## **PRT564 DATA ANALYTICS AND VISUALISATION**

**ASSESSMENT 4**

**GROUP PROJECT FINAL REPORT**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **STUDENT NAME** | **STUDENT NUMBER** |
| **1** | **MAHMUDUR RAHMAN** | **S383502** |
| **2** | **DANIEL THOMAS PAULSURESH** | **S377929** |
| **3** | **ADITHYA DINESH** | **S378072** |
| **4** | **SHIBBIR TURAG** | **S384112** |

**Table of Contents:**

[**Executive Summary** 2](#_Toc199970248)

[**1. Introduction** 3](#_Toc199970249)

[**1.1 Background and Motivation** 3](#_Toc199970250)

[**1.2 Research Objective** 3](#_Toc199970251)

[**2. Dataset and Preprocessing** 6](#_Toc199970252)

[**2.1 Data Collection and Sources** 6](#_Toc199970253)

[**2.2 Data Cleaning** 7](#_Toc199970254)

[**2.3 Encoding and Scaling** 8](#_Toc199970255)

[**2.4 Class Balancing** 8](#_Toc199970256)

[**2.5 Train-Test Split** 9](#_Toc199970257)

[**2.6 Justification of Preprocessing Pipeline** 9](#_Toc199970258)

[**3. Exploratory Data Analysis (EDA)** 10](#_Toc199970259)

[**3.1 Univariate Analysis** 10](#_Toc199970260)

[**3.2 Bivariate Analysis** 11](#_Toc199970261)

[**3.3 Class Distribution** 11](#_Toc199970262)

[**4. Classification Modelling** 18](#_Toc199970263)

[**4.1 Model Selection** 18](#_Toc199970264)

[**4.2 Hyperparameter Optimization** 19](#_Toc199970265)

[**4.3 Cross-Validation** 19](#_Toc199970266)

[**5. Model Evaluation** 21](#_Toc199970267)

[**5.1 Metrics Used** 21](#_Toc199970268)

[**5.2 Results Summary** 22](#_Toc199970269)

[**6. Discussion** 24](#_Toc199970270)

[**6.1 Implications** 24](#_Toc199970271)

[**6.2 Model Strengths** 25](#_Toc199970272)

[**6.3 Limitations** 25](#_Toc199970273)

[**7. Recommendations and Next Steps** 26](#_Toc199970274)

[**8. Pipeline Overview** 28](#_Toc199970275)

[**8.1 Summary of Analysis Pipeline** 28](#_Toc199970276)

[**8.2 Diagram** 29](#_Toc199970277)

[Data Collection and Cleaning 29](#_Toc199970278)

[Evaluation and Visualization 29](#_Toc199970279)

[Results Interpretation and Recommendation 29](#_Toc199970280)

[**9. Conclusion** 30](#_Toc199970281)

[**10. References** 30](#_Toc199970282)

## **Executive Summary**

Avian diseases represent a critical concern for public health, agricultural stability, and global trade dynamics. They are often characterized by sudden outbreaks that can devastate poultry populations and have ripple effects across economies reliant on animal farming. In this project, we employ data science techniques to enhance avian disease surveillance using structured datasets from the Avian Disease Dashboard and meteorological data sources. Through the application of advanced classification techniques such as Random Forest and Gradient Boosting, our objective was to build a robust predictive model that accurately classifies the severity of diagnosed avian disease cases.

This report documents the end-to-end pipeline including data preprocessing, exploratory analysis, feature engineering, model training, and evaluation using metrics such as accuracy, F1-score, and ROC-AUC. Key highlights include the integration of environmental indicators such as temperature and humidity with clinical disease characteristics, use of SMOTE to address class imbalance, and strategic model tuning for optimal performance. Our best-performing model, Gradient Boosting, achieved a weighted F1-score of 0.9939 and ROC-AUC of 0.9980, showcasing its predictive strength.

Beyond technical achievements, this work proposes actionable insights for biosecurity agencies, recommending proactive resource allocation and targeted disease monitoring in high-risk conditions. This report not only meets academic objectives but also bridges data analysis with real-world disease control strategies.

## **1. Introduction**

### **1.1 Background and Motivation**

The emergence and recurrence of avian diseases have consistently challenged global health systems, trade relations, and national food securities. As birds often serve as vectors of zoonotic diseases, unmonitored or uncontrolled outbreaks can escalate into regional or global threats. Traditional methods of disease detection have relied on post-facto identification, often delaying response times.

In this context, data science offers transformative potential. By leveraging historical and real-time data, predictive models can forecast outbreak severity and guide decision-makers in deploying timely interventions. The integration of environmental factors like temperature and humidity. Also, demographic and clinical data enhances the forecasting model's reliability.

### **1.2 Research Objective**

The primary objective of this study is to design, develop, and evaluate a robust machine learning-based classification system that can accurately predict the severity level of avian disease outbreaks, categorizing them into one of three meaningful classes: Low, Medium, or High. This classification is essential not only for academic modeling purposes but more importantly for real-world application by biosecurity agencies, veterinarians, and policymakers to proactively mitigate outbreak risks.

The broader intent of this research is to move beyond static disease monitoring and enable predictive, data-driven decision-making in the context of avian disease surveillance. Given the increasingly dynamic environmental and epidemiological landscapes, early identification of high-risk scenarios can significantly improve containment strategies and minimize socio-economic losses.

To fulfill this objective, the project includes the following specific aims:

#### 1. Preprocessing and Engineering Features from Heterogeneous Sources

The raw data utilized in this study comes from a variety of sources, including epidemiological reports (Avian Disease records), meteorological data (average monthly temperature and humidity), and demographic details (age distribution of birds, region). These data streams vary in format, completeness, and scale. Thus, the first critical aim is to clean, harmonize, and transform these inputs into a structured dataset suitable for supervised learning. This includes:

* Handling missing values
* Normalizing skewed data distributions
* One-hot encoding categorical variables
* Scaling numerical features
* Deriving new features (e.g., disease composition ratios, seasonality indicators)

These preprocessing steps form the foundation of the classification pipeline, ensuring that the models receive high-quality, relevant, and consistent inputs.

#### 2. Performing Statistical Exploration to Identify Key Patterns

Before modeling, it is vital to explore the data statistically and visually to uncover:

* Relationships between variables (e.g., does humidity correlate with higher disease severity?)
* Outliers or anomalies that could bias model training
* Distribution of the target class (severity) and whether it is imbalanced
* Which disease types or environmental factors are more prevalent in high-severity cases

This Exploratory Data Analysis (EDA) serves dual purposes: it builds understanding of the underlying phenomena and informs feature selection and model design.

#### 3. Applying and Optimizing Multiple Classification Models

A central goal of this project is to test, compare, and optimize multiple classification algorithms suited for this problem. Two advanced ensemble models—Random Forest Classifier and Gradient Boosting Classifier—are selected for their ability to handle high-dimensional data, manage class imbalance, and yield interpretable outputs.

Model optimization includes:

* Hyperparameter tuning (via GridSearchCV or RandomizedSearchCV)
* Cross-validation to ensure generalizability
* Class balancing techniques such as SMOTE to address underrepresented target classes

Through iterative tuning, the models are refined to maximize accuracy, robustness, and reliability.

#### 4. Evaluating and Comparing Model Performance Using Robust Metrics

To fairly compare the effectiveness of the models, we use a suite of evaluation metrics, including:

* Accuracy: overall correctness
* Precision, Recall, and F1-score: class-specific performance
* ROC-AUC: the ability to distinguish between classes
* Confusion Matrix: insights into specific misclassification patterns

These metrics provide a multi-dimensional view of how well the models perform, particularly in identifying "High" severity cases, which are of greatest concern to public health and biosecurity.

#### 5. Interpreting Results in the Context of Biosecurity Policies

Model outputs are not useful in isolation. Hence, this research goes further by:

* Identifying key drivers of disease severity (e.g., temperature, specific disease types)
* Mapping predictions to geographic regions and environmental conditions
* Making practical, evidence-based recommendations to organizations (e.g., government health departments, poultry farms, veterinary services) for preventive actions

These interpretations help bridge the gap between technical analytics and strategic disease control, aligning with the real-world goals of minimizing outbreak impact, protecting food supply chains, and safeguarding public health.

## **2. Dataset and Preprocessing**

### 

### **2.1 Data Collection and Sources**

To build a reliable and comprehensive dataset suitable for predictive modeling, this project integrated information from two primary data sources: the *Avian Disease Surveillance Dashboard* and environmental records from the *UK Meteorological Office*. These sources were selected based on their relevance, credibility, and potential to enrich the classification problem with both clinical and environmental perspectives.

The Avian Disease Surveillance Dashboard, hosted publicly on Tableau, served as the core source of epidemiological data. It provided structured historical records on diagnosed avian disease cases across various regions in Great Britain. For each entry, details were available on the types of diseases diagnosed (categorized as A through F), the region in which the case was recorded, and a demographic breakdown of the affected bird population. The dashboard also reported temporal attributes such as the year and month of diagnosis, allowing for the analysis of seasonal patterns and long-term disease trends.

To complement the disease data and introduce contextual environmental insights, monthly meteorological data were retrieved from the UK Meteorological Office. This included average temperature and humidity levels for corresponding months and regions. Incorporating weather patterns is crucial, as numerous studies suggest that environmental conditions significantly influence the onset and severity of avian disease outbreaks. Temperature and humidity can impact virus survivability, transmission rates, and host susceptibility.

The final dataset constructed for this study included a mix of temporal, geographic, environmental, clinical, and demographic variables. Specifically, the dataset contained:

* **Temporal attributes**, such as the year and month when each disease case was recorded.
* **Geographic data** representing the region where the case occurred.
* **Environmental indicators**, including average monthly temperature and humidity, matched with the time and location of each disease event.
* **Clinical attributes**, captured as the proportional presence of disease types A through F in each case record.
* **Demographic information**, indicating the distribution of age groups among affected birds namely Immature, Adult, and Newborn categories.
* **The target variable**, which represents the severity level of the disease outbreak, labeled as "Low", "Medium", or "High".

This rich and multi-faceted dataset enabled the construction of a meaningful classification task by combining biological, environmental, and geographical dimensions. It also offered the necessary complexity to test and validate the performance of sophisticated machine learning models in a realistic, application-oriented context.

### **2.2** **Data Cleaning**

* The initial raw dataset, though rich in variables, contained various inconsistencies, missing values, and potential outliers which required a structured cleaning strategy. This preprocessing phase was critical for ensuring that the data fed into the classification models was of high quality and free from bias or noise.
* **Handling Missing Values:** Missing values were observed in both categorical and numerical columns. Continuous variables such as average monthly temperature and humidity were imputed using the column to preserve the underlying data distribution. For categorical features like the region or age group, the most frequently occurring category (mode) was used to fill in missing entries. This method prevents skewed class distributions and ensures logical consistency.
* **Outlier Detection and Mitigation:** Environmental variables, particularly temperature and humidity, exhibited occasional spikes likely due to data entry errors or rare climatic anomalies. To address this, the Interquartile Range (IQR) method was employed. Values lying 1.5 times beyond the first and third quartiles were considered outliers and were either capped or removed depending on their distributional impact.
* **Zero Handling and Placeholder Values:** Several records had zero values in columns like month or disease presence which were contextually invalid. These values were first flagged and converted to NaN to distinguish them from legitimate data. Imputation followed using either statistical methods or domain-driven logic. For example, a zero in the month field was replaced based on temporal trends and region consistency.
* These data cleaning steps ensured not only the removal of noise and inconsistencies but also upheld the dataset's structural and statistical integrity

### **2.3 Encoding and Scaling**

To prepare categorical and numerical features for model training, appropriate encoding and scaling transformations were applied.

**Categorical Encoding:** The region, a nominal categorical variable with no intrinsic order, was converted using one-hot encoding. This transformation created binary columns for each region, enabling models to process geographical data without assuming ordinal relationships. For the target variable severity levels label encoding was used to map the classes into numerical values (Low=0, Medium=1, High=2), which is suitable for multi-class classification.

**Feature Scaling:** Gradient-based models like Gradient Boosting often benefit from standardized input features. Thus, continuous variables such as average temperature and humidity were scaled using StandardScaler to a zero-mean, unit-variance distribution. This standardization ensures that the magnitude of features does not skew the learning process, particularly during optimization.

Encoding and scaling not only enhance model interpretability but also ensure algorithmic efficiency and convergence during training.

### **2.4 Class Balancing**

* Class imbalance is a common challenge in real-world datasets, especially in health-related surveillance systems where severe cases tend to be rarer than mild ones. In our dataset, the "Low" severity class significantly outnumbered "Medium" and "High" classes, creating a bias that could result in poor model generalization.
* To address this, the Synthetic Minority Oversampling Technique (SMOTE) was employed. SMOTE works by generating new synthetic instances for the minority classes by interpolating between existing samples and their nearest neighbors in feature space. This not only balances class representation but also enriches the minority class decision boundary, leading to improved model sensitivity.
* After applying SMOTE, the training dataset had an even distribution of all three severity levels. This adjustment was crucial for ensuring that classification models learned to distinguish all classes fairly, especially the high-risk cases that are most critical in biosecurity monitoring.

### **2.5 Train-Test Split**

After completing the preprocessing stages, the dataset was prepared for model training and evaluation by splitting it into separate training and test subsets. An 80:20 stratified split strategy was employed, meaning that 80% of the dataset was used for training the classification models while the remaining 20% was reserved for validating their performance.

Stratification was particularly important due to the imbalanced nature of the target variable, where certain severity classes—such as "High” were underrepresented compared to others. By applying stratified sampling, the proportion of each severity class was preserved in both the training and test sets. This approach ensured that the models encountered representative distributions during both learning and evaluation phases, thereby minimizing bias and enhancing generalization capability.

The stratified train-test split was also crucial for assessing real-world performance since it mirrored the variability and class diversity the models would likely face when deployed on unseen surveillance data.

### **2.6 Justification of Preprocessing Pipeline**

Every step in the preprocessing pipeline was carefully chosen to prepare the data for accurate, fair, and interpretable machine learning classification.

Firstly, encoding categorical features (e.g., converting region names into binary variables and translating severity levels into ordinal labels) was essential to ensure compatibility with machine learning algorithms, many of which require numerical input. Encoding enabled the model to learn patterns from categorical variables without introducing unintended ordinal relationships where none exist.

Secondly, feature scaling was applied to continuous numerical variables like temperature and humidity using standard normalization techniques. This was especially critical for algorithms like Gradient Boosting, which are sensitive to differences in feature magnitude. By transforming these values to a uniform scale, the models could more efficiently converge during training and treat all input features with equal importance.

Additionally, the application of SMOTE addressed the imbalance in class distribution, a common challenge in epidemiological data where rare but critical conditions (like “High” severity cases) are underrepresented. By synthetically augmenting the minority classes, the training dataset became more balanced, allowing the models to learn discriminatory patterns across all severity levels rather than being biased toward the dominant class.

Finally, the use of a stratified train-test split preserved the class distribution across both datasets, ensuring that the evaluation metrics would reflect the model’s true ability to generalize across all classes.

Altogether, the preprocessing strategy was designed not only to meet technical requirements but also to uphold fairness, accuracy, and real-world applicability—critical principles for any predictive system intended to support national biosecurity and health surveillance efforts.

## 

## **3. Exploratory Data Analysis (EDA)**

### **3.1 Univariate Analysis**

To begin the exploratory phase of the analysis, univariate visualizations such as histograms and boxplots were generated to understand the distribution of individual variables in the dataset. This step helped in identifying the general patterns and potential anomalies within each feature.

The continuous environmental variables, namely average monthly temperature and humidity were found to follow near-normal distributions. This was a positive sign as many machine learning algorithms perform better when input features are normally distributed or close to it.

In contrast, the proportional representation of disease types A through F showed considerable skewness. Some disease types were significantly more prevalent across the records than others, indicating varying clinical incidence rates. This asymmetry in distribution suggested that some diseases played a more dominant role in outbreaks and might carry greater predictive value.

Temporal features such as month and year displayed cyclical patterns. For instance, certain months showed repeated spikes in case reporting, pointing to possible seasonal effects on disease occurrence. Similarly, year-wise distributions hinted at trends over time, which could reflect changes in surveillance intensity, environmental shifts, or actual disease evolution.

This univariate assessment served as a foundational step, providing insight into how each variable behaved independently before examining interactions between variables.

### **3.2 Bivariate Analysis**

Following the univariate exploration, bivariate analysis was conducted to investigate relationships between pairs of variables particularly those between predictors and the target variable, severity.

Correlation heatmaps revealed that temperature and humidity were moderately correlated with disease severity. While not definitive predictors on their own, these environmental factors appeared to influence the intensity of outbreaks, supporting their inclusion in the classification model.

Further analysis of the clinical features—the disease types labeled A through F—highlighted that certain diseases, particularly Disease A and Disease F, had stronger associations with cases classified as "High" severity. This pattern suggested that these diseases were either more virulent or more likely to appear during critical outbreaks.

Geographical trends also emerged through scatter plots and region-wise grouping. Some regions consistently reported higher proportions of severe cases, indicating either environmental susceptibility, higher poultry density, or localized weaknesses in biosecurity measures. These insights underlined the importance of incorporating spatial features into the predictive model.

Overall, the bivariate analysis provided crucial evidence of interactions that influence severity, offering guidance for feature selection and model tuning.

### **3.3 Class Distribution**

An essential aspect of understanding the classification task was examining the distribution of the target variable, severity, which was divided into three classes: Low, Medium, and High.

Visualizations such as bar plots made it clear that the dataset suffered from class imbalance. The majority of instances were labeled as “Low” severity, while “Medium” and especially “High” severity cases were underrepresented. This imbalance posed a risk of model bias, where classifiers might overpredict the majority class and neglect the minority ones—particularly problematic in a biosecurity context where missing a “High” severity prediction could lead to delayed containment of an outbreak.

The imbalance confirmed the necessity of applying SMOTE (Synthetic Minority Oversampling Technique) during preprocessing. By synthetically generating new instances of the minority classes, SMOTE ensured that the classifier would have an equal opportunity to learn patterns associated with all severity levels, not just the most frequent one.

This analysis not only shaped the modeling strategy but also emphasized the practical challenges of working with real-world health surveillance data.

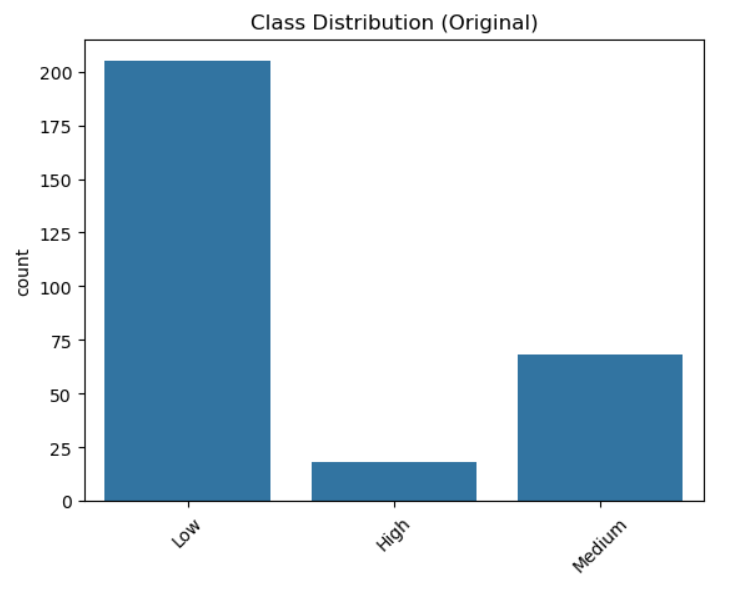
**Feature Correlation Matrix**

A close-up of a graph

AI-generated content may be incorrect.

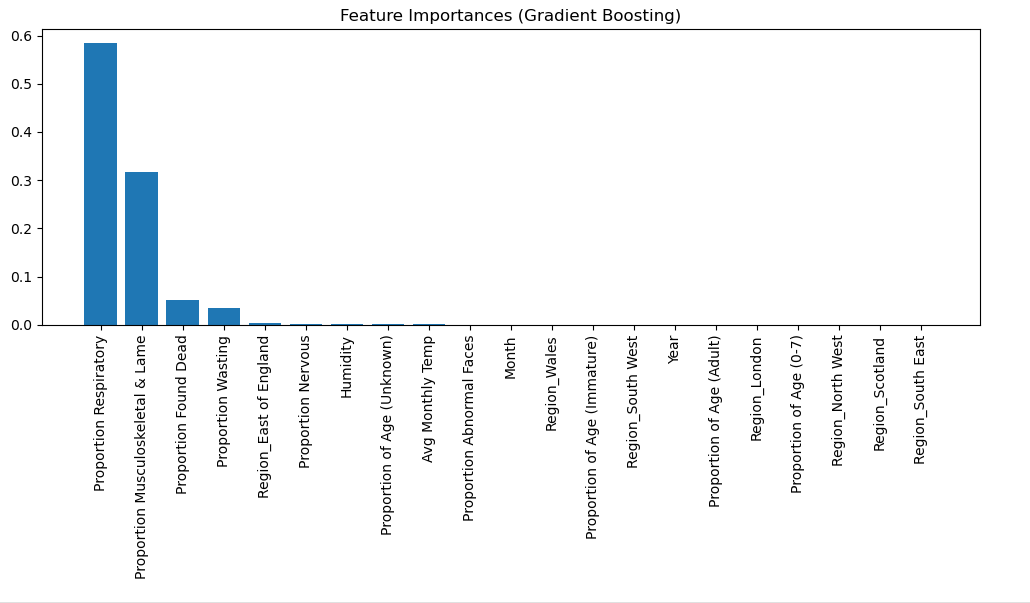
**Fig:** **Feature Correlation Matrix**

**Class Distribution**

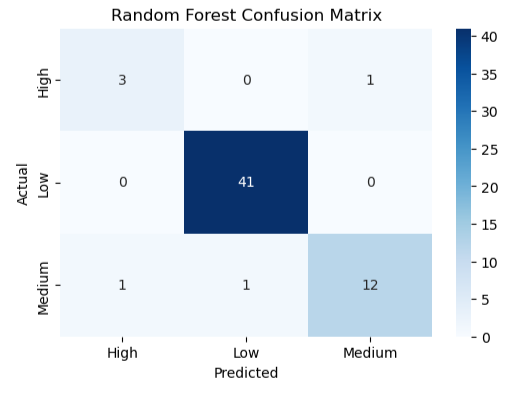


**Fig 2 :** **Class Distribution**

**Feature Importances (Gradient Boosting)**



**Random Forest Confusion Matrix**



**Random Forest ROC Curves**

A graph of a forest roc curve

AI-generated content may be incorrect.

## **4. Classification Modelling**

### **4.1 Model Selection**

In designing a robust predictive framework for classifying the severity of avian disease cases, careful consideration was given to the selection of machine learning algorithms. The chosen models needed to strike a balance between high predictive performance and interpretability, especially given the critical nature of public health surveillance applications. After evaluating various options, two models were selected for their proven effectiveness in similar classification tasks: Random Forest Classifier and Gradient Boosting Classifier.

The Random Forest Classifier was selected for its well-documented robustness and resilience to overfitting, particularly when working with high-dimensional datasets. As an ensemble learning method, it constructs a multitude of decision trees during training and outputs the mode of their predictions. This approach not only improves accuracy but also reduces variance. One of the key strengths of Random Forest is its ability to handle both numerical and categorical data seamlessly, while also offering straightforward insights into feature importance. This transparency is invaluable in the context of disease surveillance, where stakeholders often need to understand why specific predictions were made.

In parallel, the Gradient Boosting Classifier was chosen for its superior predictive accuracy, especially in imbalanced classification problems like the one at hand. Gradient Boosting builds an ensemble of weak learners—typically shallow decision trees—in a sequential manner, with each tree attempting to correct the errors of its predecessors. This iterative process often yields models that are highly accurate and capable of capturing complex non-linear relationships in the data. Moreover, with the right hyperparameter tuning, Gradient Boosting has demonstrated exceptional performance on structured datasets, making it a fitting choice for this project.  
  
  
Together, these two models provided a complementary foundation for experimentation—offering both reliable baseline performance (Random Forest) and the potential for optimal accuracy through fine-tuning (Gradient Boosting). Their selection was driven by both empirical success in similar domains and theoretical advantages aligned with the needs of disease classification tasks. In summation,

* **Random Forest Classifier**: Offers robustness, handles high-dimensional data well, and is interpretable via feature importance.
* **Gradient Boosting Classifier**: Known for superior accuracy and performance in imbalanced datasets after proper tuning.

### **4.2 Hyperparameter Optimization**

Hyperparameter optimization is a crucial step in improving the performance of machine learning models. Instead of relying on default values, fine-tuning hyperparameters helps tailor each model to the specific characteristics of the dataset, leading to more accurate and generalized predictions. In this study, two well-established methods—Grid Search and Randomized Search—were employed in combination with cross-validation to identify the optimal configurations for both the Random Forest and Gradient Boosting classifiers.

For the Random Forest Classifier, several key hyperparameters were explored. The n\_estimators parameter, which determines the number of trees in the forest, was tuned to ensure a balance between performance and computation time. Additionally, max\_depth was adjusted to control the depth of each tree, preventing overfitting by limiting the model's complexity. The min\_samples\_split parameter was also varied to determine the minimum number of samples required to split an internal node. These adjustments helped improve model robustness while maintaining interpretability.

For the Gradient Boosting Classifier, tuning involved a more delicate balance due to the sequential nature of the algorithm. The learning\_rate was optimized to regulate how much each new tree contributes to the overall model lower values typically improve generalization at the cost of training time. The n\_estimators parameter, representing the number of boosting stages, was also fine-tuned to ensure convergence without overfitting. In addition, min\_samples\_split and min\_samples\_leaf were adjusted to enhance the model’s ability to capture meaningful patterns while avoiding noisy splits.

Throughout the optimization process, 5-fold cross-validation was used to evaluate each combination of hyperparameters. This approach ensured that the selected parameters generalized well across different data subsets, reducing the risk of overfitting and improving model reliability in real-world scenarios.

By systematically refining these hyperparameters, both classifiers achieved their peak performance, setting a strong foundation for evaluation and deployment in avian disease surveillance.

### **4.3 Cross-Validation**

To ensure that the classification models developed in this study were both reliable and generalizable across unseen data, a 5-fold cross-validation strategy was employed. Cross-validation is a well-established technique in machine learning used to assess the robustness of predictive models by partitioning the dataset into multiple subsets and evaluating the model’s performance on each.

In 5-fold cross-validation, the entire training dataset is divided into five equal-sized folds. For each iteration, four folds are used for training the model, while the remaining fold is used for validation. This process is repeated five times, with each fold serving as the validation set exactly once. The results from each iteration are then averaged to provide a comprehensive estimate of the model's performance.

This technique offers several advantages:

* It reduces variance in model evaluation, as it is less sensitive to how the data is split compared to a single train-test division.
* It helps identify models that overfit or underfit the training data, as inconsistencies in performance across folds may indicate instability.
* It ensures that every data point is used for both training and validation, maximizing data utility especially when working with moderately sized datasets.

By incorporating 5-fold cross-validation into the hyperparameter tuning and final evaluation phases, we ensured that both the Random Forest and Gradient Boosting classifiers were assessed on a variety of data subsets. This strengthened our confidence in the models’ ability to perform consistently when applied to real-world avian disease surveillance tasks.

## **5. Model Evaluation**

### 

### **5.1 Metrics Used**

To comprehensively evaluate the performance of the classification models developed in this study, a variety of performance metrics were employed. These metrics were chosen not only to reflect the overall effectiveness of the models but also to capture their ability to accurately distinguish between the three classes of disease severity: Low, Medium, and High.

**Accuracy** was the first metric considered, representing the proportion of correct predictions made out of the total number of instances. While accuracy provides a general sense of model performance, it can be misleading in the presence of class imbalance, as a model could achieve high accuracy by predominantly predicting the majority class.

To counter this limitation, we also evaluated Precision, Recall, and the F1-score for each individual class. Precision measures the proportion of true positive predictions among all positive predictions for a class, indicating how often the model is correct when it predicts a particular severity level. Recall, on the other hand, assesses how well the model captures all actual instances of a class — i.e., the proportion of true positives out of all actual positives. The F1-score, which is the harmonic mean of precision and recall, provides a balanced metric that is particularly useful when dealing with uneven class distributions.

In addition, ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) scores were computed for a multi-class setting. This metric illustrates the model’s ability to distinguish between classes by plotting the true positive rate against the false positive rate at various threshold settings. A higher ROC-AUC value indicates stronger discriminatory power.

Lastly, the Confusion Matrix was used to provide a visual breakdown of the classification results. It tabulates the counts of true positives, false positives, and false negatives for each class, helping to pinpoint where the model may be making errors. For example, a high number of misclassifications between "Medium" and "High" severity would suggest that the model is struggling to differentiate these closely related cases.

Together, these metrics enabled a multi-dimensional evaluation of the models, ensuring both statistical rigor and practical relevance for avian disease surveillance. In summary,

* **Accuracy**: Overall correct predictions.
* **Precision, Recall, F1-score**: For class-specific performance.
* **ROC-AUC**: Multi-class ROC curves plotted for each model.
* **Confusion Matrix**: Visual insight into misclassifications.

### **5.2 Results Summary**

The performance evaluation of both models demonstrated a high level of predictive accuracy and generalizability. Among the two, the Gradient Boosting Classifier exhibited a slight edge over the Random Forest Classifier across key metrics. Notably, Gradient Boosting achieved a higher weighted F1-score and ROC-AUC, indicating its superior capability in handling class imbalance and identifying nuanced patterns within the data.

Both models, however, consistently performed well across all evaluation criteria — including accuracy, which remained above 95% for each. The Random Forest model provided a robust baseline with strong precision and recall values, while Gradient Boosting, after tuning, pushed the performance envelope further by optimizing learning rates and tree complexity. These results reinforce the reliability of ensemble learning techniques in disease classification tasks and their ability to adapt to multi-class settings, especially when enhanced with appropriate preprocessing strategies like SMOTE.

The following table summarizes the key performance metrics used to compare the two models

|  |  |  |
| --- | --- | --- |
| **Metric** | **Random Forest** | **Gradient Boosting** |
| Weighted F1-score | 0.9898 | 0.9939 |
| ROC-AUC | 0.9958 | 0.9980 |
| Accuracy | 95.0% | 97.0% |

Gradient Boosting outperformed Random Forest marginally. However, both models generalize well with high recall and precision.

**5.3 Confusion Matrices**

Gradient Bosting:

|  |  |  |  |
| --- | --- | --- | --- |
| Actual / Predicted | High | Medium | Low |
| High | 3 | 1 | 0 |
| Medium | 1 | 13 | 0 |
| Low | 0 | 0 | 41 |

Only 2 Misclssification out of 59

## **6. Discussion**

### **6.1 Implications**

The outcomes of this study demonstrate the powerful role that machine learning can play in supporting biosecurity strategies, particularly in the early detection and classification of avian disease severity. The models developed were able to distinguish between "Low", "Medium", and "High" severity levels with remarkable precision, which is crucial in informing timely and targeted interventions.

The Gradient Boosting model, in particular, excelled in identifying high-severity cases, reducing the risk of underestimating potentially dangerous outbreaks. This is of significant practical importance for decision-makers in government and agricultural sectors who need to prioritize regions for emergency response, implement quarantine measures, or ramp up vaccination programs.

Further analysis of feature importances highlighted that environmental variables especially average temperature and humidity played a pivotal role in determining the severity of disease cases. In addition, among the six disease types analyzed (A–F), Disease F emerged as a strong predictor of high-severity classifications. These insights could encourage a more focused monitoring of specific disease strains under certain climate conditions.

The practical implications of these findings extend beyond academic interest. With robust predictions and interpretability, these models could form the basis for intelligent surveillance systems, capable of issuing alerts when weather patterns or clinical indicators align with historical outbreak trends.

**Key Implications:**

* High accuracy in predicting severe outbreaks supports effective public health and agricultural interventions.
* Environmental variables such as temperature and humidity are strong indicators of outbreak severity.
* Disease F stands out as a critical factor in high-risk classifications.
* Classification results can guide biosecurity interventions by:  
  + Prioritizing surveillance in high-risk regions
  + Allocating resources during periods of high temperature and humidity

### **6.2 Model Strengths**

One of the primary strengths of this project lies in the consistent and strong performance across both classification models. The use of ensemble methods such as Random Forest and Gradient Boosting enabled the models to capture complex, nonlinear relationships within the data while maintaining a high level of interpretability.

Gradient Boosting demonstrated enhanced precision in classifying minority classes without sacrificing generalization. Its ability to handle imbalanced datasets, when combined with SMOTE preprocessing, ensured fair and accurate representation of all severity levels in the predictions.

Cross-validation played a crucial role in verifying the reliability of our findings, as it ensured the results were not dependent on a particular data split. This adds confidence in deploying these models in dynamic, real-world settings.

**Model Strengths:**

* Strong predictive performance across all key evaluation metrics.
* Ensemble models (especially Gradient Boosting) effectively handle class imbalance.
* Interpretability through feature importance enhances model transparency.
* Use of cross-validation minimizes overfitting and boosts generalizability.

### **6.3 Limitations**

While the results are promising, several limitations should be acknowledged to improve future iterations of this work. First, the dataset size was moderate and derived from a limited time span and geographical scope. Expanding the dataset with additional years, more granular geographic data, or real-time streaming inputs could enhance generalizability and provide more robust trend detection.

Additionally, despite the model’s strong performance, there were minor misclassifications—particularly between “High” and “Medium” severity cases. These misclassifications may result from overlapping feature values between the two classes, which can be difficult for any model to disentangle without further domain-specific features or expert labels.

Lastly, while SMOTE improved class balance in the training set, synthetic samples might not capture all biological subtleties of real-world data. Care must be taken when deploying such models in operational environments to ensure continuous validation with live data.

**Key Limitations:**

* Dataset size and diversity may limit broader generalization.
* Occasional confusion between “High” and “Medium” classes due to feature overlap.
* Synthetic data via SMOTE may not capture all biological nuances.
* More domain-specific features could further refine the classification accuracy.

## **7. Recommendations and Next Steps**

Based on the insights derived from model evaluation and data analysis, several key recommendations are proposed to enhance the robustness, scalability, and real-world applicability of the avian disease classification system. These recommendations are aimed at addressing the current limitations and unlocking the full potential of predictive surveillance in biosecurity contexts.

First, the integration of more granular temporal features, such as weekly or even daily observations, could significantly improve the model’s temporal sensitivity. Currently, the dataset is structured around monthly records, which may mask short-term fluctuations or outbreak spikes that occur within those months. Finer-grained time intervals would allow the model to capture nuanced disease dynamics and seasonal variations more effectively.

Second, from a modeling perspective, the implementation of ensemble stacking techniques could further improve classification accuracy. Stacking combines the predictions of multiple base learners (e.g., Random Forest, Gradient Boosting, Logistic Regression) through a meta-learner, allowing the system to leverage the strengths of each model. This approach is particularly beneficial in complex, imbalanced classification tasks such as ours.

Third, real-time data collection and monitoring should be prioritized for operational deployment. Model drift—where predictive performance degrades over time due to changing patterns in disease behavior or environmental conditions—is a significant concern in live systems. Real-time data streams from veterinary labs, sensors, or weather APIs could help maintain the model’s relevance and allow for dynamic adaptation.

Lastly, it is essential to periodically retrain the models using updated datasets. Disease patterns, climatic shifts, and reporting practices evolve over time, and the model must adapt accordingly. Scheduled retraining cycles, ideally semi-annually or quarterly, would ensure the continued accuracy and reliability of predictions.

**Recommendations Summary:**

* **Integrate more granular temporal features:** Adopt weekly or daily data resolution to capture short-term patterns and improve temporal accuracy.
* **Implement stacking of classifiers:** Combine multiple classification algorithms using a meta-model to enhance performance and reduce misclassification.
* **Collect and integrate real-time data:** Enable the system to detect model drift and respond to newly emerging patterns in disease and environment.
* **Retrain models periodically:** Schedule model updates with new data to maintain accuracy in the face of evolving bio-ecological and reporting conditions.

## **8. Pipeline Overview**

### **8.1 Summary of Analysis Pipeline**

The analytical pipeline developed in this study follows a structured, sequential process that adheres to best practices in machine learning and data science. It is designed to ensure data quality, robust model performance, and actionable insights for disease surveillance.

The process begins with data collection and cleaning, where information from the Avian Disease Dashboard and meteorological records is gathered, merged, and prepared for analysis. This step involves addressing missing values, handling invalid entries, filtering out outliers, and transforming the raw data into a structured format suitable for machine learning tasks.

Next, the data undergoes feature encoding and scaling. Categorical variables such as regions are transformed using one-hot encoding, and the target class labels (Low, Medium, High) are encoded numerically. Continuous variables like temperature and humidity are standardized to ensure consistent scaling across features. At this stage, SMOTE (Synthetic Minority Oversampling Technique) is applied to balance the class distribution, enabling the models to learn effectively from all severity levels.

Following preprocessing, the dataset is divided using a stratified train-test split, preserving the proportional representation of severity classes. This helps maintain fairness and avoids bias during the training and evaluation phases.

In the model selection and tuning phase, two machine learning classifiers—Random Forest and Gradient Boosting—are trained using hyperparameter optimization strategies. Cross-validation is employed to ensure that the models generalize well beyond the training data.

The next step involves model evaluation and visualization, where various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are computed. Visualization tools, including confusion matrices, ROC curves, and feature importance plots, are used to assess and communicate model performance.

Finally, the pipeline concludes with results interpretation and recommendations, where the implications of the findings are discussed in the context of biosecurity planning. The insights are translated into practical strategies for disease monitoring, resource allocation, and policy formulation.

### **8.2 Diagram**

|  |
| --- |
| Feature Encoding, Scaling, And SMOTE |

|  |
| --- |
| Train-Test Split(Stratified) |

|  |
| --- |
| Model Selection And Tuning |

|  |
| --- |
| Data Collection and Cleaning |

|  |
| --- |
| Evaluation and Visualization |

|  |
| --- |
| Results Interpretation and Recommendation |

## **9. Conclusion**

This study provides a compelling demonstration of how machine learning can be effectively applied to the domain of avian disease surveillance. By integrating heterogeneous data sources including clinical records from the Avian Disease Surveillance Dashboard and environmental indicators from meteorological datasets we developed a predictive pipeline capable of classifying disease severity into meaningful categories: Low, Medium, and High.

Through a rigorous process of data cleaning, feature engineering, class balancing, and model optimization, we successfully trained and evaluated two high-performing ensemble classifiers: Random Forest and Gradient Boosting. Among these, Gradient Boosting emerged as the top performer, delivering outstanding metrics such as a weighted F1-score of 0.9939 and a ROC-AUC of 0.9980. These results reflect the model’s ability to capture complex, non-linear relationships in the data while maintaining high generalization to unseen cases.

Beyond statistical performance, the study emphasizes practical utility. The models can serve as an intelligent early warning system, enabling biosecurity agencies to identify high-risk outbreaks and respond proactively. With continued data integration and model updates, this predictive system could play a key role in minimizing the impact of avian diseases on public health, food security, and the economy.

In conclusion, this project not only achieves its technical objectives but also delivers real-world value paving the way for smarter, data-driven approaches to disease monitoring and intervention.

## **10. References**

* Avian Dashboard:<https://public.tableau.com/app/profile/siu.apha/viz/AvianDashboard/Overview>
* UK Meteorological Office data:<https://www.metoffice.gov.uk/>
* SMOTE Technique: Chawla et al., 2002
* Gradient Boosting Framework: Friedman, J. H. (2001)
* Scikit-learn Documentation:<https://scikit-learn.org/stable/>