.bar(shap_values)

    rf_train_shap_values.append(class_1_shap_values)
    rf_train_feature_values.append(X_train.values)
    rf_train_individual_shap_values.append(shap_values)
```

```
for model, X_train, X_test, scaler in zip(
    mlp_models,
    mlp_X_train,
    mlp_X_test,
    mlp_scalers
    ):
    X_train_transformed = pd.DataFrame(
        scaler.transform(X_train),
        columns=X_train.columns,
        index=X_train.index
        )
    background = shap.sample(X_train_transformed, 100)
    explainer = shap.Explainer(model, background)
    shap_values = explainer(X_train_transformed)

    shap.plots.bar(shap_values)

    mlp_train_shap_values.append(shap_values.values)
    mlp_train_feature_values.append(X_train.values)
    mlp_train_individual_shap_values.append(shap_values)
    mlp_feature_names = X_train.columns  # set once
```

Figure 4.25 Implementation of Shapley Values Generation

4.4.3 Response Curve Implementation

To implement response curve analysis, the minimum, maximum, and mean values of each variable are extracted from the 5 training datasets. The target variable is then systematically varied across its range while all other variables are held constant at their mean values. This method isolates the effect of the target variable, allowing the model's response to changes in that variable to be assessed independently [47].

```
def plot_response_curves(models, datasets, feature_name, num_points=100, model_type="rf",train=True):
    """
    Generates response curves for a given feature across multiple models.
    """
    plt.figure(figsize=(10, 6))

    # Define feature range (min/max across all datasets)
    feature_min = min(X[feature_name].min() for X in datasets)
    feature_max = max(X[feature_name].max() for X in datasets)
    feature_range = np.linspace(feature_min, feature_max, num_points)

    all_predictions = []  # Store predictions from all models

    # Generate response curves for each model
    for i, (model, X) in enumerate(zip(models, datasets)):
        X_baseline = X.mean().to_frame().T  # Fix other features at their mean
        predictions = []

        for value in feature_range:
            X_temp = X_baseline.copy()
            X_temp[feature_name] = value  # Set feature to current value

            # Predict probability of class 1 (presence)
            if model_type == "mlp":
              pred = model.predict(X_temp,verbose=0)[0].flatten()
              # print(pred)
            if model_type == "rf":
              pred = model.predict_proba(X_temp)[0, 1]
              # print(pred)
            predictions.append(pred)
```

Figure 4.26 Implementation of Response Curve Generation

4.4.3 Habitat Suitability Map Implementation

Habitat suitability maps were generated by averaging the predicted probabilities from five model replications, each trained with distinct pseudo-absence datasets. This ensemble approach improves the robustness of predictions by reducing the variability introduced by random pseudo-absence generation. The final continuous suitability maps were then categorized into four habitat classes based on defined thresholds: high (0.75–1.00), medium (0.50–0.75), poor (0.25–0.50), and unsuitable (0.00–0.25) [48].

To evaluate changes in habitat suitability under future climate and land-use conditions, suitability maps were generated for three scenarios: current conditions, current conditions, and two future scenarios for the period 2061–2080—Shared Socioeconomic Pathway 2–4.5 (SSP2-4.5,) and Shared Socioeconomic Pathway 5–8.5 (SSP5-8.5). SSP2-4.5 represents a moderate-emissions pathway characterized by balanced socioeconomic development and land-use pressure, while SSP5-8.5 reflects a high-emissions, fossil-fuel-driven trajectory with intensive land-use expansion. Climate projections were derived from the Coupled Model Intercomparison Project Phase 6 (CMIP6) model known as Max Planck Institute Earth System Model (MPI-ESM1-2-HR), which provides high-resolution simulations with enhanced Earth system representation [20]. Corresponding land-use projections were obtained from the Land-Use Harmonization (LUH2) dataset [22, 54], with SSP2-4.5 aligned to the Model for Energy Supply Strategy Alternatives and their General Environmental Impact–Global Biosphere Management Model (MESSAGE-GLOBIOM) assumptions and SSP5-8.5 based on the Regional Model of Investment and Development–Model of Agricultural Production and its Impact on the Environment (REMIND-MAGPIE) scenario.

```
stacked_arrays = []
feature_columns = []
for feature_name, path in raster_files.items():
    with rasterio.open(path) as src:
        arr = src.read(1)
        stacked_arrays.append(arr)
        feature_columns.append(feature_name)
        profile = src.profile  # Save metadata from the first raster

stacked_array = np.stack(stacked_arrays, axis=0)
n_layers, n_rows, n_cols = stacked_array.shape

# Flatten and prepare for prediction
flat_data = stacked_array.reshape(n_layers, -1).T
flat_df = pd.DataFrame(flat_data, columns=feature_columns)
nan_mask = flat_df.isna().any(axis=1)
valid_pixels = ~nan_mask
X_valid = flat_df[valid_pixels]

# Predict using model ensemble and average
preds = np.array([model.predict_proba(X_valid)[:, 1] for model in model_list])
y_pred = preds.mean(axis=0)

# Reconstruct prediction map
prediction_map = np.full(flat_data.shape[0], np.nan)
prediction_map[valid_pixels.values] = y_pred
prediction_map = prediction_map.reshape(n_rows, n_cols)

# Categorize habitat suitability
category_map = np.full(prediction_map.shape, np.nan)
category_map[(prediction_map >= 0.00) & (prediction_map < 0.25)] = 0
category_map[(prediction_map >= 0.25) & (prediction_map < 0.5)] = 1
category_map[(prediction_map >= 0.5) & (prediction_map < 0.75)] = 2
category_map[(prediction_map >= 0.75)] = 3
```

Figure 4.27 Habitat Suitability Map Implementation

CHAPTER 5 EVALUATION AND DISCUSSION

5.1 Model Performance Comparison

The performance of the Multilayer Perceptron (MLP) and Random Forest (RF) models was evaluated using two primary metrics: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUCPR). Across all datasets, RF models consistently outperformed MLP models in AUROC, as summarized in Figure 5.1.

In terms of AUC-PR, RF models also outperformed MLP models in all datasets except for Dataset 3, where the MLP achieved a slightly higher AUC-PR score (0.897 compared to RF's 0.889), as shown in Figure 5.2. Despite this exception, Random Forest models demonstrated superior overall performance across the majority of folds.

Further analysis of the average performance metrics, summarized in Table 5.1, reinforces this conclusion. Random Forest achieved an average AUROC of 0.8784 ± 0.0289 and an average AUC-PR of 0.8718 ± 0.0300, while the MLP achieved an average AUROC of 0.8359 ± 0.0292 and an average AUC-PR of 0.8398 ± 0.0530. Notably, Random Forest not only achieved higher mean scores but also demonstrated lower standard deviations compared to the MLP, indicating more stable and consistent performance across datasets.

These findings are consistent with previous studies. Study [28] indicated that deep neural networks, particularly those with multiple hidden layers, tend to overfit and suffer from reduced generalization ability on smaller datasets, while also acknowledging that Random Forest models often maintain strong performance under such conditions. In addition, study [26] further supported the robustness of Random Forest across a wide range of dataset sizes among different machine learning models.

The dataset used in this study comprised 1,490 records, with approximately 1,000 samples allocated to the training set in each fold. This dataset size is considered relatively small for training deep neural networks, which may explain why the MLP

model with three hidden layers underperformed relative to the Random Forest models in this study.

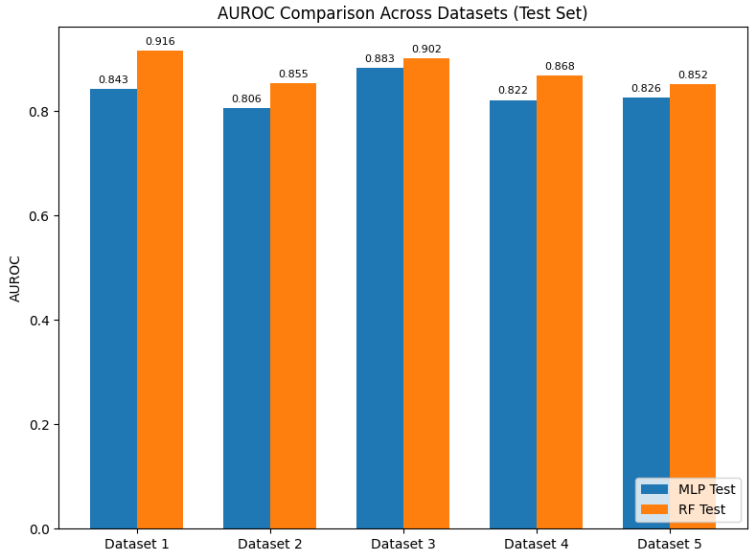


Figure 5.1 AUROC Performance of MLP and RF among All Test Set

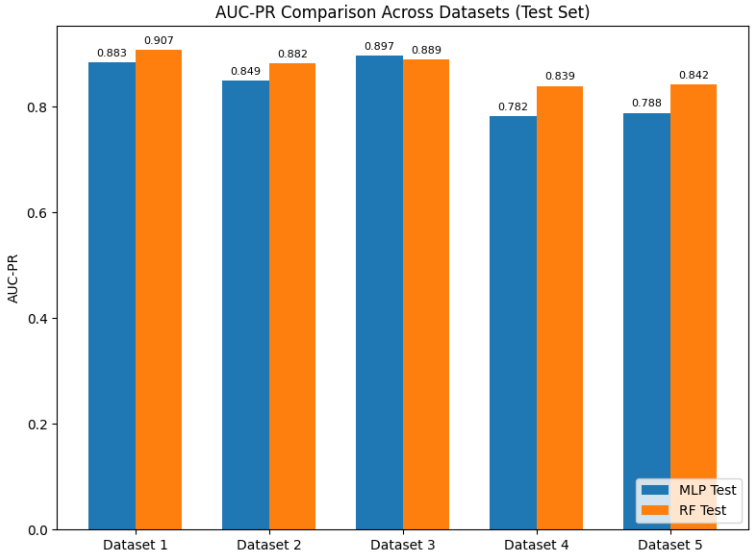


Figure 5.2 AUC-PR Performance of MLP and RF among All Test Set

Table 5.1 Average Performance of MLP and RF using Test Sets

Model	Average AUROC (Mean ± Std)	Average AUC-PR (Mean ± Std)
MLP	0.8359 ± 0.0292	0.8398 ± 0.0530
Random Forest	0.8784 ± 0.0289	0.8718 ± 0.0300

5.2 Predictor Importance Analysis

Although the RF models were ultimately selected as the final working models over the MLP models, predictor importance analysis was conducted on both architectures to allow for comparison and deeper interpretation.

The average absolute Shapley values across all datasets for both RF and MLP models are summarized in Figure 5.3. In both models, urban_median (proportion of urban land cover) emerged as the most impactful predictor, with average Shapley values of 0.10 for RF and 0.12 for MLP. This was followed by bio12 (Annual Precipitation), which achieved Shapley values of 0.08 in RF and 0.09 in MLP. In contrast, bio02 (Mean Diurnal Range) was the least important predictor in the RF model, with a Shapley value of 0.02, and the second least important in the MLP model, with a value of 0.04.

These findings differ from the Mean Decrease in Impurity (MDI) analysis presented earlier in Figure 4.23, where bio12 (Annual Precipitation) and bio13 (Precipitation of Wettest Month) were ranked as the most important variables, while primf_median (Primary Forest Cover) and bio03 (Isothermality) were among the least influential. One possible explanation for this discrepancy relates to the concept of feature cardinality. Cardinality refers to the number of unique values a feature can assume. The Scikit-learn documentation [55] notes that features with higher cardinality often receive artificially elevated importance scores in impurity-based methods such as MDI. This phenomenon is illustrated in Figure 5.4, which shows a positive relationship between Mean Decrease in Impurity scores and feature cardinality, with bio12 and bio13 exhibiting the highest cardinality among all predictors.

Consequently, while the MDI results provide some insights, they may not fully reflect true predictor importance due to this inherent bias. Greater emphasis is placed on the Shapley values analysis, which is less sensitive to feature cardinality and distribution effects. Based on the Shapley findings, urban_median and bio12 are confirmed as the two most important predictors, while bio02 is identified as the least important. Subsequent Response Curve analysis will therefore focus on interpreting the effects of these three key predictors on species distribution.

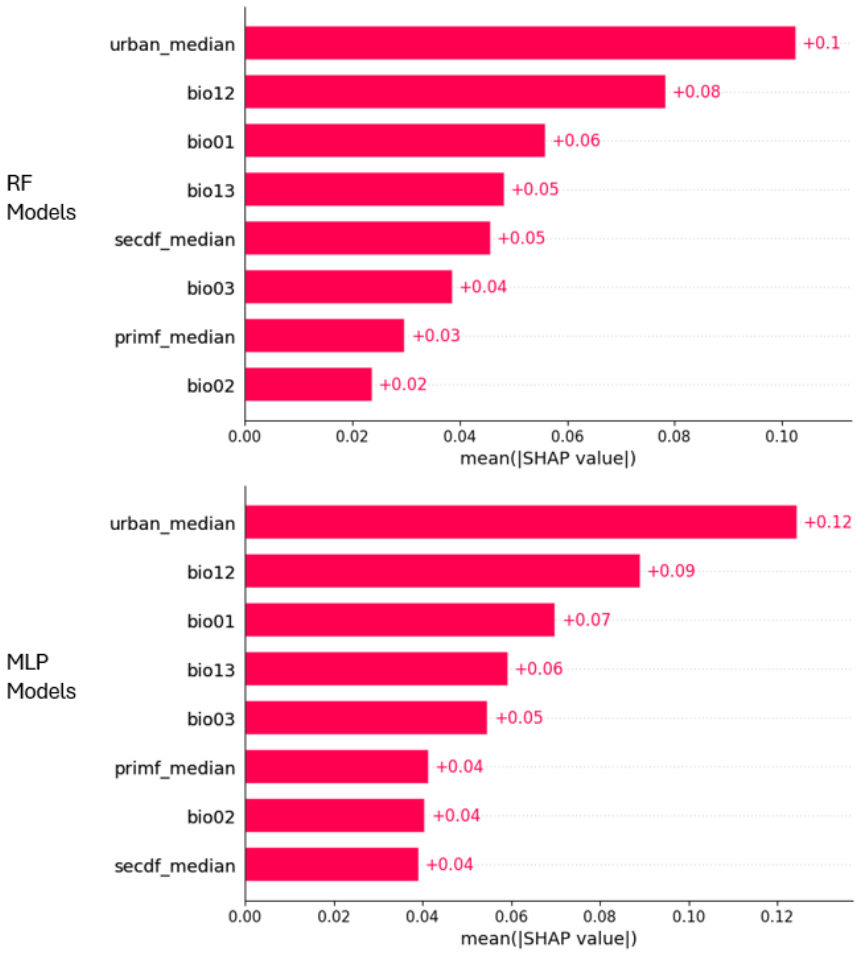


Figure 5.3 Mean Shapley Values of RF and MLP models

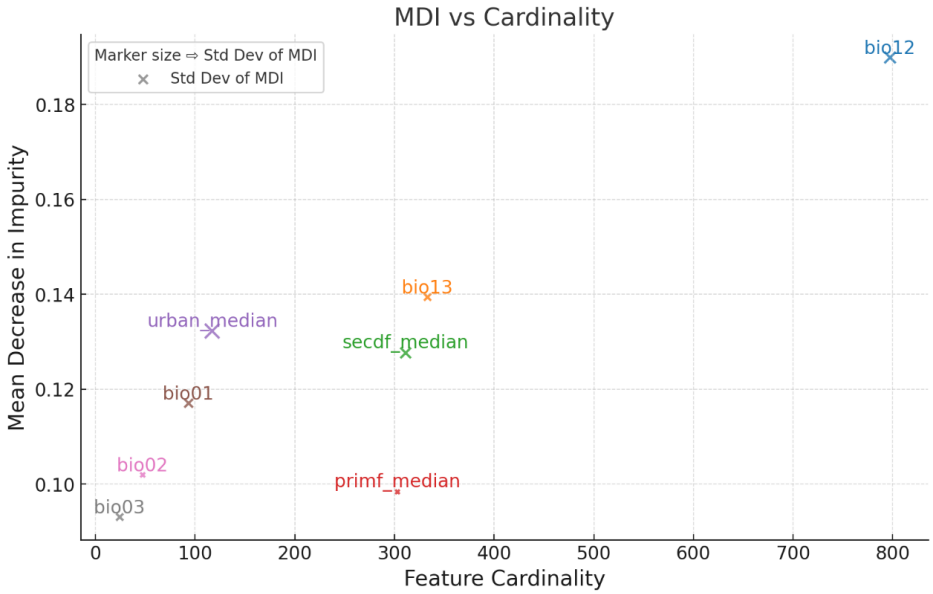


Figure 5.4 Relationship between Feature Cardinality and MDI

A. Urban_Median (Proportion of grid cell covered by urban areas)

Figure 5.5 illustrates the response curves for the predictor urban_median across five RF and five MLP models, with the black lines representing the mean response. The maximum observed value of urban_median in the dataset is approximately 0.7. Both models exhibit a strong positive relationship between urban cover and predicted species presence, particularly within the 0.0–0.3 range. This indicates that low-to-moderate levels of urbanization may provide suitable habitat conditions for species within the genus Ketupa, potentially due to factors such as increased prey availability, alternative roosting structures, or habitat heterogeneity. Similar trends have been reported for other urban-adapted owl species such as Strix aluco (Tawny Owl), which frequently occupies peri-urban landscapes that blend forest patches with built environments [56], and Ninox strenua (Powerful Owl), which has been observed persisting in urban areas where prey density remains sufficient [57].

Beyond an urban_median value of 0.3, the response curves plateau until predicted suitability of 0.7. This plateau likely reflects a lack of occurrence data in highly urbanized areas, as most high urban_median values in the dataset correspond to a small number of presence records located in the Kuala Lumpur region. Previous research [58, 59] has demonstrated that global biodiversity platforms such as GBIF are prone to spatial sampling biases, where data collection tends to concentrate near urban centers, while remote or less accessible habitats are often underrepresented. As a result, models may overestimate suitability in urban areas if trained on such spatially biased datasets without sufficient contrasting absence data.

In the present study, the distribution of urban_median values is visualized in Figure 5.6, revealing that most presence points fall within the 0.0–0.3 range, with a secondary cluster near 0.7. This uneven distribution, particularly the limited number of samples representing highly urbanized areas, restricts the model's ability to effectively differentiate habitat suitability across the full urbanization gradient. Consequently, the elevated suitability estimates beyond 0.3 should be interpreted with caution, as they may reflect underlying spatial bias rather than true ecological preference.

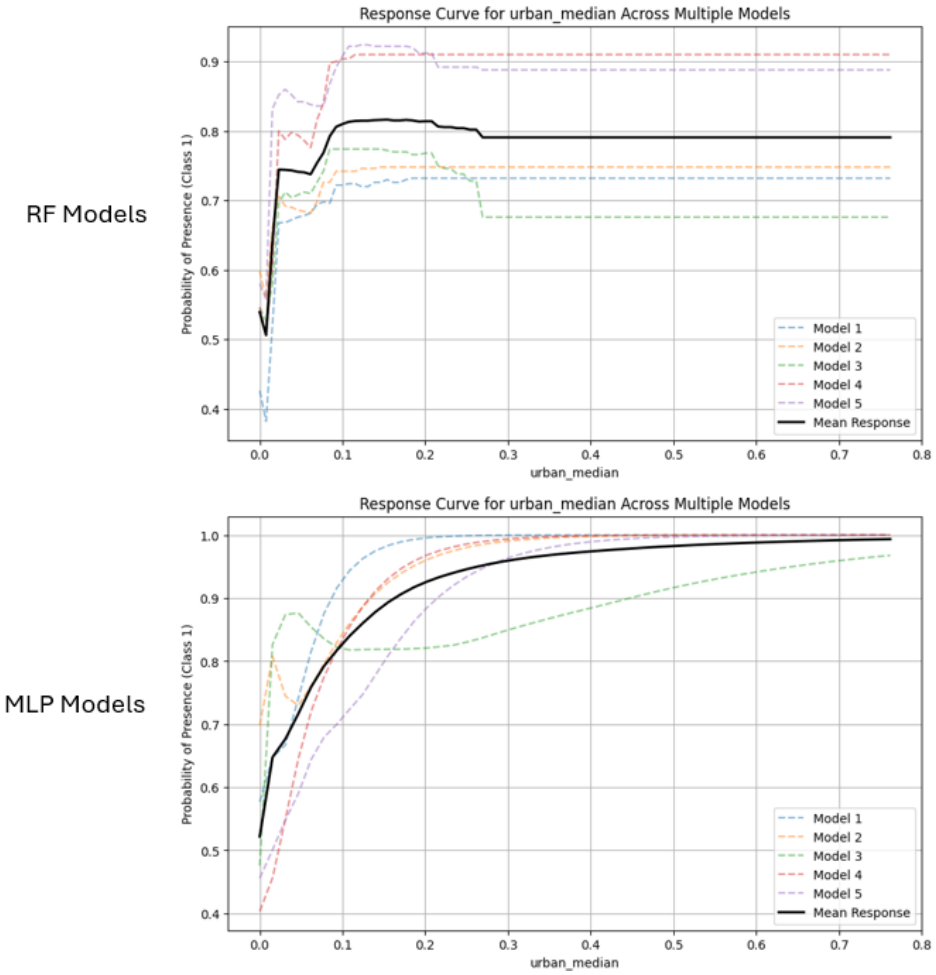


Figure 5.5 Response Curve for Urban_median of RF and MLP models

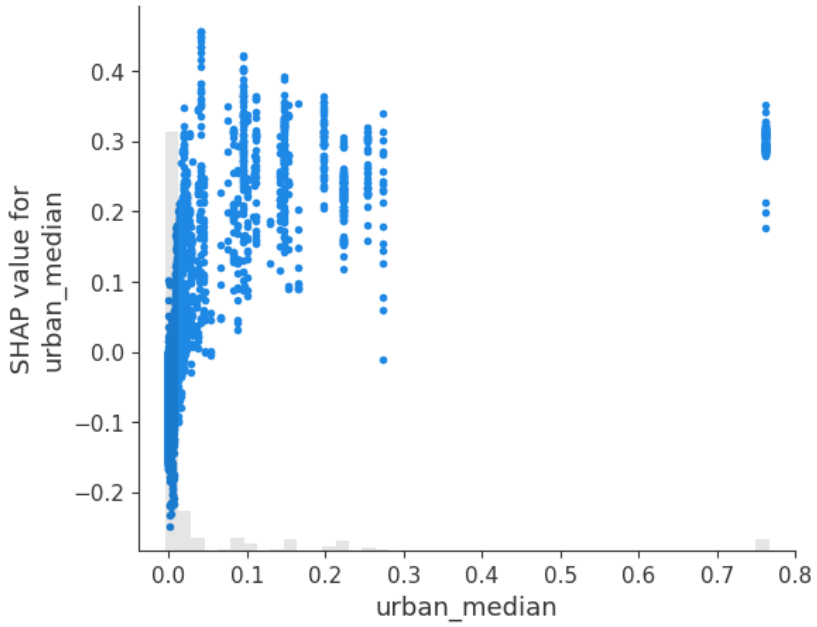


Figure 5.6 Relationship between Urban_median and Shapley values

B. Bio12 (Annual Precipitation)

The response curve for bio12 in Figure 5.7 aligns well with the ecological information provided about the Ketupa' habitat preferences. This genus is known to inhabit areas such as mangroves, tropical forests, and freshwater wetlands [5, 60]. Mangroves typically experience an annual precipitation range of 2000–3000 mm, while tropical montane forests have an annual precipitation range of 1000–3000 mm [61, 62]. This ecological evidence aligns with the results of the response curve. The response curve indicates that the habitat suitability is moderate for precipitation levels below 2000 mm, gradually increasing to peak suitability around 2500–3000 mm. This trend is further supported by the scatter plot of Shapley values and bio12 in Figure 5.8, where Shapley contributions are also highest within this precipitation range.

This pattern suggests that the conditions found in mangroves and the wetter regions of tropical montane forests likely represent core ecological niche for genus Ketupa. However, beyond 3000 mm, the response curve shows a sharp decline in predicted suitability. This indicates that extremely wet environments, such as lowland rainforests with precipitation exceeding 3500 mm, may not be as suitable for Ketupa. Additionally, areas with precipitation over 3500 mm are predominantly located in the inland regions of Sarawak, known for being the highest annual rainfall areas in Malaysia, particularly on hill slopes [52]. It is possible that genus Ketupa has been poorly surveyed in these inland Sarawak areas, leading to an apparent drop in suitability due to insufficient data from these regions [58, 59]. This could represent a modeling artifact rather than an ecological reality.

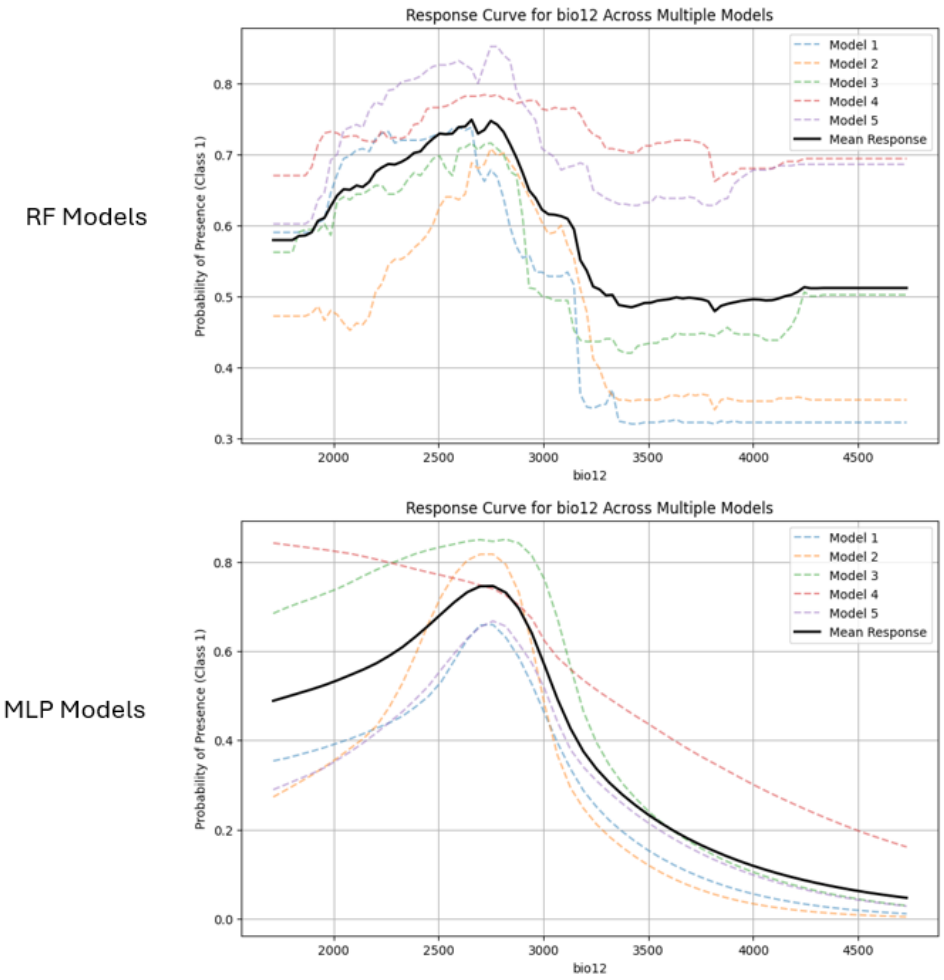


Figure 5.7 Response Curve for Bio12 (Annual Precipitation) of RF and MLP models

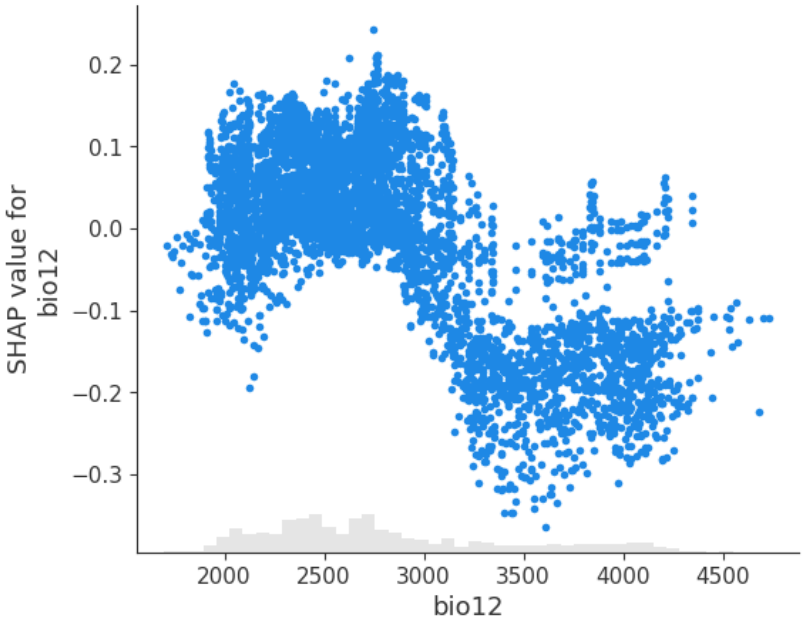


Figure 5.8 Relationship between Bio12 (Annual Precipitation) and Shapley Values

C. Bio02 (Mean Diurnal Range)

The response curves for bio02 (Mean Diurnal Range) exhibited relatively inconsistent patterns compared to other predictors, with notable differences observed between the Random Forest (RF) and Multi-Layer Perceptron (MLP) models. Both models indicated an optimum suitability range between 70 and 90 (equivalent to 7–9 °C), yet the overall curve shapes diverged significantly. RF models displayed a broad dome-shaped pattern, where suitability increased from approximately 0.65 at bio02 = 70 (7 °C) to a peak near 0.75 around bio02 = 75–100 (7.5–10 °C), followed by a gradual decline back to 0.65 around bio02 = 110 (11 °C). In contrast, MLP models produced a sharper, more pronounced peak, beginning with low suitability (~0.3), gradually increasing to ~0.7 between 80 and 90 (8–9 °C), and then decreasing and plateauing beyond 90 (9 °C).

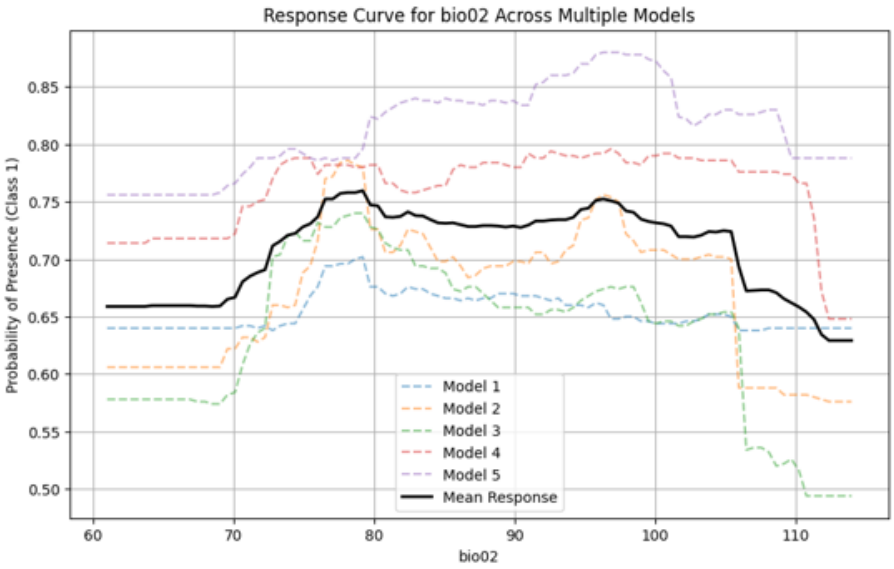
This shared peak suggests that moderate diel temperature variation (7–9 °C) may be associated with more favorable ecological conditions, potentially due to factors such as increased prey activity or reduced thermoregulatory stress. However, this does not necessarily indicate that genus Ketupa prefers habitats within this bio02 range or that bio02 is a key driver of their distribution. The high variability among individual model curves, especially within the MLP ensemble, indicates low importance in prediction. The inconsistencies in response patterns further suggest that bio02 has a minimal overall influence on habitat suitability, as small changes in training conditions result in substantial fluctuations in its modeled contribution.

These findings contrast sharply with studies from other regions. For instance, in the case of the Mexican Spotted Owl (Strix occidentalis lucida), bio02 was identified as the most important climatic predictor, contributing 44.8% to model gain in a MaxEnt analysis [63]. Suitability in that study peaked at 11 °C and dropped significantly outside the optimal 7–12 °C range.

This discrepancy highlights the critical need for region-specific species distribution models. Unlike owls in temperate regions, Ketupa genus in Malaysia inhabit tropical environments with relatively stable temperature regimes and narrower thermal fluctuations. According to [15], although SDMs have been widely applied across various regions, environmental conditions and ecological dynamics can differ significantly from one country to another. This underscores the importance of

conducting localized studies, even when similar research has already been performed elsewhere. Substantial gaps remain in SDM applications across different biological groups and geographic regions, particularly in highly biodiverse areas such as Malaysia

RF Models



MLP Models

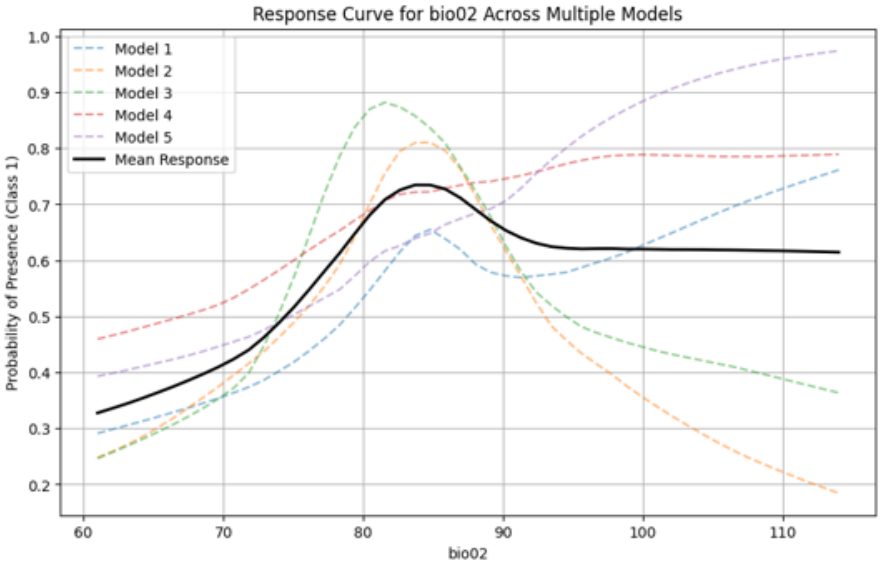


Figure 5.9 Response Curve for Bio02 (Mean Diurnal Range) of RF and MLP models

5.3 Habitat Suitability Map Comparison

Figure 5.10 presents habitat suitability maps for the genus Ketupa under current conditions and two future scenarios: SSP2-4.5 and SSP5-8.5 projected for the period 2061–2080. Figure 5.11 further summarizes the distribution of habitat suitability classes across these scenarios using a stacked bar chart. Under current conditions, 45.23% of the area is classified as unsuitable, 33.13% as poor, 14.42% as moderate, and 7.22% as high suitability. In contrast, both future scenarios reveal a substantial shift: in SSP2-4.5, unsuitable areas disappear entirely, while poor and moderate suitability increase to 48.07% and 51.47%, respectively. High suitability drops sharply to just 0.46%. A similar trend is seen in SSP5-8.5, where poor and moderate suitability rise to 51.63% and 48.36%, and high suitability declines further to only 0.01%.

While the disappearance of unsuitable areas may initially appear positive, this change does not indicate improved habitat conditions for Ketupa. The sharp decline in high suitability zones across both SSP scenarios suggests the loss of optimal habitats. Although the increase in moderate and poor suitability reflects an overall shift toward the center of the suitability scale, this flattening of the gradient diminishes the ecological distinction between favorable and unfavorable regions, effectively compressing the species' niche space.

One possible explanation for this convergence is the contrasting influence of future urbanization and climate variables. In both SSP2-4.5 and SSP5-8.5, urban cover is projected to increase, and urban_median is the most influential variable in the model—positively associated with habitat suitability. This urban expansion likely elevates predicted suitability across large areas. However, climatic conditions are projected to worsen, particularly in terms of precipitation and temperature variability, which exert negative effects on suitability through variables such as bio12 and bio02. These opposing influences may cancel each other out, pushing predictions away from both extreme unsuitable and highly suitable, clustering them around moderate values.

Overall, despite the apparent reduction in unsuitable classifications, the near-complete loss of high suitability areas and the compression of predictions into intermediate classes suggest potential habitat degradation and loss of ecological optimality. These findings do not support an optimistic outlook for genus Ketupa under future conditions. Instead, they highlight the importance of cautious interpretation and reinforce the need

for early conservation planning, especially in regions currently identified as high-quality habitats that may be at risk of decline.

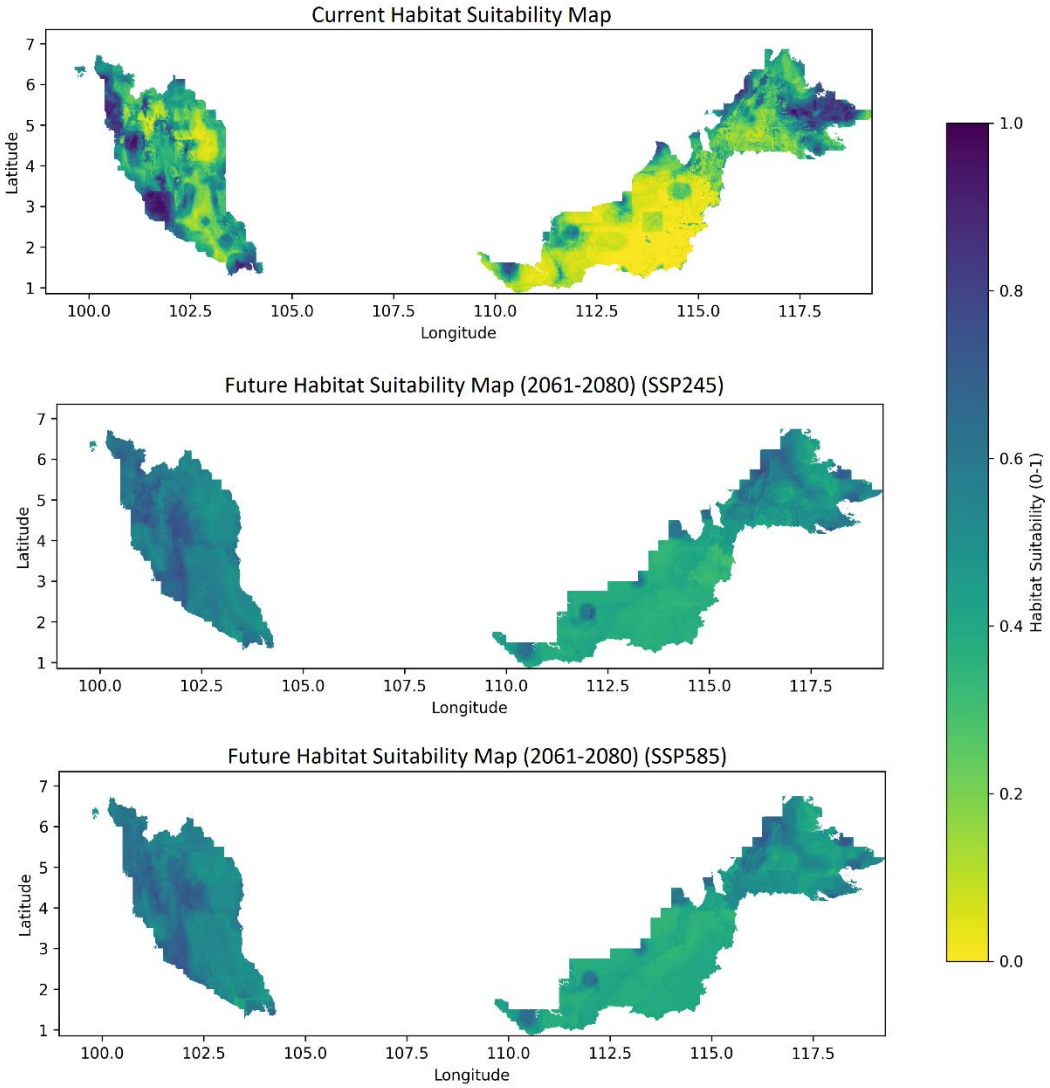


Figure 5.10 Habitat Suitability Maps for both scenarios

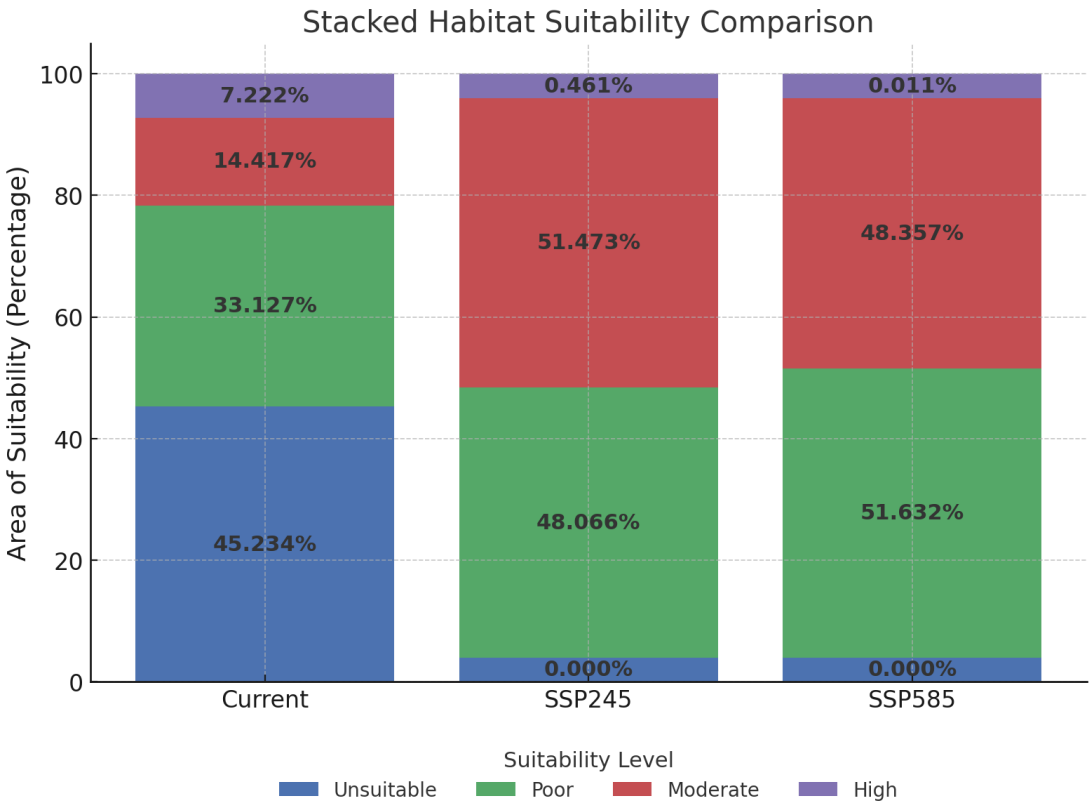


Figure 5.11 Stacked Habitat Suitability Comparison

CHAPTER 6 CONCLUSION AND FUTURE WORKS

6.1 Conclusion

This study focused on modelling the habitat suitability of the Ketupa genus in Malaysia, addressing key challenges in species distribution modelling (SDM), such as spatial autocorrelation and sampling bias. By comparing Random Forest (RF) and Multi-Layer Perceptron (MLP) models, the study aimed to establish a robust and interpretable SDM framework tailored to a tropical biodiversity context. To further enhance prediction reliability and ecological understanding, environmental impact analyses were conducted using three complementary techniques: response curves, Gini impurity reduction, and Shapley values.

Among the data splitting strategies, spatial block sampling outperformed random sampling by mitigating spatial dependency and improving model generalization across independent geographic regions. In terms of predictive performance, Random Forest achieved superior results, with an average AUROC of 0.8784 ± 0.0289 and AUC-PR of 0.8718 ± 0.0300, compared to MLP's AUROC of 0.8359 ± 0.0292 and AUC-PR of 0.8398 ± 0.0530.

The predictor importance analysis consistently identified urban area cover (urban_median) as the most influential environmental variable, followed by bio12 (annual precipitation). Conversely, bio02 (mean diurnal range) was found to be the least significant, a result that contrasts with studies on temperate-region owls such as the Mexican Spotted Owl [63], where bio02 was dominant. This highlights the importance of region-specific SDMs, particularly in tropical environments like Malaysia, where ecological responses to environmental variables may differ significantly.

Finally, habitat suitability mapping revealed a worrying trend under future climate and land-use scenarios for the period 2061–2080. While 7.22% of the landscape is currently categorized as high suitability, this proportion drops drastically to 0.46% under SSP2-4.5 and further to 0.01% under SSP5-8.5. These results indicate a potential loss of optimal habitats and a narrowing of suitable ecological conditions, despite increased urbanization. Overall, the findings suggest that Ketupa species in Malaysia may face increasing habitat constraints in the future, reinforcing the need for early conservation planning and more localized ecological research.

6.2 Future Works

As data scarcity and sampling bias remain persistent challenges in biological research, including species distribution modelling (SDM) [28, 58, 59], emerging deep learning techniques such as zero-shot and few-shot learning offer promising solutions. Zero-shot learning leverages ecological representations learned from a broad range of species to infer plausible habitat suitability patterns, even in the absence of direct occurrence data for the target species [64]. Similarly, few-shot SDM approaches, such as the Few-Shot Spatial Implicit Neural Representations (FS-SINR) proposed by Lange et al. [65], have demonstrated strong potential for predicting habitat suitability using only a limited number of observations. FS-SINR, a Transformer-based model, encodes geographic coordinates and optional metadata into spatially aware embeddings, allowing for accurate inference without requiring model retraining. These approaches combine the flexibility of deep learning with the ability to handle sparse datasets, making them particularly suitable for modelling the distributions of rare or newly described species.

Although this study quantified the projected decline in high-suitability habitats under Shared Socioeconomic Pathway 2–4.5 (SSP2-4.5) and Shared Socioeconomic Pathway 5–8.5 (SSP5-8.5) for the period 2061–2080, the specific environmental drivers responsible for these changes remain uncertain. Future studies should consider further analysis on impact of variables, in which individual environmental variables are held constant between current and future projections to isolate their independent effects on habitat suitability. In addition, model explainability tools such as differential SHAP (Shapley values) analysis across timeframes can help identify which variable shifts contribute most significantly to habitat loss. Combining this approach with spatial trend analysis of climate and land-use data may offer more mechanistic insight into the causes of future distributional shifts and further improve model interpretability.

REFERENCES

[1] C. Bellard, C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp, 'Impacts of climate change on the future of biodiversity', *Ecology Letters*, vol. 15, no. 4, pp. 365–377, Apr. 2012, doi: 10.1111/j.1461-0248.2011.01736.x.

[2] Ipcc, *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*, 1st ed. Cambridge University Press, 2022. doi: 10.1017/9781009157940.

[3] Hong, Shiao-Yu, Sun, Yuan-Hsun, Wu, Hsin-Ju, and Chen, Chao-Chieh, 'Spatial distribution of the Tawny Fish Owl Ketupa flavipes shaped by natural and man-made factors in Taiwan', *Forktail*, vol. 29, pp. 48–51, 2013.

[4] M. Arszulowicz, 'Ketupa zeylonensis', Animal Diversity Web. Accessed: Jan. 24, 2025. [Online]. Available: https://www.thevibes.com/articles/news/100693/sarawak-govt-confirms-construction-of-three-more-hydroelectric-dams

[5] Thai National Parks, 'Buffy Fish Owl', Thai National Parks. Accessed: Dec. 04, 2024. [Online]. Available: https://www.thainationalparks.com/species/buffy-fish-owl

[6] MyBIS, 'Malaysia Biodiversity Information System', Malaysia Biodiversity Information System. Accessed: Jan. 24, 2025. [Online]. Available: https://www.mybis.gov.my/

[7] J. Kitzes and R. Shirley, 'Estimating biodiversity impacts without field surveys: A case study in northern Borneo', *Ambio*, vol. 45, no. 1, pp. 110–119, Feb. 2016, doi: 10.1007/s13280-015-0683-3.

[8] S. Then, 'Sarawak govt confirms construction of three more hydroelectric dams', The Vibes. [Online]. Available: https://www.thevibes.com/articles/news/100693/sarawak-govt-confirms-construction-of-three-more-hydroelectric-dams

[9] M. C. Urban, 'Accelerating extinction risk from climate change', *Science*, vol. 348, no. 6234, pp. 571–573, May 2015, doi: 10.1126/science.aaa4984.

[10] 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: 10.1016/j.ecolmodel.2005.03.026.

[11] J. Elith and J. R. Leathwick, 'Species Distribution Models: Ecological Explanation and Prediction Across Space and Time', *Annu. Rev. Ecol. Evol. Syst.*, vol. 40, no. 1, pp. 677–697, Dec. 2009, doi: 10.1146/annurev.ecolsys.110308.120159.

[12] F. Recknagel, 'Applications of machine learning to ecological modelling', *Ecological Modelling*, vol. 146, no. 1–3, pp. 303–310, Dec. 2001, doi: 10.1016/S0304-3800(01)00316-7.

[13] J. Zhang and S. Li, 'A Review of Machine Learning Based Species' Distribution Modelling', in *2017 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, Wuhan: IEEE, Dec. 2017, pp. 199–206. doi: 10.1109/ICIICII.2017.76.

[14] R. E. Matadamas, P. L. Enríquez, L. Guevara, and A. G. Navarro-Sigüenza, 'Stairway to extinction? Influence of anthropogenic climate change on distribution patterns of montane Strigiformes in Mesoamerica', *ACE*, vol. 17, no. 2, p. art37, 2022, doi: 10.5751/ACE-02314-170237.

[15] M. Fois, A. Cuena-Lombraña, G. Fenu, and G. Bacchetta, 'Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions', *Ecological Modelling*, vol. 385, pp. 124–132, Oct. 2018, doi: 10.1016/j.ecolmodel.2018.07.018.

[16] M. Ryo, B. Angelov, S. Mammola, J. M. Kass, B. M. Benito, and F. Hartig, 'Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models', *Ecography*, vol. 44, no. 2, pp. 199–205, Feb. 2021, doi: 10.1111/ecog.05360.
[17] S. Christin, É. Hervet, and N. Lecomte, 'Applications for deep learning in ecology', *Methods Ecol Evol*, vol. 10, no. 10, pp. 1632–1644, Oct. 2019, doi: 10.1111/2041-210X.13256.
[18] L. Brugere, Y. Kwon, A. E. Frazier, and P. Kedron, 'Improved prediction of tree species richness and interpretability of environmental drivers using a machine learning approach', *Forest Ecology and Management*, vol. 539, p. 120972, Jul. 2023, doi: 10.1016/j.foreco.2023.120972.
[19] GBIF.Org User, 'Occurrence Download'. The Global Biodiversity Information Facility, p. 2493397, 2024. doi: 10.15468/DL.RM4MYH.
[20] S. E. Fick and R. J. Hijmans, 'WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas', *Intl Journal of Climatology*, vol. 37, no. 12, pp. 4302–4315, Oct. 2017, doi: 10.1002/joc.5086.
[21] T. G. Farr *et al.*, 'The Shuttle Radar Topography Mission', *Reviews of Geophysics*, vol. 45, no. 2, p. 2005RG000183, Jun. 2007, doi: 10.1029/2005RG000183.
[22] G. Hurtt *et al.*, 'Harmonization of Global Land Use Change and Management for the Period 850-2015'. Earth System Grid Federation, 2019. doi: 10.22033/ESGF/INPUT4MIPS.10454.
[23] A. Guisan, T. C. Edwards, and T. Hastie, 'Generalized linear and generalized additive models in studies of species distributions: setting the scene', *Ecological Modelling*, vol. 157, no. 2–3, pp. 89–100, Nov. 2002, doi: 10.1016/S0304-3800(02)00204-1.
[24] J. R. Leathwick, J. Elith, and T. Hastie, 'Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions', *Ecological Modelling*, vol. 199, no. 2, pp. 188–196, Nov. 2006, doi: 10.1016/j.ecolmodel.2006.05.022.
[25] E. Chollet Ramampiandra, A. Scheidegger, J. Wydler, and N. Schuwirth, 'A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation', *Ecological Modelling*, vol. 481, p. 110353, Jul. 2023, doi: 10.1016/j.ecolmodel.2023.110353.
[26] R.-Y. Duan, X.-Q. Kong, M.-Y. Huang, W.-Y. Fan, and Z.-G. Wang, 'The Predictive Performance and Stability of Six Species Distribution Models', *PLoS ONE*, vol. 9, no. 11, p. e112764, Nov. 2014, doi: 10.1371/journal.pone.0112764.
[27] R. Zbinden, B. Kellenberger, L. H. Hughes, and D. Tuia, 'Exploring the potential of neural networks for Species Distribution Modeling', presented at the ICLR 2023 Workshop on Tackling Climate Change with Machine Learning, 2023. [Online]. Available: https://www.climatechange.ai/papers/iclr2023/46
[28] D. J. Benkendorf and C. P. Hawkins, 'Effects of sample size and network depth on a deep learning approach to species distribution modeling', *Ecological Informatics*, vol. 60, p. 101137, Nov. 2020, doi: 10.1016/j.ecoinf.2020.101137.
[29] D. R. Roberts *et al.*, 'Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure', *Ecography*, vol. 40, no. 8, pp. 913–929, Aug. 2017, doi: 10.1111/ecog.02881.
[30] K. H. D. Tang, 'Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations', *Science of The Total Environment*, vol. 650, pp. 1858–1871, Feb. 2019, doi: 10.1016/j.scitotenv.2018.09.316.

[31] P. A. Harrison, P. M. Berry, N. Butt, and M. New, 'Modelling climate change impacts on species' distributions at the European scale: implications for conservation policy', *Environmental Science & Policy*, vol. 9, no. 2, pp. 116–128, Apr. 2006, doi: 10.1016/j.envsci.2005.11.003.
[32] C. C. Lee and S. C. Sheridan, 'Trends in weather type frequencies across North America', *npj Clim Atmos Sci*, vol. 1, no. 1, p. 41, Nov. 2018, doi: 10.1038/s41612-018-0051-7.
[33] Y. Guo, Z. Zhao, F. Zhu, and X. Li, 'Climate change may cause distribution area loss for tree species in southern China', *Forest Ecology and Management*, vol. 511, p. 120134, May 2022, doi: 10.1016/j.foreco.2022.120134.
[34] X. Ying, 'An Overview of Overfitting and its Solutions', *J. Phys.: Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
[35] J. A. Lee-Yaw, J. L. McCune, S. Pironon, and S. N. Sheth, 'Species distribution models rarely predict the biology of real populations', *Ecography*, vol. 2022, no. 6, p. e05877, Jun. 2022, doi: 10.1111/ecog.05877.
[36] C. Hari *et al.*, 'Future climate and land use change will equally impact global terrestrial vertebrate diversity', Dec. 16, 2024. doi: 10.1101/2024.12.13.627895.
[37] C. Seo, J. H. Thorne, L. Hannah, and W. Thuiller, 'Scale effects in species distribution models: implications for conservation planning under climate change', *Biol. Lett.*, vol. 5, no. 1, pp. 39–43, Feb. 2009, doi: 10.1098/rsbl.2008.0476.
[38] M. Barbet-Massin, F. Jiguet, C. H. Albert, and W. Thuiller, 'Selecting pseudo-absences for species distribution models: how, where and how many?', *Methods Ecol Evol*, vol. 3, no. 2, pp. 327–338, Apr. 2012, doi: 10.1111/j.2041-210X.2011.00172.x.
[39] R. Séchaud *et al.*, 'Behaviour-specific habitat selection patterns of breeding barn owls', *Mov Ecol*, vol. 9, no. 1, p. 18, Dec. 2021, doi: 10.1186/s40462-021-00258-6.
[40] K. Pearson, 'Note on Regression and Inheritance in the Case of Two Parents', *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.
[41] P. Brun *et al.*, 'Model complexity affects species distribution projections under climate change', *Journal of Biogeography*, vol. 47, no. 1, pp. 130–142, Jan. 2020, doi: 10.1111/jbi.13734.
[42] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[43] T. Fawcett, 'An introduction to ROC analysis', *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
[44] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.
[45] H. R. Sofaer, J. A. Hoeting, and C. S. Jarnevich, 'The area under the precision-recall curve as a performance metric for rare binary events', *Methods Ecol Evol*, vol. 10, no. 4, pp. 565–577, Apr. 2019, doi: 10.1111/2041-210X.13140.
[46] J.-H. Hur, S.-Y. Ihm, and Y.-H. Park, 'A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing', *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–13, 2017, doi: 10.1155/2017/6817627.
[47] M. Bazzichetto *et al.*, 'Sampling strategy matters to accurately estimate response curves' parameters in species distribution models', *Global Ecol Biogeogr*, vol. 32, no. 10, pp. 1717–1729, Oct. 2023, doi: 10.1111/geb.13725.
[48] N. Z. Ab Lah, Z. Yusop, M. Hashim, J. Mohd Salim, and S. Numata, 'Predicting the Habitat Suitability of Melaleuca cajuputi Based on the MaxEnt Species Distribution Model', *Forests*, vol. 12, no. 11, p. 1449, Oct. 2021, doi: 10.3390/f12111449.

[49] E. M. Baglaeva, A. P. Sergeev, A. V. Shichkin, and A. G. Buevich, 'The Effect of Splitting of Raw Data into Training and Test Subsets on the Accuracy of Predicting Spatial Distribution by a Multilayer Perceptron', *Math Geosci*, vol. 52, no. 1, pp. 111–121, Jan. 2020, doi: 10.1007/s11004-019-09813-9.

[50] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita, 'BLOCK CV : An R package for generating spatially or environmentally separated folds for *k* -fold cross-validation of species distribution models', *Methods Ecol Evol*, vol. 10, no. 2, pp. 225–232, Feb. 2019, doi: 10.1111/2041-210X.13107.

[51] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[52] Malaysian Meteorological Department, 'Climate of Malaysia', Dec. 2024. Accessed: Dec. 04, 2024. [Online]. Available: https://www.met.gov.my/en/pendidikan/iklim-malaysia/

[53] S. M. Lundberg and S.-I. Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[54] G. Hurtt *et al.*, 'Harmonization of Global Land Use Change and Management for the Period 2015-2300'. Earth System Grid Federation, 2019. doi: 10.22033/ESGF/INPUT4MIPS.10468.

[55] F. Pedregosa *et al.*, 'Feature importance evaluation with Random Forests'. [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

[56] N. Pagaldai, J. Arizaga, M. V. Jiménez-Franco, and I. Zuberogoitia, 'Colonization of Urban Habitats: Tawny Owl Abundance Is Conditioned by Urbanization Structure', *Animals*, vol. 11, no. 10, p. 2954, Oct. 2021, doi: 10.3390/ani11102954.

[57] B. Isaac, R. Cooke, D. Ierodiaconou, and J. White, 'Does urbanization have the potential to create an ecological trap for powerful owls (Ninox strenua)?', *Biological Conservation*, vol. 176, pp. 1–11, Aug. 2014, doi: 10.1016/j.biocon.2014.04.013.

[58] J. Beck, M. Böller, A. Erhardt, and W. Schwanghart, 'Spatial bias in the GBIF database and its effect on modeling species' geographic distributions', *Ecological Informatics*, vol. 19, pp. 10–15, Jan. 2014, doi: 10.1016/j.ecoinf.2013.11.002.

[59] D. E. Bowler *et al.*, 'Temporal trends in the spatial bias of species occurrence records', *Ecography*, vol. 2022, no. 8, p. e06219, Aug. 2022, doi: 10.1111/ecog.06219.

[60] Bird Society of Singapore, 'Buffy Fish Owl', Singapore Birds. Accessed: Dec. 04, 2024. [Online]. Available: https://singaporebirds.com/species/buffy-fish-owl

[61] M. Kappelle, 'TROPICAL FORESTS | Tropical Montane Forests', in *Encyclopedia of Forest Sciences*, Elsevier, 2004, pp. 1782–1792. doi: 10.1016/B0-12-145160-7/00175-7.

[62] A. Goessens *et al.*, 'Is Matang Mangrove Forest in Malaysia Sustainably Rejuvenating after More than a Century of Conservation and Harvesting Management?', *PLoS ONE*, vol. 9, no. 8, p. e105069, Aug. 2014, doi: 10.1371/journal.pone.0105069.

[63] M. A. Salazar-Borunda, M. E. Pereda-Solís, P. M. López-Serrano, J. A. Chávez-Simental, J. H. Martínez-Guerrero, and L. A. Tarango-Arámbula, 'El cambio climático afectará la distribución del búho manchado mexicano (Strix occidentalis lucida Nelson 1903)', *Rev Cha Se Cie For y del Amb*, vol. 28, no. 2, pp. 305–318, Jan. 2023, doi: 10.5154/r.rchscfa.2021.10.066.

[64] R. Dinnage, 'NicheFlow: Towards a foundation model for Species Distribution Modelling', Oct. 18, 2024. doi: 10.1101/2024.10.15.618541.

[65] C. Lange *et al.*, ‘Few-shot Species Range Estimation’, 2025, *arXiv*. doi: 10.48550/ARXIV.2502.14977.

APPENDIX

Poster

ASSESSING BIODIVERSITY LOSS DUE TO ENVIRONMENTAL CHANGES USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Problem Statement

Biodiversity in Malaysia, particularly species within the genera Ketupa, faces threats from habitat loss and climate change. Accurate species distribution modeling is challenging due to spatial autocorrelation, biased sampling, and the limitations of presence-only data, reducing reliability for conservation planning and climate impact assessments.

Objectives

1. Develop Species Distribution Models (SDMs) using Random Forest (RF) and Multi-Layer Perceptrons (MLP).
2. Apply Explainable AI (XAI) techniques to enhance interpretability.
3. Compare random sampling and spatial block sampling to address spatial autocorrelation.
4. Analyze environmental variables and create Habitat Suitability Maps (HSMs) for current and future climate scenarios for period 2061-2080.

Environmental Impact Analysis

1. Urban land cover had the greatest influence on owl habitat suitability, followed by annual precipitation (Bio12). In contrast, mean diurnal temperature range (bio02) had minimal impact.
2. High suitability areas are mainly found in lowland forests and wetlands, with presence near urban edges, reflecting both natural and semi-modified habitats.

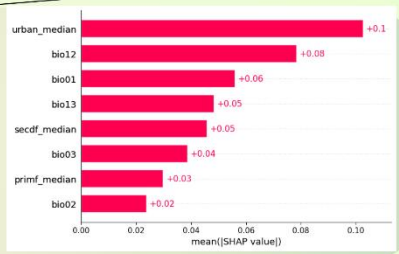


Figure 1 Shapley Values

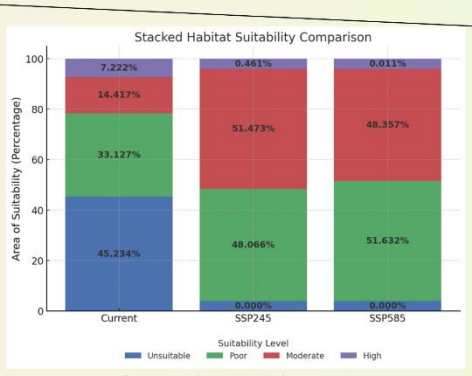


Figure 2 Habitat Suitability Comparison

Results

1. Spatial block sampling outperformed random sampling by reducing spatial dependency,
2. Random Forest performed best, with AUROC 0.8784 and AUC-PR 0.8718, outperforming MLP (AUROC 0.8359, AUC-PR 0.8398).
3. High suitability areas drop from 7.22% (current) to 0.46% (SSP2-4.5) and 0.01% (SSP5-8.5) by 2061-2080, showing drastic habitat loss.

Presenter: Chia Cheng Gun

Supervisor:Dr Mogana a/p Vadiveloo