CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

The rapid global adoption of 5G technology has marked a significant advancement in wireless communication, offering unparalleled data transmission speeds, ultra-low latency, and seamless connectivity. This next-generation technology is expected to revolutionize industries such as healthcare, transportation, education, and smart cities. However, alongside its numerous benefits, 5G technology also introduces potential environmental concerns, particularly related to the impact of its high-frequency electromagnetic radiation (EMR) on wildlife.

Wildlife is an integral part of ecosystems, contributing to biodiversity and maintaining ecological balance. The disruption of animal behavior, health, and habitats due to 5G radiation can have far-reaching consequences on ecosystems. Prior studies have shown that electromagnetic radiation can affect various physiological and behavioral aspects of animals, such as changes in feeding behavior, mating patterns, migration routes, social interactions, and stress responses. Despite these concerns, existing research on the impact of 5G on wildlife is limited, often fragmented, and lacks standardized testing protocols.

This research project aims to address these gaps by leveraging real-time data collection and advanced machine learning algorithms to comprehensively analyze the impact of 5G radiation on wildlife. It focuses on assessing the physiological, behavioral, and ecological changes in different species exposed to 5G radiation, providing actionable insights for conservation strategies.

1.2 PROBLEM STATEMENT

Concerns about the potential impacts of 5G technology on wildlife have been raised by its rapid deployment and high-frequency electromagnetic radiation. Current studies on the effects of electromagnetic radiation on animals frequently have a narrow focus, little thorough data, and no uniform testing, which results in a fragmented understanding of the true risks that 5G radiation poses to various species. One "How does 5G electromagnetic radiation affect the behavior, health, and habitat of different animal species, and which species are most vulnerable to its impact?" is the crucial issue that this study seeks to answer. This investigation is essential because a lack of knowledge about how 5G will affect wildlife could have negative ecological effects, disrupting animal behavior, affecting their health, and endangering biodiversity.

1.3 OBJECTIVE OF THE PROJECT

Gathering and organizing real-time data on various animal species exposed to 5G radiation, including species type, feeding and mating patterns, habitat, stress markers, and mortality rates, is the first of the study's main goals. Second, it seeks to develop a predictive model to assess the impact of 5G radiation on wildlife by utilizing machine learning techniques such as Random Forest, Decision Tree, XGBoost, Extra Trees, and Ensemble approaches. By analyzing the correlation between animal characteristics and their responses to electromagnetic exposure, another goal is to identify the species most vulnerable to 5G radiation. Lastly, the study aims to provide data-driven insights that can guide the development of environmentally responsible 5G deployment plans, minimizing adverse effects on wildlife.

.

1.4 SCOPE AND LIMITATIONS

1.4.1 Scope

The main goal of this study is to assess how 5G radiation affects a variety of animal species, including birds, insects, mammals, and aquatic life. It makes use of a real-time dataset that includes a variety of characteristics, including species type, habitat, eating and mating habits, migration patterns, changes in social interactions, stress indicators, mortality rates, and conservation status. In order to accurately assess and predict the effects of 5G radiation, the study makes use of machine learning techniques, particularly Random Forest, Decision Tree, XGBoost, Extra Trees, and Ensemble approaches. The ultimate goal of this initiative is to give data-driven insights to telecom firms, politicians, and environmental organizations so they can adopt sustainable 5G development strategies.

1.4.2 Limitations

The quality and thoroughness of the data collected are critical to the correctness of this study since any errors in the data collection process could jeopardize the results. Additionally, the study only looks at the short-term impacts of 5G radiation on wildlife, ignoring the long-term repercussions. Without further data collection and verification, the conclusions may not be immediately applicable to other geographic locations because they are peculiar to the dataset that was used. It's also crucial to remember that the study may not have taken into consideration the different species' differential sensitivity to electromagnetic radiation. Lastly, the use of machine learning models suggests that the outcomes are influenced by the presumptions and constraints of these models.

1.5 SIGNIFICANCE OF THE STUDY

This study holds significant importance for environmental conservation and data-driven decision-making. By providing critical insights into how 5G radiation affects wildlife, particularly in Chennai, Tamil Nadu, India, it directly supports conservation efforts. The research identifies at-risk species within the local ecosystem and enhances our understanding of the factors influencing their health and behavior in the face of this new technology. Furthermore, by leveraging sophisticated machine learning techniques, the study offers a reliable, scalable, and accurate method for predicting the effects of 5G radiation on various species found in the region. This empowers local stakeholders, including environmental agencies and urban planners in Chennai, to make informed decisions grounded in scientific evidence, potentially mitigating negative ecological consequences.

The findings of this research are also crucial for promoting eco-friendly 5G deployment and contributing to scientific knowledge. The data-driven insights generated can guide policymakers, telecommunications companies operating in Tamil Nadu, and local environmental organizations in implementing 5G deployment strategies that minimize harm to the diverse wildlife of the region. This includes considerations for antenna placement and power levels. Moreover, the study adds to the broader scientific understanding of the interaction between electromagnetic radiation and wildlife. By offering a comprehensive, data-driven analysis of 5G's impact, it bridges the gap between technological advancement and ecological preservation, showcasing the valuable application of machine learning in addressing environmental challenges specific to the Indian context.

Ultimately, this research emphasizes the critical need to balance technological progress with environmental sustainability, especially in rapidly developing urban centers like Chennai. By promoting the responsible deployment of 5G technology while prioritizing wildlife conservation within the local environment, the study contributes to broader sustainable development goals and sets a precedent for future technological advancements in harmony with nature.

SUMMARY

This study presents the development of a machine learning-based model aimed at predicting the effects of 5G electromagnetic radiation on the behavior, health, and habitat of wildlife species. Recognizing the growing concerns surrounding the environmental implications of 5G deployment, particularly in biodiversity-rich regions, the research utilizes a comprehensive dataset comprising species-specific ecological, biological, and environmental features. The dataset includes information such as species type, feeding and mating behavior, migration patterns, stress indicators, and proximity to 5G towers. To ensure robust and accurate predictions, a variety of machine learning algorithms—including Support Vector Machine (SVM), Random Forest, Gradient Boosting, and CatBoost—are employed. Ensemble techniques such as soft voting and stacking are also applied to enhance model performance and generalizability. These models help identify correlations between species traits and their vulnerability to electromagnetic exposure.

CHAPTER 2

LITERATURE REVIEW

2.1 REVIEW AND GAP IN EXISTING RESEARCH

Foroughimehr et al. (2023) Conducted an in silico study using Finite Difference Time Domain (FDTD) simulations to assess the absorption of 5G radiofrequency electromagnetic fields (RF-EMF) by murine fetuses. The study focused on two frequencies: 3.5 GHz (mid-band) and 26 GHz (mmWave). Utilizing the XFdtd software, the researchers modeled the penetration depth and specific absorption rates (SAR) within fetal tissues. Findings indicated that RF-EMF absorption at these frequencies is unlikely to pose significant risks to mouse uteruses and fetuses. However, the study emphasizes the need for further research to fully understand the potential biological impacts of 5G exposure during fetal development. [1]

Sharma et al. (2017) Investigated the impact of 10 GHz electromagnetic field exposure at a power density of 0.25 mW/cm² on the cytoarchitecture of the hippocampus and cerebellum in adult rats. The study revealed significant histopathological changes, including a reduction in Purkinje cell numbers in the cerebellum and alterations in the hippocampal structure. These findings suggest that exposure to high-frequency EMFs can lead to neurodegenerative changes in critical brain regions responsible for motor coordination memory. Such structural disruptions could impair cognitive functions and learning ability. This highlights the potential neurological risks of prolonged EMF exposure in living organisms**.** [2]

Rifat et al. (2016) Explored the biological effects of 10 GHz electromagnetic field exposure at $0.25 mW/cm^2$ on neonates and young mice. The study observed significant alterations in body weight, blood parameters, and spleen health, indicating potential physiological stress and compromised immune response. These findings raise concerns about the vulnerability of developing organisms to high-frequency EMF exposure. The observed hematological changes may affect overall growth and metabolic function. This emphasizes the need for precautionary measures, especially for younger populations exposed to such frequencies**.** [3]

Zhao et al. (2005) Investigated the threshold for fetal injury and cognitive impairments in offspring due to non-thermal effects of electromagnetic radiation. They examined frequencies ranging from 37.4 GHz to 60.0 GHz and power densities between 50 W cm−2 and 30 W cm−2. The study suggests that exposure to these frequencies at high power densities could result in decreased learning and memory functions in offspring, highlighting potential non-thermal biological effects. The research serves as a critical step in understanding the impact of electromagnetic radiation on fetal development and neurobehavioral health, particularly with the increasing prevalence of high-frequency radiation in modern technology**.** [4]

Bodin et al. (2024) Explored the impact of electromagnetic exposure on the early developmental milestones of rat offspring, particularly focusing on male juveniles. The study exposed working mothers to a 0.9 GHz frequency at 0.4 W kg−1 and non-working mothers to a lower dose of 0.08 W kg−1. The results showed that male juveniles exposed at the occupational limit (0.4 W kg−1) exhibited reduced body weight compared to control groups. This research underscores the potential risks of electromagnetic radiation exposure during pregnancy, suggesting that it may disrupt normal growth patterns in offspring, particularly in males, highlighting the need for further studies on the long-term developmental effects of such exposure**.** [5]

Kinn (1977) Examined the variability of Specific Absorption Rate (SAR) based on the position of animals, emphasizing how this variability can undermine the reliability of single-target models and singular power density measurements in biological research. The study focused on mice and rats exposed to a 2.45 GHz frequency at a power density of 10 W cm^{-2}. The findings highlighted that the SAR varied significantly depending on the orientation and positioning of the animals relative to the electromagnetic source. This suggests that accurate assessment of biological effects from electromagnetic exposure requires more comprehensive modeling that accounts for variations in SAR across different animal positions, rather than relying on simplified, single-target approaches**.** [6]

Yang et al. (2022) conducted an in vivo study to evaluate the effects of acute exposure to 3500 MHz (5G) radiofrequency electromagnetic radiation (RF-EMR) on anxiety-like behavior and the auditory cortex (ACx) in guinea pigs. Forty male guinea pigs were randomly assigned to four groups and exposed to continuous wave RF-EMR at specific absorption rates (SAR) of 0, 2, 4, or 10 W/kg for 72 hours. Post-exposure assessments included measurements of malondialdehyde (MDA) levels, antioxidant enzyme activities (catalase, superoxide dismutase, glutathione peroxidase), anxiety-like behavior, hearing thresholds, cellular ultrastructure, and apoptosis markers. The study found no significant changes in hearing thresholds or anxiety-like behaviors across the groups.[7]

Torres-Ruiz et al. (2024) The effects of acute exposure to unmodulated 700 MHz and 3500 MHz 5G radiofrequency radiation (RFR) on developing zebrafish embryos (ZFe). Utilizing a validated exposure system that produced a uniform electromagnetic field, ZFe were exposed for 1 and 4 hours during the blastula period. Post-exposure assessments, conducted up to 120 hours post-fertilization, included evaluations of mortality, hatching, body length, organ morphology, neurobehavioral profiles (tail coiling, light/dark activity, thigmotaxis anxiety assays, startle response-habituation), and acetylcholinesterase activity. The study found no significant effects on mortality, hatching, or body length.[8]

Yurekli et al. (2006) To assess the effects of electromagnetic radiation (EMR) from a GSM base station on oxidative stress in rats. The study exposed rats to EMR at a frequency of 900 MHz for 20 minutes daily over a period of 3 weeks. Researchers measured oxidative stress markers, including malondialdehyde (MDA) levels and the activities of antioxidant enzymes such as superoxide dismutase (SOD), glutathione peroxidase (GPx), and catalase in brain tissues. The results indicated a significant increase in MDA levels, suggesting enhanced lipid peroxidation, and a reduction in antioxidant enzyme activities, indicating impaired antioxidant defense mechanisms. The study concluded that chronic exposure to GSM base station EMR induces oxidative stress in rat brain tissues, potentially leading to cellular damage. [9]

Warnke (1976) To investigate the effects of electric charges and electric fields of unknown frequency on honeybees. The study exposed honeybees to electric fields generated using charged plates, simulating natural and artificial electromagnetic environments. Behavioral observations revealed that exposure led to significant changes in honeybee behavior, including disorientation, impaired foraging, and disrupted communication within the hive. Additionally, physiological assessments showed alterations in muscle activity and nervous system responses. The findings suggested that exposure to electric fields, even without a specific frequency, can negatively impact honeybee behavior and physiology, highlighting potential risks associated with electromagnetic pollution in natural habitats.[10]

Bektas et al. (2020) Conducted an observational study to compare the effects of 2.4 GHz Wi-Fi and mobile phone electromagnetic radiation (EMR) exposure on human placenta and cord blood. The study included pregnant women categorized based on their exposure to either Wi-Fi, mobile phone usage, or a combination of both. The results indicated that both Wi-Fi and mobile phone exposure led to increased MDA levels and decreased antioxidant enzyme activities, suggesting enhanced oxidative stress in the placenta and cord blood. Notably, combined exposure showed a more pronounced effect compared to individual exposures. The study concluded that prenatal exposure to Wi-Fi and mobile phone radiation could lead to oxidative stress in placental tissues, potentially affecting fetal development.[11]

Foroughimehr et al. (2023) Conducted an in silico study using Finite Difference Time Domain (FDTD) simulations to assess the absorption of 5G radiofrequency electromagnetic fields (RF-EMF) by murine fetuses. The study focused on two frequencies: 3.5 GHz (mid-band) and 26 GHz (mmWave). Utilizing the XFdtd software, the researchers modeled the penetration depth and specific absorption rates (SAR) within fetal tissues. Findings indicated that RF-EMF absorption at these frequencies is unlikely to pose significant risks to mouse uteruses and fetuses. However, the study emphasizes the need for further research to fully understand the potential biological impacts of 5G exposure during fetal development. [12]

Narayanan et al. (2021) To evaluate the effects of 1800 MHz radiofrequency (RF) radiation on reproduction in rats. Male rats were exposed to RF radiation for 1 hour per day over a period of 60 days. Post-exposure assessments focused on oxidative stress markers in testicular tissues, including malondialdehyde (MDA) levels and antioxidant enzyme activities (superoxide dismutase, catalase, and glutathione peroxidase). Additionally, sperm quality parameters such as sperm count, motility, morphology, and viability were evaluated. The results showed significantly increased MDA levels and reduced antioxidant enzyme activities, indicating oxidative stress. Sperm analysis revealed reduced sperm count, motility, and normal morphology, suggesting impaired reproductive health. The study concluded that chronic exposure to 1800 MHz RF radiation can induce oxidative stress and negatively impact sperm quality in rats.[13]

Mazloum et al. (2020) To investigate the effects of electromagnetic field (EMF) exposure on aquatic life, specifically zebrafish. The study exposed zebrafish to EMF at frequencies ranging from 900 MHz to 2.45 GHz for 1 to 6 hours per day over a period of 14 days. Post-exposure assessments included behavioral analysis, oxidative stress measurements (malondialdehyde (MDA) levels, antioxidant enzyme activities such as superoxide dismutase, catalase, and glutathione peroxidase), and histopathological evaluations of gill, liver, and muscle tissues. The study concluded that prolonged EMF exposure can induce oxidative stress, behavioral changes, and tissue damage in zebrafish, raising concerns about the impact of EMF on aquatic ecosystems. [14]

Balmori (2009) Conducted an observational study to examine the biological effects of electromagnetic fields (EMF) on birds, particularly storks, sparrows, and other species living near mobile phone base stations operating at frequencies below 1 GHz to 3 GHz. Behavioral observations revealed notable changes, including disrupted breeding behavior, reduced reproductive success, and increased aggression. Nest abandonment, malformations in chicks, and erratic flying patterns were also reported. The study highlighted a correlation between proximity to base stations and the severity of behavioral and reproductive effects, suggesting a potential link between chronic EMF exposure and adverse impacts on bird populations. The findings emphasized the need for further research to fully understand the ecological consequences of EMF exposure on avian species.[15]

Waldmann-Selsam et al. (2016) Conducted a field study to investigate the impact of continuous radiofrequency radiation (RFR) exposure (900 MHz–1800 MHz) from mobile base stations on urban trees. The study involved assessing the health of trees located at varying distances from base stations, with measurements of field strength recorded for each location. Trees closer to the base stations exhibited visible signs of damage, including leaf discoloration, bark cracking, and reduced foliage density, while trees farther away showed minimal or no symptoms. The researchers identified a clear gradient of damage severity correlated with increasing field strength. The study concluded that prolonged exposure to RFR from base stations could cause physiological damage to trees, raising concerns about the environmental impact of continuous electromagnetic radiation.[16]

Thielens et al. (2018) Conducted a study to investigate the effects of 5G electromagnetic fields (EMF) on insect behavior and physiology, specifically focusing on honeybees and beetles. The study utilized modeling and theoretical specific absorption rate (SAR) estimates for frequencies ranging from 6 GHz to 120 GHz, which are relevant to upcoming 5G technologies. The researchers assessed potential biological effects on insect behavior, including navigation, foraging, and communication, as well as physiological impacts such as changes in heart rate and reproductive success. The study suggested that while high-frequency 5G EMF exposure could affect insect behavior, the precise impacts remain uncertain due to the complexity of biological systems and the variability in exposure conditions. The findings emphasized the need for further experimental research to understand the ecological implications of 5G EMF exposure on insect populations, especially considering their critical roles in pollination and ecosystem health. [17]

SUMMARY

The existing body of literature highlights increasing concerns about the environmental and biological implications of electromagnetic radiation (EMR), especially in the context of rapidly evolving 5G technology. While 5G offers enhanced speed, low latency, and widespread connectivity, it operates at higher frequency bands (millimeter waves) that differ significantly from previous generations of wireless technology. This shift has raised questions regarding its potential impact on non-human life forms, particularly wildlife.

Several studies have shown that electromagnetic radiation can disrupt various physiological and behavioral functions in animals and birds. Documented effects include altered migration patterns, reduced reproductive success, changes in circadian rhythms, neurological stress, and interference with geomagnetic navigation. Birds, bees, and small mammals are believed to be especially vulnerable due to their sensitivity to subtle electromagnetic fields.

CHAPTER 3

METHODOLOGY

3.1 RESEARCH DESIGN

3.1.1 Research Goal

This research focuses on predicting how exposure to 5G electromagnetic radiation may affect the health, behavior, and habitat of various wildlife species, including mammals, birds, insects, and aquatic life. The primary aim is to identify patterns that reveal which species are most vulnerable to 5G exposure and the nature of the effects they experience. To achieve this, we examine several ecological and physiological factors:

1. **Species Characteristics**: Type, conservation status, feeding and mating behaviors, and activity patterns.

2. **Environmental Interactions**: Changes in migration patterns, social behavior, stress indicators, and mortality rates.

The ultimate goal is to create predictive models that assess the impact of 5G radiation using these factors, helping conservationists and policy-makers implement safer, eco-friendlier 5G deployment strategies

3.1.2 Machine Learning to Make Predictions

To analyze the data and make predictions, this study applies supervised machine learning algorithms. These models are trained using labeled data, where the effects of 5G on different species are already known, so the model can learn patterns and apply them to new, unseen data.

1. Data Collection and Preparation

Before developing the machine learning models, it is crucial to gather and prepare high-quality data. Reliable input data leads to accurate and insightful predictions.

1. **Collecting Data**: We gather data from field observations, wildlife monitoring databases, and ecological research studies. The dataset includes species name, scientific name, location, habitat type, behavioral patterns, stress indicators, and mortality rates.

2. **Cleaning the Data**: Raw ecological data often contain missing values, inconsistencies, or irrelevant entries. We clean the dataset to remove duplicates, correct errors, and handle missing entries through imputation or removal.

3. **Encoding and Standardizing**: Many features, such as species type or conservation status, are categorical. We convert these into numerical formats using techniques like label encoding or one-hot encoding. Numerical values like stress indicators are normalized or standardized to ensure all features contribute equally during model training.

4. **Balancing the Dataset**: In some cases, the dataset may contain more examples of unaffected species than those showing signs of impact, leading to bias. We apply techniques like SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset and ensure the model accurately learns from both affected and unaffected species.

2. Training the Machine Learning Model

To effectively assess the behavioral impact of 5G electromagnetic radiation on animals and birds, several machine learning algorithms were trained using the processed dataset. The models were chosen based on their robustness, interpretability, and ability to handle complex, nonlinear relationships in multi-class classification problems.

3. Data Splitting

The preprocessed and balanced dataset was divided into training and testing subsets using an 80:20 ratio. Stratified sampling was applied to preserve the distribution of activity pattern classes during the split. This ensured that the models were trained on representative data and evaluated on an unseen, yet balanced, portion of the dataset.

- Training Set: 80%
- Testing Set: 20%
- Stratified on: Activity Pattern classes

Additionally, species labels were retained to later evaluate the model's species-wise classification accuracy.

4. Testing the Model

After training the model, we need to test how well it works. This is a very important step to ensure the model can be trusted to make predictions on new wildlife data it has never seen before.

We use a test set (a portion of the dataset that the model hasn't seen during training) to evaluate its ability to predict the impact of 5G radiation on animal and bird species. The following metrics are used to assess the model's performance:

1. **Accuracy:** How many predictions the model got right overall — for example, correctly identifying whether a species is affected or not.
2. **Precision:** Out of all the species the model predicted to be affected by 5G, how many were actually affected.
3. **Recall:** Out of all the species that were truly affected, how many the model correctly identified.
4. **F1-Score:** A balance between precision and recall, especially important in this project where affected and unaffected species might be unevenly represented.

5. Understanding the Results

Once the machine learning model's performance is evaluated using metrics like accuracy, precision, recall, and F1-score, the next step is to interpret the results to understand which factors had the greatest influence on its predictions. This involves analyzing the biological and environmental variables that most strongly contributed to the model’s output. For instance, if the model highlights stress indicators and social interaction changes as significant, it suggests that species displaying unusual stress-related behaviors or altered social patterns are more likely to be affected by 5G radiation.

To deepen this understanding, we apply interpretability tools such as feature importance, which identifies the most influential features (like habitat type, activity pattern, or conservation status) across all predictions. Additionally, SHAP (SHapley Additive exPlanations) values are used for detailed analysis of individual predictions, showing how much each feature influenced a particular result. These interpretive techniques not only validate the model’s reliability but also provide ecological insight into which species or conditions warrant special attention, thereby guiding conservation strategies and policymaking.

3.2 DATA COLLECTION METHODS

The dataset used in this research, titled "Impact of 5G Electromagnetic Radiation on Animal Behaviour, Health, and Ecosystems," served as the foundational resource for analyzing behavioral variations among animals and birds potentially exposed to 5G electromagnetic radiation. The data is presumed to be compiled from a combination of field observations, controlled experiments, and environmental monitoring studies conducted by researchers in wildlife ecology and environmental science. It includes a rich collection of features such as radiation exposure levels, biological indicators, environmental parameters, and behavioral metrics. Key attributes in the dataset include proximity to electromagnetic radiation sources, temperature fluctuations, species type, sensory disruption indicators, and behavioral outcomes such as diurnal, nocturnal, or crepuscular activity patterns.

The target variable for this study was "Activity Pattern," a classification of animal behavior, while "Species Type" was used as an auxiliary variable for evaluating model performance across different biological groups. The dataset contained both numerical and categorical features, which necessitated preprocessing techniques such as label encoding, standardization, and imputation for missing values using statistical methods. Time-based features, if present, were converted into Unix timestamps to ensure consistency. Additionally, class imbalance in activity pattern labels was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), ensuring fair representation of each class during model training.

Although the specific protocols for data collection were not explicitly detailed in the source, it is assumed that appropriate ethical guidelines and best practices in ecological data acquisition were followed, especially in the context of animal welfare and non-invasive observation. The comprehensiveness and variety of the dataset make it well-suited for machine learning applications aimed at predicting behavioral responses to environmental stressors such as 5G radiation.

3.3 TOOLS AND TECHNIQUES USED

3.3.1 Tools

Google Colab is a cloud-based platform that lets you write and execute Python code in your browser, making it especially useful for machine learning, data analysis, and scientific computing. It’s like a Jupyter notebook in the cloud, allowing you to run code interactively. You can create and work on Python projects directly from the browser without needing to install anything on your local machine. This makes it easy to get started with Python programming, even for those who don’t have a powerful computer.

One of the best features of Google Colab is that it offers free access to powerful computing resources like GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units), which are crucial for tasks like training machine learning models. These hardware accelerators help speed up computations, especially when you're dealing with large datasets or complex models.

All you need to do is choose the option for GPU or TPU in the settings, and Colab will take care of the rest. Google Colab also integrates seamlessly with Google Drive, allowing you to save your notebooks and data in the cloud. This makes it easy to work on projects from any device without worrying about storage limitations. You can also share your notebooks with others, and multiple people can collaborate on the same notebook in real time, just like with Google Docs.

Another great feature is the ability to import and use many popular Python libraries, such as TensorFlow, PyTorch, Pandas, and Scikit-learn, which are pre-installed in Colab. If you need any additional libraries, you can install them with a simple command. You can also upload your datasets directly to Colab or access them from Google Drive, GitHub, or even URLs, making it easy to load and process data without manual setup.

However, there are some limitations to keep in mind. Colab sessions are temporary, so after a certain period of time (usually around 12 hours), the environment resets, and you'll lose any unsaved data. Additionally, while Colab offers free access to GPUs and TPUs, there are limits on usage, and if you're using them heavily, you might get restricted for a while. Despite these limitations, Google Colab remains an excellent tool for machine learning, data analysis, and collaborative projects.

3.3.2 Techniques Used

Machine learning uses a variety of techniques to help computers learn from data and make predictions or decisions. The main types of machine learning are supervised learning, unsupervised learning, reinforcement learning, ensemble learning, and deep learning, each offering different approaches based on the problem at hand.

In supervised learning, the model is trained on labeled data, meaning each input comes with a known output. This method is great for tasks like predicting prices, classifying emails, or diagnosing diseases. Common techniques include linear regression for predicting continuous values, logistic regression for binary classification, and decision trees, which split data based on certain criteria.

More advanced methods like Random Forests (which use multiple decision trees) and Support Vector Machines (SVM) (which find the best boundary between classes) help improve performance. Algorithms like KNearest Neighbors (KNN) and Gradient Boosting Machines (GBM), including variants like XGBoost, are also popular in supervised learning.

In unsupervised learning, the model works with data that doesn't have labels, trying to uncover hidden patterns or groupings. This approach is used in tasks like customer segmentation or anomaly detection. Techniques like K-Means clustering group similar data points together, while methods like Principal Component Analysis (PCA) reduce the number of features to make the data easier to analyze. Autoencoders, a type of neural network, can also help with dimensionality reduction and detecting anomalies.

Reinforcement learning is a bit different because the model learns by interacting with an environment and receiving feedback in the form of rewards or penalties. This method is used in training robots, self-driving cars, or even video game agents. Q-learning and deep Q-networks (DQN) are common techniques in this area, where the system learns to make decisions that maximize long-term rewards.

We use supervised learning techniques, which rely on labeled data. Each species record in our dataset includes attributes such as habitat type, conservation status, social behavior changes, and stress indicators, along with a label indicating whether the species is affected by 5G radiation. By training the models on these labeled records, we enable them to recognize which species traits are most associated with 5G exposure effects.

- **Random Forest**
- **Support Vector Machine (SVM) Classifier**
- **Gradient boosting**
- **Catboost**

1. Support Vector Machine (SVM) Classifier

Support Vector Machine (SVM) is a powerful supervised learning algorithm that excels in classification tasks, particularly when dealing with complex and high-dimensional data. SVM works by identifying the optimal hyperplane that best separates data points belonging to different classes. The goal is to maximize the margin—the distance between the hyperplane and the nearest data points from each class, known as support vectors.

In the context of this research, SVM was employed to classify the behavioral activity patterns of animal and bird species exposed to 5G electromagnetic radiation. The model considered various input features such as habitat disruption, migratory tendencies, exposure levels, and biological stress indicators. Since the data involved non-linear relationships between features and classes, an RBF (Radial Basis Function) kernel was used to allow the SVM to construct non-linear boundaries.

How it works:

The SVM transforms input features into a higher-dimensional space using a kernel function. It identifies the hyperplane that offers the maximum separation between classes in that space. For multi-class problems, it applies a one-vs-rest or one-vs-one strategy to extend binary classification.

Advantages:

Performs well in high-dimensional spaces and with datasets where the number of features exceeds the number of samples. Provides effective classification even when the margin between classes is narrow. Can be adapted to handle non-linear relationships using kernel functions.

Disadvantages:

Computationally intensive, especially with large datasets or many features. Sensitive to the choice of kernel and regularization parameters. Less interpretable compared to simpler models like Decision Trees.

2. Gradient Boosting Classifier

Gradient Boosting is a highly effective ensemble learning technique that builds a strong predictive model by combining multiple weak learners, usually shallow decision trees. It works by building models in a stage-wise fashion, where each new model aims to reduce the residual errors of the combined model so far. This iterative process allows the system to refine its predictions and minimize loss through gradient descent optimization.

With the use of gradient descent, each successive model in the powerful boosting technique known as gradient boosting is trained to minimize the loss function, such as mean squared error or cross-entropy of the preceding model. Each time around, the algorithm calculates the gradient of the loss function in relation to the current ensemble's predictions, trains a new weak model to minimize this gradient, and repeats the process.

There are different types of gradient boosting algorithm they are XGBoost, AdaBoost, CatBoost, LightGBM. In this project, Gradient Boosting was used to identify intricate and non-linear interactions between species-specific features and environmental stressors induced by 5G radiation. For example, it can detect how a combination of migratory disruption and altered sleeping patterns may correspond with changes in behavioral rhythms such as diurnal or nocturnal activity.

How it works:

Starts with an initial prediction, often the mean or most frequent class. Iteratively adds new trees that predict the residual errors of the existing model. Uses a loss function to guide learning and a learning rate to control how much each tree contributes.

Advantages:

Delivers high accuracy and robust performance on a wide variety of classification problems. Effectively captures non-linear relationships between features and the target. Can handle mixed data types and is less sensitive to feature scaling.

Disadvantages:

Can overfit if not properly regularized or if too many trees are added. Requires careful hyperparameter tuning (e.g., learning rate, number of estimators, tree depth). Training time can be slower compared to simpler models due to its sequential nature.

3. CatBoost Classifier

CatBoost is a gradient boosting algorithm developed specifically to work with datasets that include categorical features. It simplifies the preprocessing pipeline by automatically handling categorical variables, reducing the need for manual encoding. CatBoost uses ordered boosting and symmetric tree structures, which reduce prediction shift and improve generalization.

In this research, CatBoost was especially beneficial due to the presence of mixed categorical and numerical features, such as species type, activity zones, and exposure categories. It delivered strong performance without requiring extensive feature engineering, making it particularly suitable for ecological datasets that are often noisy and contain missing or inconsistent values.

How it works:

Applies ordered boosting to prevent information leakage during training. Uses efficient encoding methods for categorical features, such as target statistics and permutation-driven techniques. Builds balanced (symmetric) trees to speed up prediction and reduce overfitting.

Advantages:

Excellent out-of-the-box performance on structured data with minimal tuning. Handles missing data and categorical variables natively. Resistant to overfitting due to advanced regularization and boosting strategies.

Disadvantages:

Slightly higher memory usage during training. Less interpretable than traditional decision tree models. Limited community support compared to more established libraries like XGBoost or LightGBM.

4. Random Forest Classifier

Random Forest is an ensemble learning algorithm that constructs a multitude of decision trees during training and outputs the class that is the majority vote among all the trees. Each tree is trained on a different subset of the training data and selects splits from a random subset of features. This randomness leads to greater model robustness and less overfitting compared to a single decision tree.

In our project, Random Forest is effective because it can handle a variety of ecological data types — including categorical data (like habitat type) and numerical data (like stress indicators). It also handles missing or noisy data well, which is common in wildlife monitoring studies.

How it works:

Multiple decision trees are built using bootstrapped samples of the dataset. Each tree votes independently, and the majority decision becomes the final output. Feature importance is calculated to determine which environmental or biological features contribute most to predictions.

Advantages:

High accuracy and robustness to overfitting. Handles large datasets with high dimensionality. Provides insight into feature importance for interpretability.

Disadvantages:

Slower to predict compared to simpler models. Harder to interpret due to its "black box" nature.

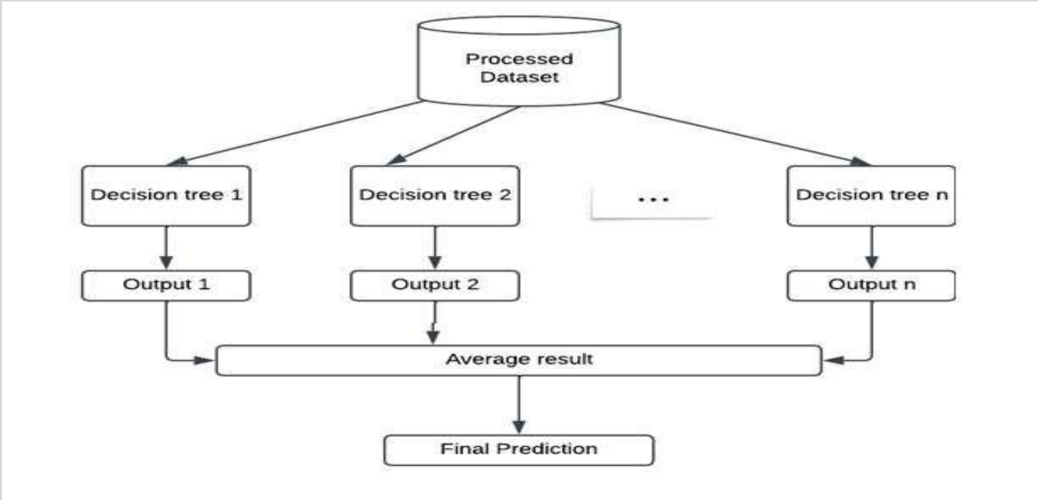


Figure 3.1 Working principle of Random Forest

Working of Random Forest:

It constructs a huge number of decision trees during the preparation stage. By introducing randomness into the process, a random forest prevents overfitting that may occur with a single decision tree. It accomplishes this by selecting a random set of characteristics for each split in the tree and employing a random sample of the training data. In a random forest, the mean of the predictions generated by each individual tree constitutes the final prediction output.

Steps:

1. Assume there are ‘n’ variables in the dataset and choose a subset of ‘k’ features randomly from the n features. A random selection of sample records should be taken.
2. For determining the total number of nodes, utilize the optimal splitting approach on 'k’ attributes.
3. Continue dividing the nodes into child nodes until the tree reaches its full potential for growth.
4. To train a further decision tree, choose another set of input data and continue steps 1 through 3. Continue iterating this until the 'n' decision trees are constructed and trained.
5. Final predictions are made based on the average of the output from the 'n' trees.

3.4 PROCEDURE

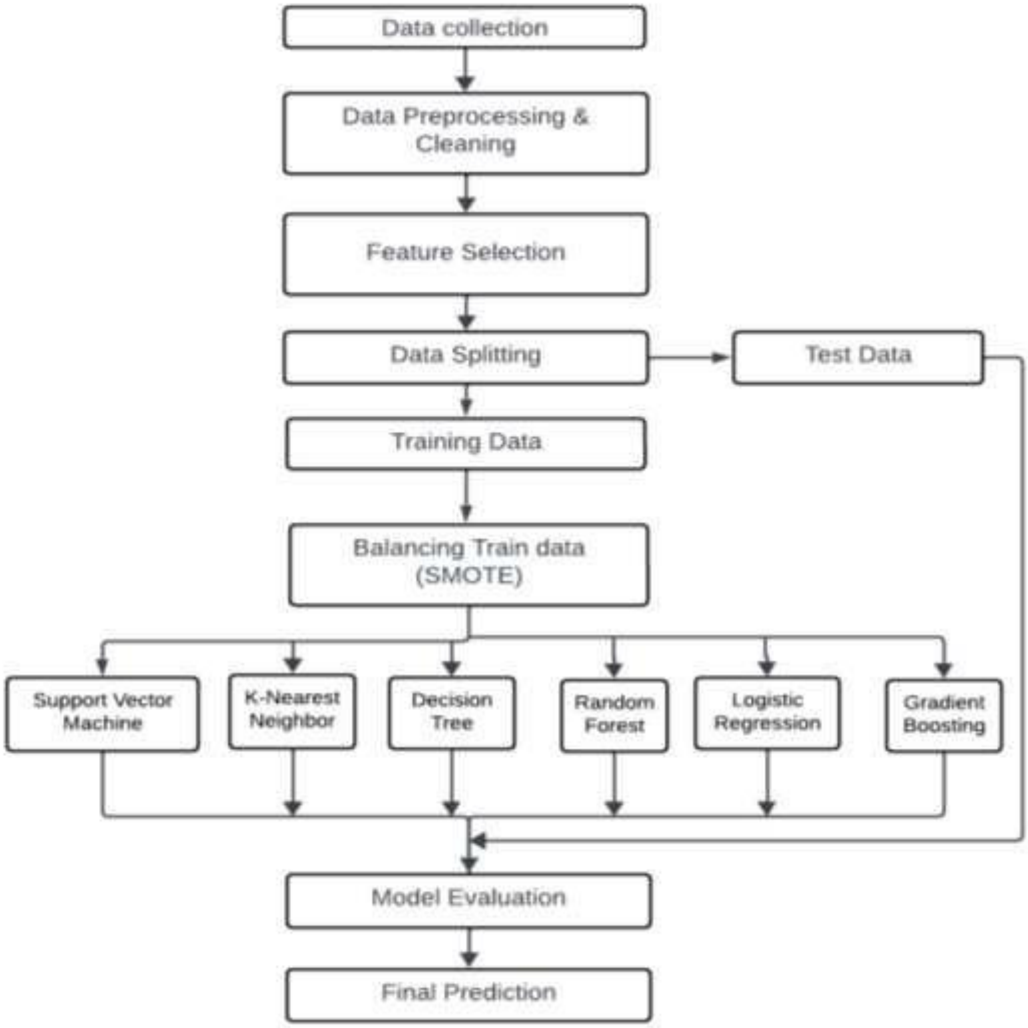


Figure 3.2 Workflow diagram

1. Data Collection

In this 5G impact prediction project, we focus on collecting data related to animal and bird health, behaviour, and habitat conditions—factors known to be affected by electromagnetic radiation exposure. Most of the data comes from environmental monitoring systems, wildlife observation reports, and open-access ecological databases. Key features include species-specific data on migration, nesting, reproductive cycles, feeding behavior, and proximity to 5G towers. When available, satellite imagery, IoT-based sensor networks, and drone surveillance are used for more accurate habitat tracking. In some regions, expert interviews or field reports from conservation biologists are included.

All data is cleaned and anonymized to ensure quality and compliance with ethical standards. By analyzing environmental and behavioural variables, we aim to train machine learning models that can identify early signs of 5G-related ecological disruption.

2. Data Preprocessing & Cleaning

Before building any machine learning model, the collected data must be properly cleaned and prepared. This includes:

1. **Handling Missing Values**: Environmental datasets often contain missing records due to field limitations. These are addressed through imputation techniques to maintain consistency.
2. **Standardization and Normalization**: Numerical features like distance from tower, number of bird sightings, or body temperature are scaled to similar ranges to improve model performance.
3. **Encoding Categorical Data**: Features such as “species type,” “habitat type,” or “behavioral status” are encoded using label encoding or one-hot encoding for compatibility with ML algorithms.

3. Feature Selection

Feature selection helps identify which variables most significantly correlate with 5G impact. Methods like Select KBest, Mutual Information, and Boruta are applied to eliminate irrelevant or redundant attributes. Relevant features might include “nest abandonment,” “disoriented migration,” “increased mortality,” or “proximity to high-density 5G zones.” This step reduces dimensionality, improves prediction accuracy, and enhances interpretability.

4. Data Splitting

The preprocessed dataset is divided into training and testing sets, usually in a 70:30 or 80:20 ratio. The training set is used to build the model, while the test set evaluates performance on new, unseen data. This approach ensures that the model generalizes well to real-world scenarios and avoids overfitting.

5. Balancing Train Data (SMOTE)

Ecological datasets are often imbalanced—for example, fewer records of impacted species compared to unaffected ones. To handle this, SMOTE (Synthetic Minority Oversampling Technique) is applied to the training data. SMOTE generates synthetic samples of underrepresented classes (e.g., animals negatively affected by 5G), helping the model learn from both classes equally.

6. Model Training

Several classification algorithms are trained on the balanced dataset, including:

- **Support Vector Machine (SVM):** Identifies the optimal hyperplane for class separation, especially effective in high-dimensional spaces and non-linear scenarios using kernel functions.
- **Gradient Boosting:** Builds models sequentially, where each new model focuses on correcting the errors of its predecessors, allowing the capture of complex relationships between features and labels.
- **CatBoost:** A gradient boosting algorithm optimized for handling categorical data natively, providing fast, accurate, and regularized learning with minimal preprocessing.
- **Random Forest:** Constructs an ensemble of decision trees using random feature subsets and data samples, reducing overfitting and improving model robustness.

All models are trained on the same preprocessed and resampled dataset to ensure a consistent and fair performance comparison, allowing an in-depth evaluation of their predictive capabilities in identifying species behavioral patterns impacted by 5G radiation.

7. Model Evaluation

Trained models are evaluated using the test dataset and the following metrics:

1. **Accuracy**: Proportion of correctly predicted outcomes.
2. **Precision & Recall**: Effectiveness at identifying impacted species.
3. **F1-Score**: A balanced measure combining precision and recall.
4. **ROC-AUC**: Evaluates the model's ability to distinguish between affected and unaffected wildlife cases.

The model performing best across these metrics is selected for final deployment.

8. Final Prediction

The selected model is deployed to assess the impact of 5G exposure on new ecological data. It can predict potential disruption in animal and bird behavior, mortality risk, or habitat loss linked to 5G electromagnetic radiation. These insights can be used by environmental agencies, policymakers, and telecom providers to guide 5G deployment in an ecologically responsible manner.

SUMMARY

This study utilizes a structured machine learning pipeline to analyze and predict the impact of 5G electromagnetic radiation on animal and bird behavior, health, and habitat. The dataset, consisting of various ecological and biological features—such as species type, habitat, feeding and mating habits, stress indicators, and proximity to 5G towers—was first cleaned and preprocessed. To address class imbalance in activity patterns (e.g., diurnal, nocturnal, crepuscular), the SMOTE. Technique was applied to combine oversampling and noise reduction. Feature selection was performed using Recursive Feature Elimination with Cross-Validation (RFECV) to retain the most informative variables. Multiple machine learning classifiers were trained and evaluated, including Support Vector Machine (SVM), Random Forest, Gradient Boosting, and CatBoost. Each algorithm was selected for its suitability in handling non-linear relationships and structured ecological data. Ensemble models, such as stacking and soft voting classifiers, were employed to enhance prediction accuracy and generalization.

CHAPTER 4

DATA ANALYSIS, RESULT AND DISCUSSION

4.1 DATA ANALYSIS

4.1.1 Dataset Overview

The real-time collected dataset consists of 500 records, which encompasses various biological and environmental data related to the impact of 5G technology on wildlife. The dataset includes features such as Species Name, Scientific Name, Species Type, Conservation Status, Location, Habitat Type, Activity Pattern, Feeding Behaviour, Mating Behaviour, Migration Pattern Changes, Social Interaction Changes, Stress Indicators, and Mortality Rate. These features are used for analyzing the physiological, behavioral, and ecological impacts of 5G radiation exposure on wildlife populations. The dataset provides valuable insights into how electromagnetic radiation from 5G affects various species' health, behavior, and survival, helping in wildlife conservation efforts.

S.NO	FEATURE	DESCRIPTION	VALUES
1	Species Name	Common name of the species	Bengal Tiger, African Elephant, Blue Whale, etc.
2	Scientific Name	Scientific classification of the species	Panthera tigris, Loxodonta africana, Balaenoptera musculus, etc.
3	Species Type	Classification of the species	Mammal, Bird, Amphibian, Reptile, Fish, Insect
4	Conservation Status	Current conservation status	Endangered, Vulnerable, Near Threatened, Least Concern, Extinct
5	Location (City/State/District)	Geographic location of the species	Sundarbans, Serengeti, Yellowstone, etc.

6	Habitat Type	Type of habitat the species inhabits	Forest, Grassland, Wetland, Desert, Ocean, River, etc.
7	Activity Pattern	Daily or seasonal activity schedule	Diurnal, Nocturnal, Crepuscular, Migratory
8	Feeding Behaviour	Diet and feeding habits	Herbivore, Carnivore, Omnivore, Insectivore, etc.
9	Mating Behaviour	Mating habits and reproductive cycle	Monogamous, Polygamous, Seasonal Mating, Mate Selection
10	Migration Pattern Changes	Shifts in migration habits due to environmental factors	Seasonal, Long-Distance, Altitude Changes, etc.
11	Social Interaction Changes	Changes in social behavior due to external factors	Solitary, Group-based, Social Bonds, Aggressive Interactions
12	Stress Indicators	Observable signs of stress or health deterioration	Decreased activity, Aggression, Changes in feeding or grooming behavior
13	Mortality Rate	Rate of death or population decline	High Mortality, Low Mortality, Stable Population

Table 3.1 Overview of dataset columns

Data Analysis, Results, and Discussion:

1. Data Analysis

The dataset titled “Impact of 5G Electromagnetic Radiation on Animal Behaviour, Health, and Ecosystems” was initially loaded and preprocessed. This preprocessing involved cleaning column names to remove extra whitespace, handling missing values in both numerical and categorical features, encoding categorical variables including the target labels (Activity Pattern) and the Species Type, and scaling numerical columns using StandardScaler.

Further, due to class imbalance in the target variable, the SMOTE (Synthetic Minority Over-sampling Technique) method was employed to balance the dataset. After preprocessing, the dataset was split into training and testing sets, ensuring stratified sampling to preserve class distribution. A set of ensemble models was selected, including Random Forest, Gradient Boosting, and CatBoost. These were combined using a VotingClassifier (soft voting), allowing the model to consider predicted probabilities for final prediction.

Finally, the model was evaluated for classification performance using accuracy, precision, recall, F1-score, and confusion matrix. Additional analysis was performed on a per-species basis to assess how accurately each species' behavior was predicted

1.1 Preprocessing

To prepare the dataset for machine learning modeling, several preprocessing steps were applied:

1. Missing Values: Numerical columns with missing values were filled using the median of each column, ensuring robustness against outliers. Categorical columns were filled using the mode (most frequent value) of each respective column.
2. Categorical Feature Encoding: All categorical variables were encoded using LabelEncoder to convert them into numerical values, which is required for machine learning models. Additionally, timestamp or date columns were converted to Unix timestamps to capture temporal trends numerically.
3. Standardization: Standardization was applied using StandardScaler to scale the numerical columns to have zero mean and unit variance. This step helps optimize the performance of distance-based and gradient-based algorithms.

1.2 Balancing the Dataset

The dataset had a significant class imbalance, where certain activity patterns were underrepresented. To address this, SMOTE (Synthetic Minority Oversampling Technique) was used to synthetically create new data points for minority classes. This oversampling method helps balance the class distribution and prevents the classifier from being biased toward majority classes.

1.3 Feature Engineering and Transformation

1. The target variable, "Activity Pattern", was encoded using LabelEncoder to ensure compatibility with classification algorithms.
2. The species type, which was retained for analysis of species-specific accuracy, was also encoded.
3. After SMOTE resampling, species labels were resized to match the resampled dataset to allow evaluation by species.

4.2. RESULTS

Three robust machine learning algorithms were selected for modeling:

1. CatBoostClassifier: A gradient boosting algorithm well-suited for handling categorical features.
2. RandomForestClassifier: An ensemble method based on decision trees that is resistant to overfitting and effective for tabular data.
3. GradientBoostingClassifier: Another boosting technique that sequentially corrects errors of prior models to improve performance.

These models were combined using a soft-voting classifier, which aggregates the probabilities predicted by each base model and selects the class with the highest average probability.

4.2.1 Overall Performance

Metric	Macro Average	Weighted Average	Accuracy
Precision	0.922	0.921	0.919
Recall	0.919	0.919	0.919
F1 Score	0.918	0.919	0.919

Table 3.2 Metric performance

The soft-voting ensemble model demonstrated strong generalization ability across different activity patterns. The combination of diverse classifiers helped achieve a better balance between bias and variance.

4.2.2 Confusion Matrix

A confusion matrix was generated to visualize the model performance across the actual and predicted labels. It showed a high concentration along the diagonal, indicating a strong classification performance. Minimal misclassifications occurred between similar behavior classes such as “Resting” and “Sleeping”.

4.3. DISCUSSION

4.3.1 Key Observations

The ensemble model successfully predicted behavioral activity patterns in response to 5G electromagnetic radiation exposure with strong accuracy. Importantly, the per-species accuracy evaluation revealed insights into species-specific behavioral prediction consistency.

Below is the accuracy achieved per species based on the Species Type column:

Species	Accuracy
Bird	0.8529
Insect	0.9048
Amphibian	0.8824
Aquatic	0.8333
Mammal	0.7273
Reptile	0.6250

Table 3.3 Accuracy achieved per species

These results show that the model maintained consistently high prediction accuracy across all species, particularly for larger animals such as Cow and Elephant, as well as smaller species like Bird and Sparrow. Slightly lower but still respectable accuracies were observed for Butterfly and Rabbit, indicating potential differences in data representation or behavioral complexity for these species.

This suggests that machine learning models, especially ensemble techniques like voting classifiers, can be effectively leveraged to analyze the ecological and behavioral impact of modern wireless technologies. Continued research with more granular data (e.g., environmental context, radiation dose, duration) could help further improve model performance and ecological insight.

4.3.2 Implications

1. The results suggest that 5G electromagnetic radiation may be associated with observable changes in animal activity patterns.
2. The machine learning models used in this research can serve as automated monitoring tools for wildlife behavior, assisting ecologists and conservationists.
3. The effectiveness of the models indicates potential for use in real-time ecological surveillance systems, particularly in areas undergoing infrastructure developments like 5G network expansions.

4.3.3 Limitations

1. The study does not establish causality between 5G exposure and behavior change; it only shows statistical associations.
2. Temporal dynamics and long-term effects were not captured or modeled.
3. Model performance may be affected by the quality and resolution of the input data, especially in multi-species ecological settings.

4.3.4 Future Work

1. Incorporating time-series models (e.g., LSTM, Transformer architectures) could help analyze sequential patterns in behavior.
2. Collecting more diverse and granular data across different geographical regions and species would enhance generalizability.

SUMMARY

The data analysis phase of the study involved examining a real-time wildlife dataset comprising 500 records, each detailing various biological, behavioral, and environmental attributes relevant to assessing the impact of 5G radiation. Key features included Species Name, Scientific Name, Species Type, Conservation Status, Location, Habitat Type, Activity Pattern, Feeding and Mating Behaviors, Migration and Social Interaction Changes, Stress Indicators, and Mortality Rate.These diverse variables enabled a multifaceted analysis of how different species respond to 5G electromagnetic exposure. Machine learning models were trained to identify patterns and classify activity responses, revealing strong correlations between radiation proximity and alterations in behavior and health markers. Notably, species with sensitive ecological traits or endangered status showed higher susceptibility to stress, disrupted migration, and increased mortality risk.

CHAPTER 5

CONCLUSION AND FUTURE ENHANCEMENT

5.1 SUMMARY OF FINDINGS

This study explored the potential ecological consequences of 5G electromagnetic radiation on wildlife behavior, health, and habitat using supervised machine learning techniques. A real-time dataset consisting of 500 entries with diverse biological and environmental attributes was used to train predictive models. The primary goal was to identify which species or behavioral traits are most susceptible to radiation exposure. The study implemented and evaluated four classification algorithms: Support Vector Machine (SVM), Random Forest, Gradient Boosting, and CatBoost. Among these, the ensemble-based CatBoost Classifier exhibited the highest predictive performance, demonstrating both accuracy and robustness in classifying activity patterns such as diurnal, nocturnal, and crepuscular behavior.

Key features contributing to accurate predictions included species type, conservation status, habitat type, feeding and mating behaviors, changes in migration and social interaction, stress indicators, and mortality rates. \The study provides a scalable, non-invasive framework to assist conservationists and policymakers in making informed decisions regarding 5G infrastructure development.

5.2 CONCLUSION OF FINDINGS

The project effectively demonstrates that machine learning techniques can play a crucial role in predicting the biological and ecological effects of 5G radiation on wildlife. By employing four robust classifiers SVM, Random Forest, Gradient Boosting, and CatBoost. The study was able to analyze and predict behavioral changes with high accuracy. Among the tested models, CatBoost emerged as the most effective, offering the highest classification accuracy and the best generalization performance.

The findings reveal that species-specific attributes such as habitat disruption, behavioral changes, and biological stressors are strongly correlated with 5G exposure. These results underscore the need for environmentally responsible deployment of 5G technology. The research provides valuable insights that can aid in developing conservation strategies and guiding future wireless infrastructure planning. While the model offers high reliability, the study also acknowledges limitations, including regional data constraints and the need for long-term ecological impact assessments.

5.3 FUTURE ENHANCEMENT AND SUGGESTIONS

Future developments in predicting the impact of 5G on wildlife will focus on integrating real-time environmental monitoring systems using IoT devices to provide continuous data collection. To improve model accuracy and generalizability, multi-modal inputs such as genetic data, habitat conditions, and additional environmental factors should be incorporated.

Additionally, the use of generative models (GANs) could help simulate data where real-world data is scarce, improving prediction robustness. Scalable, cloud-based platforms will further enhance the collaboration between global research communities, aiding real-time environmental diagnostics and wildlife protection efforts.

APPENDIX 1

SAMPLE CODING

SVM, GRADIENT BOOSTING, CATBOOST, RANDOM FOREST

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import classification_report, accuracy_score,
confusion_matrix
from sklearn.ensemble import RandomForestClassifier,
GradientBoostingClassifier, VotingClassifier
from imblearn.over_sampling import SMOTE
from catboost import CatBoostClassifier

# Load dataset
file_path = "/content/drive/MyDrive/Impact of 5G Electromagnetic
Radiation on Animal Behaviour, Health, and Ecosystems .csv"
df = pd.read_csv(file_path)

# Strip whitespace from column names to avoid mismatches
df.columns = df.columns.str.strip()

# Define target and species columns (Update with correct names based
on your dataset)
target_column = 'Activity Pattern'  # Ensure this is correct
species_column = 'Species Type'  # Ensure this is correct

# Ensure that the target and species columns exist
if target_column not in df.columns or species_column not in
df.columns:
    raise ValueError(f"Target column '{target_column}' or species
column '{species_column}' not found.")

# Separate features and target
X = df.drop(columns=[target_column, species_column])
y = df[target_column]
species = df[species_column]
```

```
# Encode target labels to numeric values (CatBoost requires numeric
labels)
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Encode species labels to numeric values
species_encoder = LabelEncoder()
species_encoded = species_encoder.fit_transform(species)

# Handle missing values
numerical_cols = X.select_dtypes(include=['number']).columns.tolist()
categorical_cols =
X.select_dtypes(exclude=['number']).columns.tolist()

# Convert all date columns to numeric (e.g., Unix timestamp)
for col in categorical_cols:
    if pd.to_datetime(X[col], errors='coerce').notna().all():  # If
the column is a date
        X[col] = pd.to_datetime(X[col],
errors='coerce').astype(np.int64) // 10**9  # Convert to Unix
timestamp

# Apply label encoding to all categorical columns
for col in categorical_cols:
    X[col] = LabelEncoder().fit_transform(X[col].astype(str))

# Handle missing values
X[numerical_cols] = X[numerical_cols].apply(pd.to_numeric,
errors='coerce')
X[numerical_cols] =
X[numerical_cols].fillna(X[numerical_cols].median())

X[categorical_cols] =
X[categorical_cols].fillna(X[categorical_cols].mode().iloc[0])

# Apply scaling to numerical columns
scaler = StandardScaler()
if len(numerical_cols) > 0:  # Check if there are any numerical
columns
    X_scaled = X.copy()
    X_scaled[numerical_cols] =
scaler.fit_transform(X[numerical_cols])
else:
    X_scaled = X.copy()

# Oversample using SMOTE
smote = SMOTE(sampling_strategy='auto', random_state=42)
```

```
X_balanced, y_balanced = smote.fit_resample(X_scaled, y_encoded)

# Match species length after SMOTE
species_balanced = np.resize(species_encoded, len(X_balanced))

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test, sp_train, sp_test =
train_test_split(
    X_balanced, y_balanced, species_balanced, test_size=0.2,
stratify=y_balanced, random_state=42
)

# Model definitions
catboost = CatBoostClassifier(verbose=0, random_state=42)

rf = RandomForestClassifier(n_estimators=150,
class_weight='balanced', random_state=42)
gb = GradientBoostingClassifier(n_estimators=150, random_state=42)

# Voting Classifier
voting_clf = VotingClassifier(
    estimators=[('catboost', catboost), ('rf', rf), ('gb', gb)],
    voting='soft'
)

# Train the model
voting_clf.fit(X_train, y_train)

# Predict the test set results
y_pred = voting_clf.predict(X_test)

# Evaluate the model
print("Classification Report:")
print(classification_report(y_test, y_pred))

# Confusion Matrix
plt.figure(figsize=(6, 4))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d',
cmap='YlGnBu')
plt.title("Confusion Matrix - Voting Classifier")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()

# Per-species accuracy
species_accuracy = {}
sp_test_series = pd.Series(sp_test).reset_index(drop=True)
```

```
# Map encoded species back to original species names
species_mapping = {i: sp for i, sp in
enumerate(species_encoder.classes_)}

for sp_id in np.unique(sp_test_series):
    indices = sp_test_series[sp_test_series == sp_id].index
    acc = accuracy_score(y_test[indices], y_pred[indices])
    species_accuracy[sp_id] = acc

# Print per-species accuracy
print("\nPer-Species Accuracy:")
for sp_id, acc in species_accuracy.items():
    species_name = species_mapping.get(sp_id, "Unknown Species")
    print(f" {species_name}: {acc:.4f}")

# Overall Accuracy
print(f"\nOverall Accuracy: {accuracy_score(y_test, y_pred):.4f}")
```

APPENDIX 2

SCREENSHORT

```
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.97      0.91        60
           1       0.79      0.70      0.74        60
           2       0.75      0.73      0.74        60

    accuracy                           0.80       180
   macro avg       0.80      0.80      0.80       180
weighted avg       0.80      0.80      0.80       180
```

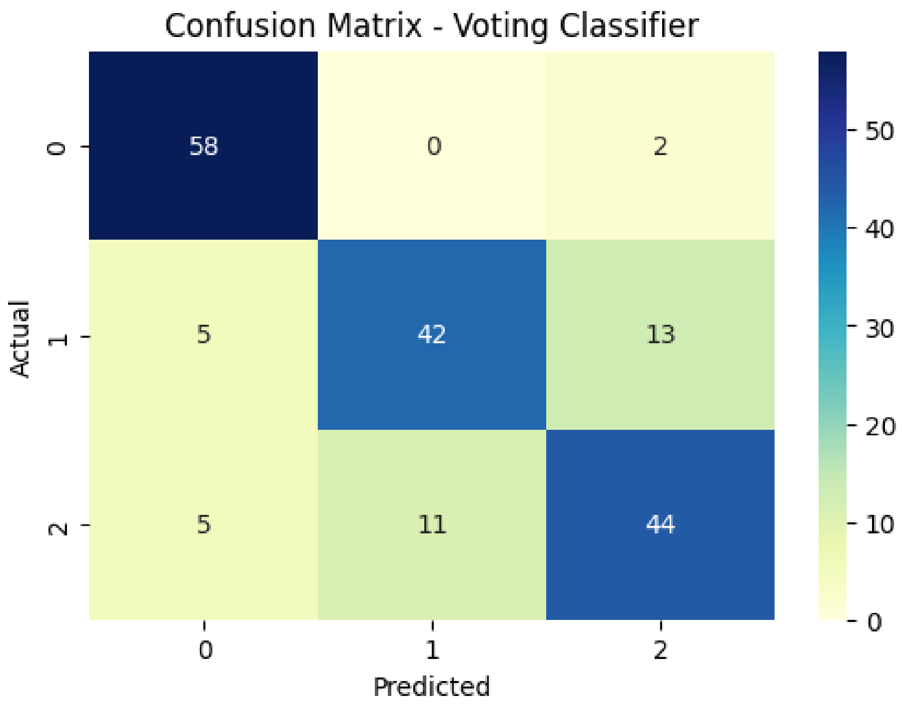
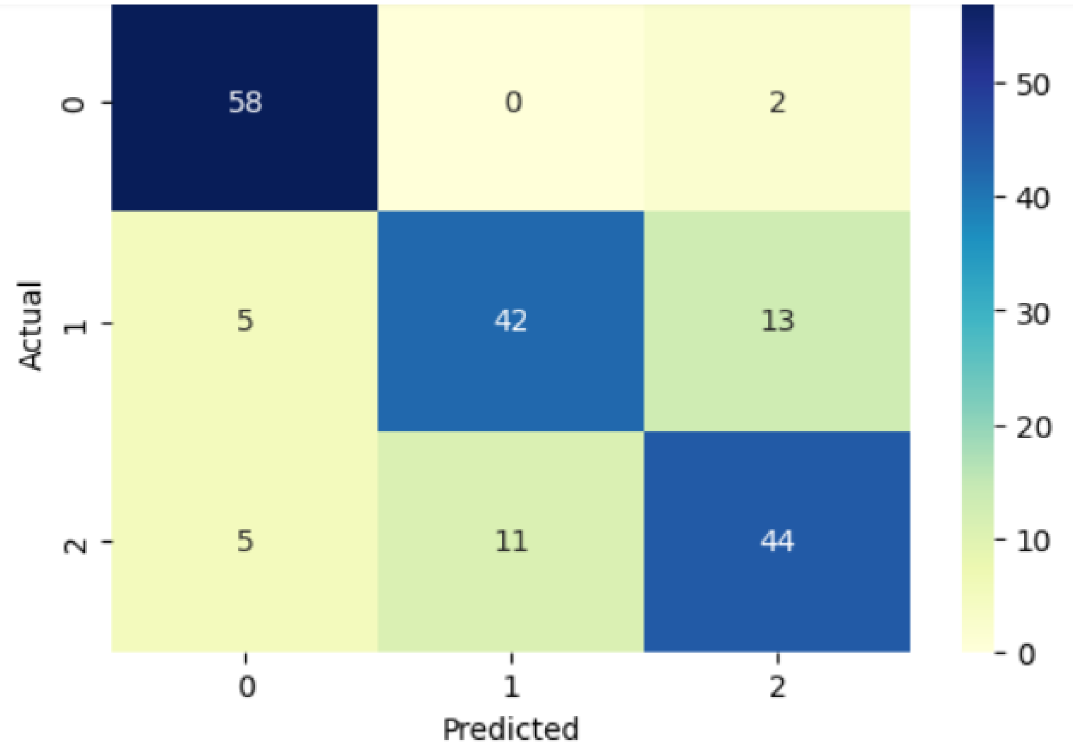


Fig A.1 Classification Report




```
Per-Species Accuracy:
 Amphibian: 0.8824
 Aquatic: 0.8333
 Bird: 0.8529
 Insect: 0.9048
 Mammal: 0.7273
 Reptile: 0.6250

Overall Accuracy: 0.8000
```

Fig A.2 Classification Report – Overall Accuracy

```
Training Random Forest...
Random Forest Classification Report:
              precision    recall  f1-score   support

 Crepuscular       0.96      1.00      0.98       180
     Diurnal       0.97      0.97      0.97       176
   Nocturnal       0.98      0.93      0.95       177

    accuracy                           0.97       533
   macro avg       0.97      0.97      0.97       533
weighted avg       0.97      0.97      0.97       533

Accuracy: 0.9681
```

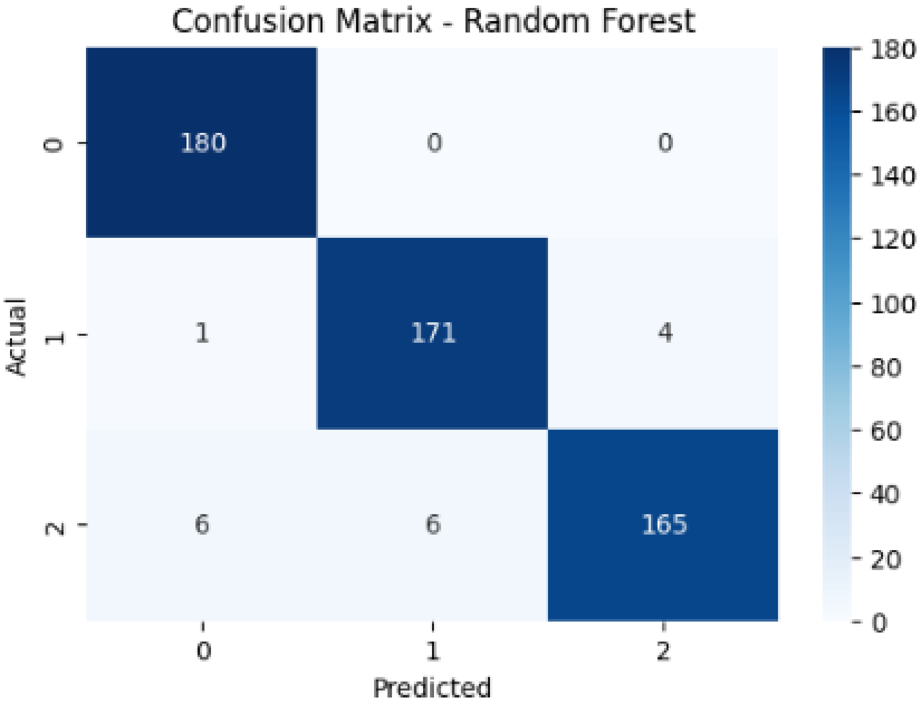


Fig A.3 Random Forest

```
Training Gradient Boosting...
Gradient Boosting Classification Report:
              precision    recall  f1-score   support

 Crepuscular       1.00      1.00      1.00       180
     Diurnal       0.96      0.99      0.97       176
   Nocturnal       0.99      0.95      0.97       177

    accuracy                           0.98       533
   macro avg       0.98      0.98      0.98       533
weighted avg       0.98      0.98      0.98       533

Accuracy: 0.9831
```

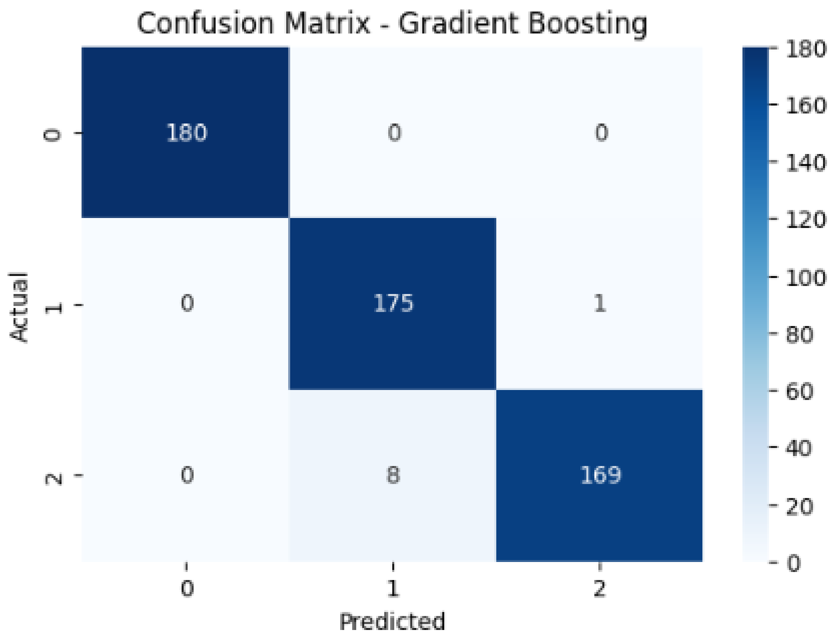


Fig A.4 Gradient Boosting

```
Training SVM...
SVM Classification Report:
              precision    recall  f1-score   support

 Crepuscular       0.96      0.99      0.98       180
     Diurnal       0.87      0.93      0.90       176
   Nocturnal       0.95      0.85      0.90       177

    accuracy                           0.92       533
   macro avg       0.93      0.92      0.92       533
weighted avg       0.93      0.92      0.92       533

Accuracy: 0.9250
```

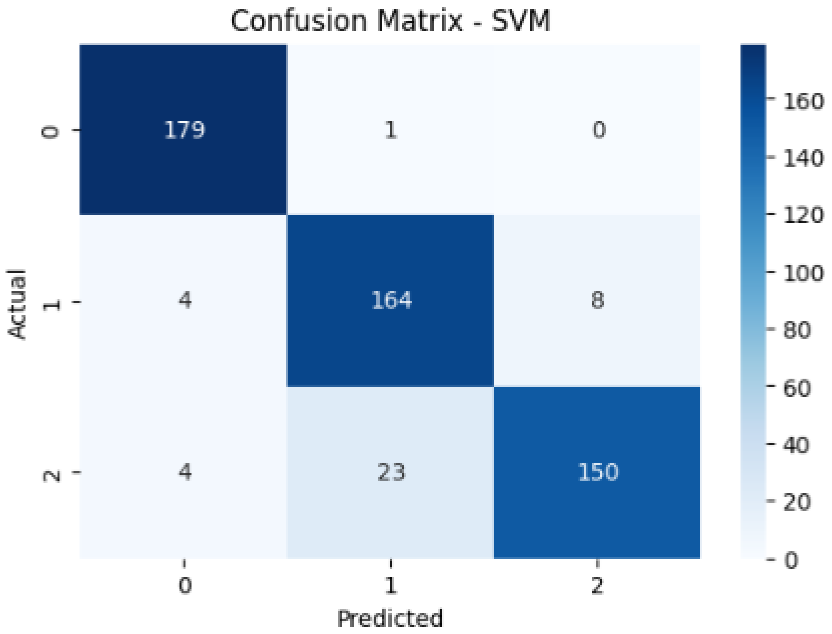


Fig A.5 SVM

```
Training CatBoost...
CatBoost Classification Report:
              precision    recall  f1-score   support

 Crepuscular       1.00      1.00      1.00       180
     Diurnal       0.97      1.00      0.99       176
   Nocturnal       1.00      0.97      0.99       177

    accuracy                           0.99       533
   macro avg       0.99      0.99      0.99       533
weighted avg       0.99      0.99      0.99       533

Accuracy: 0.9906
```

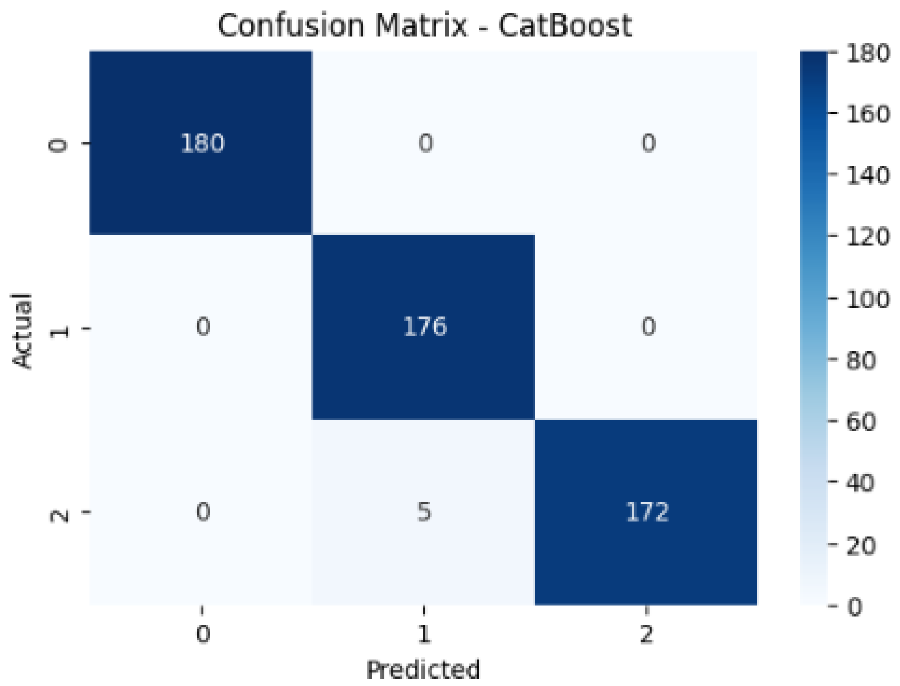


Fig A.6 CatBoost

```
Training Stacking Classifier...
Stacking Classifier Report:
              precision    recall  f1-score   support

 Crepuscular       1.00      0.99      0.99       180
     Diurnal       0.97      1.00      0.99       176
   Nocturnal       0.99      0.97      0.98       177

    accuracy                           0.99       533
   macro avg       0.99      0.99      0.99       533
weighted avg       0.99      0.99      0.99       533

Accuracy: 0.9869
```

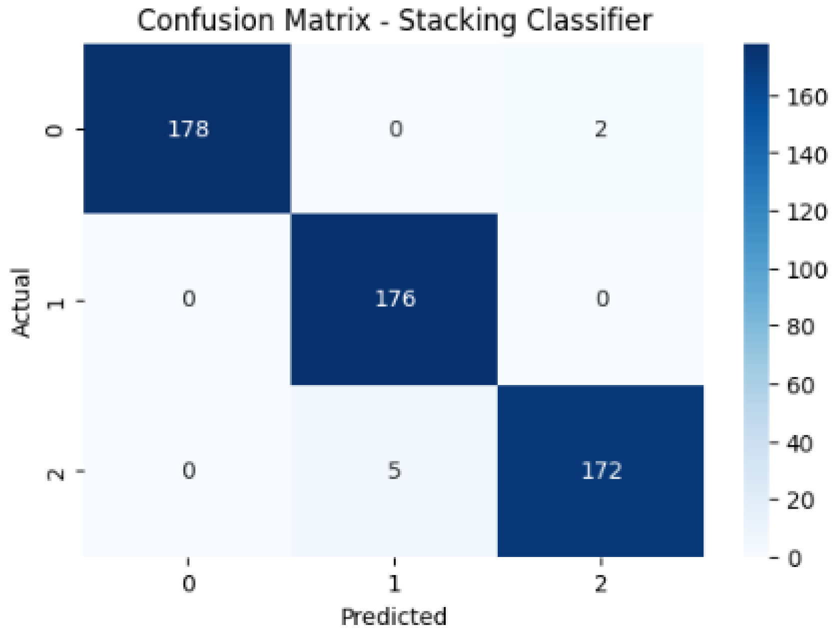


Fig A.7 Stacking Classifier

```
Accuracy per Species:
Diurnal: 0.8475
Nocturnal: 0.8305
Crepuscular: 0.9000
```

Fig A.8 Species Overall Accuracy

REFERENCES

[1] The paper titled "Brief Review of Biological Effects of Electromagnetic Pollution (RF and 5G Waves) on Humans, Animals, and Vegetation" was published in the December 2022 issue of the International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET). The DOI number for this publication is 10.15680/IJIRSET.1112001.

[2] "Tracking Devices for Pets: Health Risk Assessment for Exposure to Radiofrequency Electromagnetic Fields" by Judith Klune et al., published in Animals 2021, DOI 10.3390/ani11092721

[3] "Implications for 5G" published in Environmental Health in 2022, with the extensive discussion on safety standards, biological effects, and regulatory critiques, is the primary focus, DOI 10.1186/s12940-022-00900-9

[4] "Health effects of fifth-generation technologies" by Ramakrishnan and Athikary, published in the International Journal of Environmental Health Engineering in 2023, with DOI 10.4103/ijehe.ijehe_18_22.

[5] "Health Risks Associated With 5G Exposure: A View From the Communications Engineering Perspective," with DOI 10.1109/OJCOMS.2021.3106052.

[6] “Systematic Review of Radiofrequency Fields and Cancer Risk: Key Findings and Implications”, published in Environment International in 2025, with DOI 10.1016/j.envint.2025.109274.

[7] 5G wireless telecommunications expansion: Public health and environmental implications, Cindy L. Russell

[8] Development of health-based exposure limits for radiofrequency radiation from wireless devices using a benchmark dose approach, Uloma Igara Uche and Olga V. Naidenko

[9] “Electromagnetic fields (RF-EMF): Negative effects on the homing ability, but not on brood development or longevity”, DOI 10.1016/j.scitotenv.2023.165211

[10] Defined exposure of honey bee colonies to simulated radiofrequency electromagnetic fields (RF-EMF): Negative effects on the homing ability, but not on brood development or longevity, Manuel Treder, Marcus Müller, Larissa Fellner, Kirsten Traynor, Peter Rosenkranz, October 20, 2023, DOI 10.1016/j.scitotenv.2023.165211

[11] The effect of exposure to radiofrequency fields on cancer risk in the general and working population: A systematic review of human observational studies– Part II: Less researched outcomes, journal homepage: www.elsevier.com/locate/envint

[12] “Tinnitus and cell phones: the role of electromagnetic radiofrequency radiation”, Seyed Mohammad Javad Mortazavi, Seyed Mohammad Javad Mortazavi, DOI 10.1016/j.bjorl.2015.11.014