Vector-Borne Disease Prediction

By: Yusuf Jamal, Rajat Vatwani, Rajat Kumar, Ramit Nag
Group number: 18

ML-mid sem Project

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

Why have we chosen this problem?

- Global Health Issue: Vector-borne diseases cause 17% of all infectious diseases and result in 700,000 deaths annually, highlighting a critical need for innovative solutions.

- Early Diagnosis Challenge: Despite treatability, the lack of timely diagnosis remains a key issue due to reliance on costly and inaccessible diagnostic methods.

- Healthcare Limitations: Affected regions, especially underdeveloped areas, often lack the necessary healthcare infrastructure for effective diagnosis and intervention.

- Machine Learning Solution: We aim to leverage machine learning to build predictive models for early detection, addressing diagnostic gaps and improving patient outcomes.

Literature review(1)

Introduction and Background:

Explores the importance of leveraging machine learning to predict diseases early, driven by global health concerns.

Research and Methods:

A dataset consisting of 132 symptoms and 4920 patient records related to 41 diseases was utilized. Data preprocessing was carried out to select 95 symptoms closely associated with these diseases.

Classification Models:

Implemented Decision Tree, Random Forest, and Naive Bayes algorithms to classify diseases based on symptom inputs.

Results:

Achieved a maximum accuracy of 95.12% using the Naive Bayes model, demonstrating its effectiveness in disease prediction.

Literature review(2)

Introduction and Background:

Vector-borne diseases like malaria and dengue are responsible for significant global health issues. Machine learning can help predict outbreaks, improving prevention and control strategies.

Research and Methods:

Using data from 2013-2017 across India, they analyzed environmental and demographic factors. Machine learning algorithms, including CNN, were applied to predict outbreak severity.

Classification Models:

They used CNN with weather and disease data to classify regions into high, moderate, or low-risk areas. The model incorporated environmental and social factors.

Results:

The CNN-based model achieved an 88% prediction accuracy for vector-borne disease outbreaks, demonstrating the power of machine learning in health predictions.

Literature review(3)

Introduction and Background:

Early detection of diseases like breast cancer, diabetes, and heart disease is crucial. Machine learning is increasingly used to improve diagnosis accuracy and survival rates through classification models.

Research and Methods:

Datasets from the UCI repository were cleaned and processed. Five classification models were tested: Logistic Regression, Decision Trees, Random Forest, Support Vector Machine, and Adaptive Boosting.

Classification Models:

Each algorithm, including Logistic Regression, Random Forest, and AdaBoost, was applied to different datasets. Feature selection was performed using backward modeling and p-value tests.

Results:

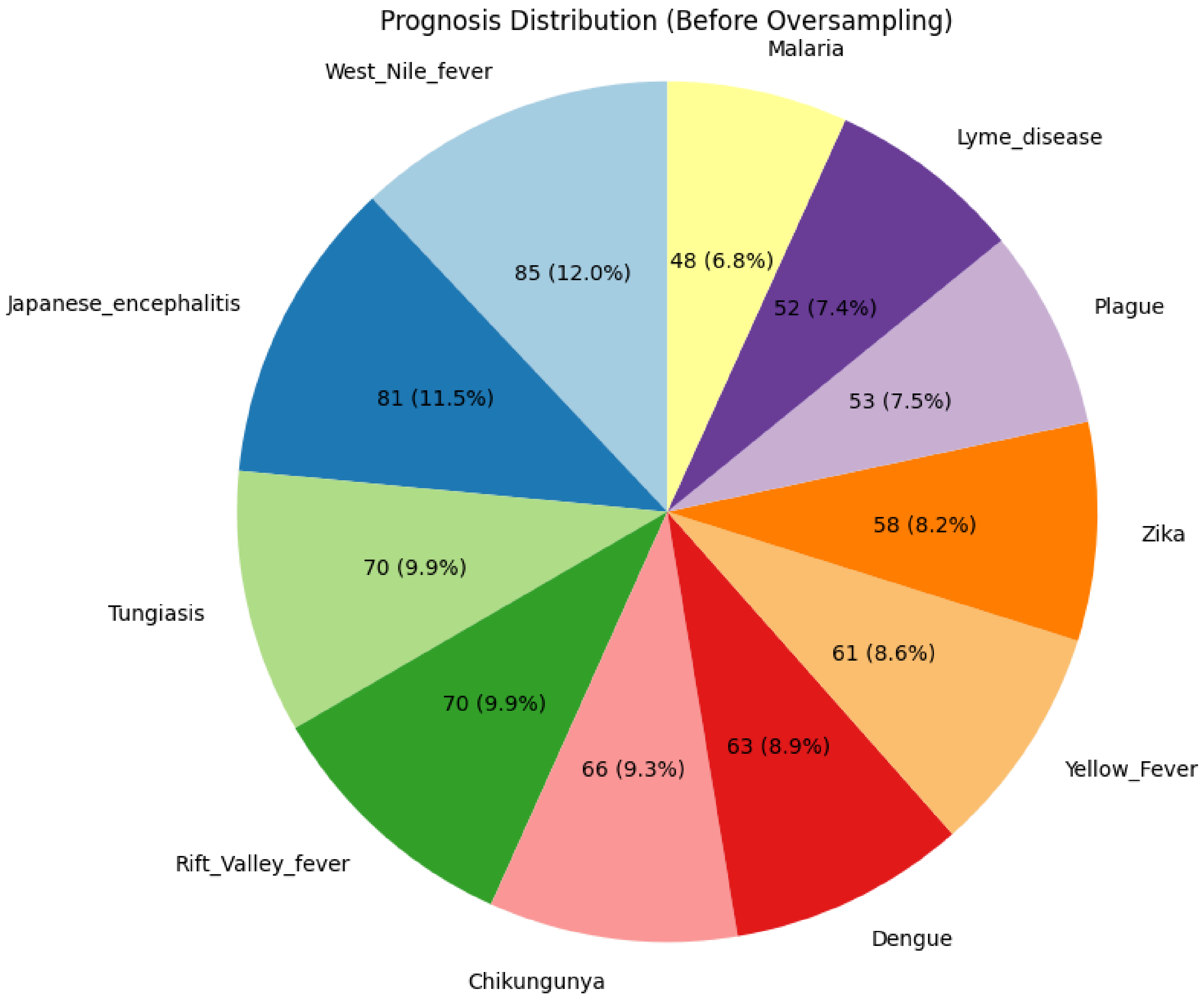
The study showed AdaBoost performed best for breast cancer (98.57% accuracy), SVM excelled in diabetes prediction, and Logistic Regression had the highest accuracy for heart disease (87.1%).

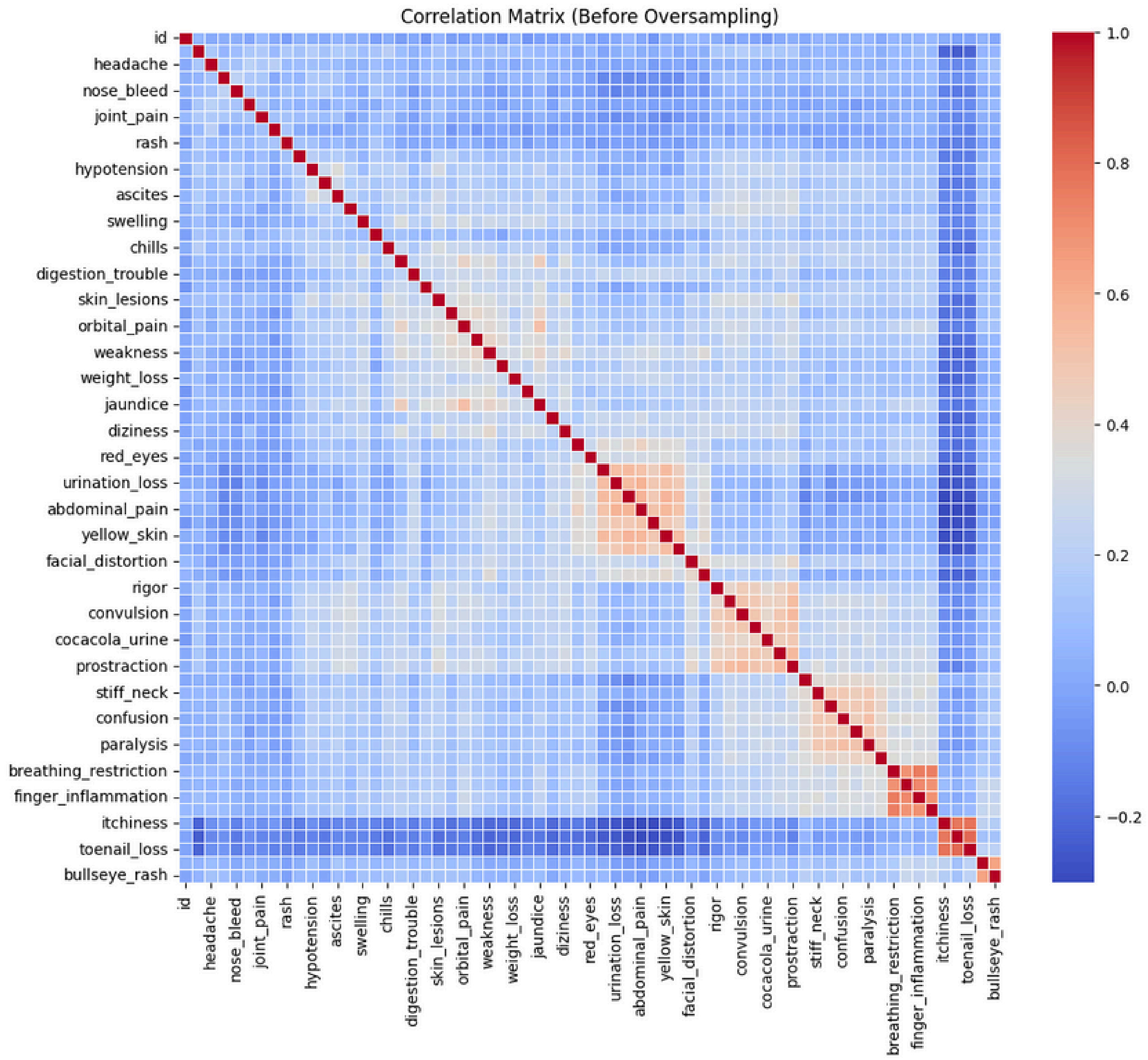
Dataset Description

- Attributes Overview: The dataset consists of 1,010 records with 66 columns. Key attributes include:
- ID: Unique identifier for each record.
- Symptoms: A set of 64 binary features indicating the presence or absence of specific symptoms, such as fever, headache, muscle pain, etc.
- Prognosis: The target variable representing the diagnosis or medical outcome of patients, with 11 distinct categories (e.g., Lyme disease, Zika, Rift Valley fever, etc.).
- Dataset Dimensions: The dataset has 64 features, highlighting the need for dimensionality reduction techniques to manage high dimensionality effectively.

Data Visualization

- A correlation matrix was computed to explore interdependencies between features, revealing collinearity and indicating relationships among symptoms.





Data Preprocessing

- Label Encoding**:** The categorical target variable, "Prognosis," was label-encoded into numerical values to make it compatible with machine learning algorithms.

- Handling Missing Values**:** Missing values in feature columns were imputed using the mean of each column to ensure data completeness.

- Data Splitting**:** The dataset was split into training and testing sets using an 80:20 ratio to enable model evaluation on unseen data.

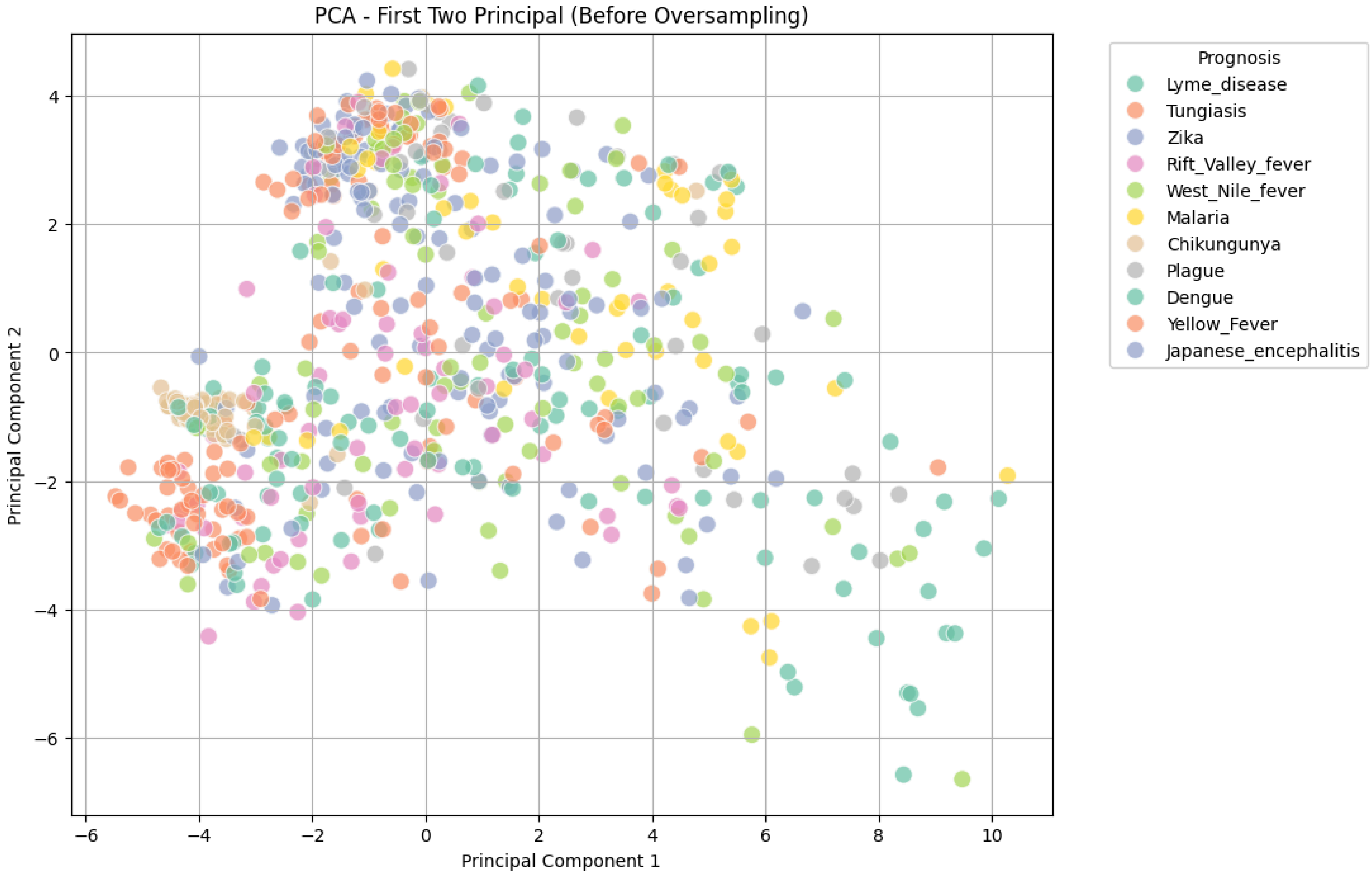
- Dimensionality Reduction**:** PCA was applied to reduce the number of features while retaining 95% of the dataset’s variance.

- Data Balancing with SMOTE**:** Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance by generating synthetic samples for minority classes.

- Data Standardization**:** All features were standardized using Standard-Scaler to place them on a similar scale, which is essential for distance-based models.

Data Preprocessing

- Principal Component Analysis (PCA) was performed for dimensionality reduction, shown in a scatter plot to illustrate the variance captured by key components



Methodology(1)

Naive Bayes:

- A probabilistic model based on Bayes' theorem, assuming all features are independent.
- Performs well in low-variance conditions, showing only a 1% accuracy improvement after bootstrapping.

Decision Tree:

- A tree-based model that recursively splits data based on feature values.
- Overfitting is a common issue, but applying bootstrapping significantly improved accuracy by 8%.

Logistic Regression:

- A linear model for binary classification, estimating class probabilities using the logistic function.
- Supports L1 regularization (Lasso), which selects key features by setting some coefficients to zero, and L2 regularization (Ridge) to reduce overfitting.
- L1 regularization improved accuracy by 2%, with the best performance at C = 1.0 .

Methodology(2)

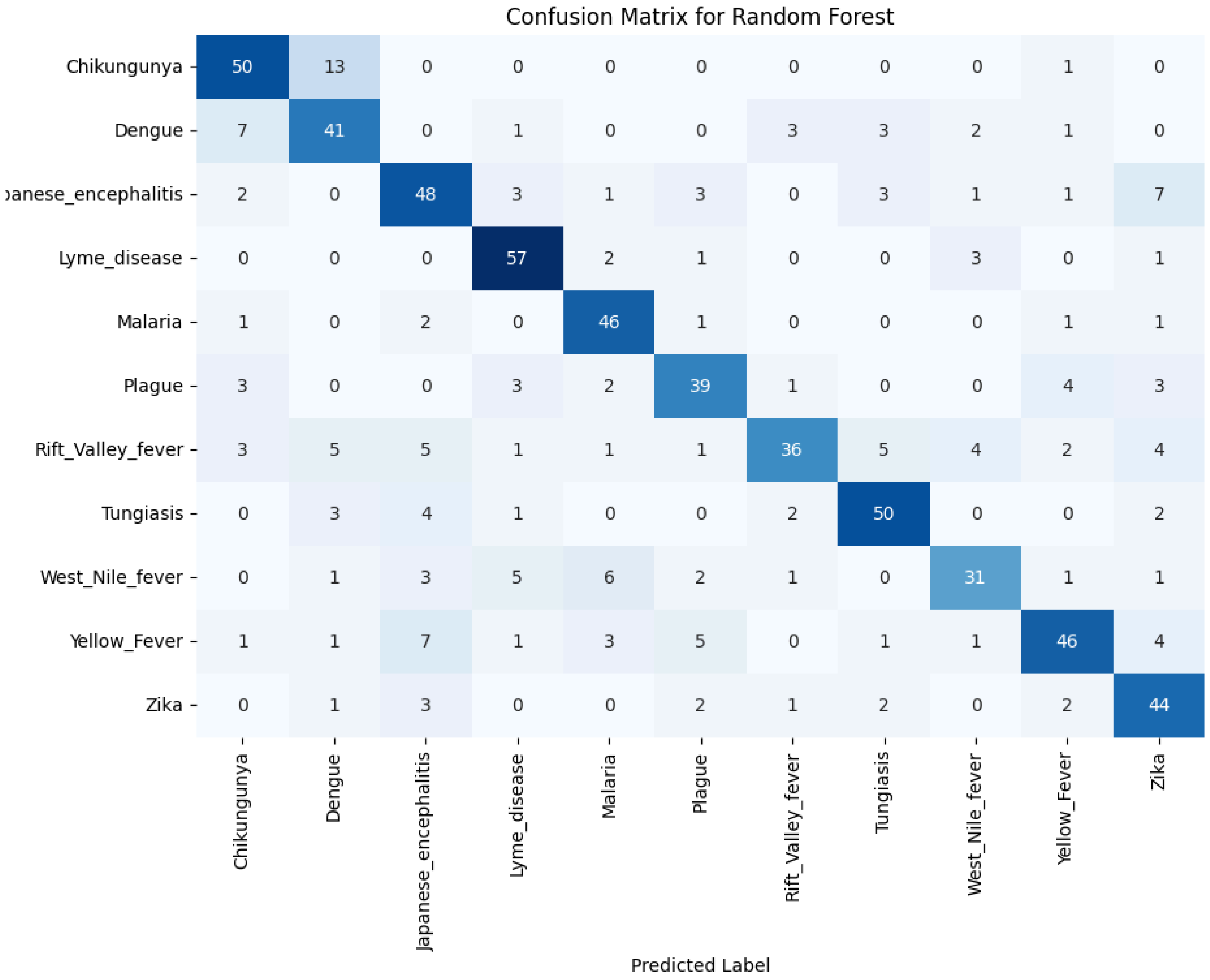
Random Forest:

- An ensemble of multiple decision trees trained on bootstrapped samples of the original dataset.

- Bootstrapping introduces diversity, leading to better generalization and higher accuracy by aggregating predictions from multiple trees.

- The confusion matrix shows the performance of a Random Forest classifier in predicting diseases, with correct predictions along the diagonal and misclassifications off the diagonal.

- High accuracies for diseases like Lyme Disease, Tungiasis, and Zika are evident, while some diseases like Chikungunya and Dengue experience more confusion, indicating room for improvement in distinguishing similar diseases.



Confusion Matrix for Random Forest

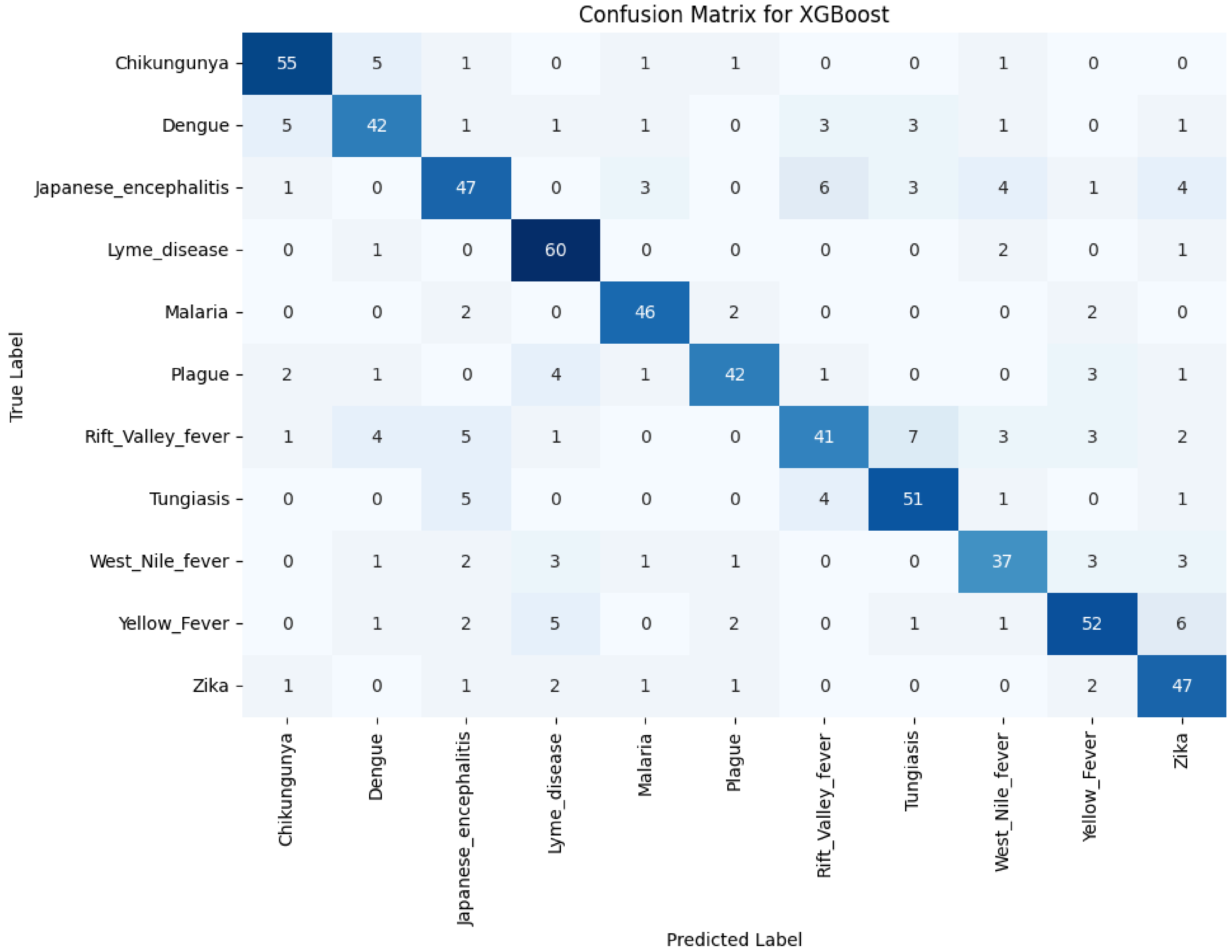
	Chikungunya	Dengue	Japanese_encephalitis	Lyme_disease	Malaria	Plague	Rift_Valley_fever	Tungiasis	West_Nile_fever	Yellow_Fever	Zika
Chikungunya	50	13	0	0	0	0	0	0	0	1	0
Dengue	7	41	0	1	0	0	3	3	2	1	0
anese_encephalitis	2	0	48	3	1	3	0	3	1	1	7
Lyme_disease	0	0	0	57	2	1	0	0	3	0	1
Malaria	1	0	2	0	46	1	0	0	0	1	1
Plague	3	0	0	3	2	39	1	0	0	4	3
Rift_Valley_fever	3	5	5	1	1	1	36	5	4	2	4
Tungiasis	0	3	4	1	0	0	2	50	0	0	2
West_Nile_fever	0	1	3	5	6	2	1	0	31	1	1
Yellow_Fever	1	1	7	1	3	5	0	1	1	46	4
Zika	0	1	3	0	0	2	1	2	0	2	44

Predicted Label

Methodology(3)

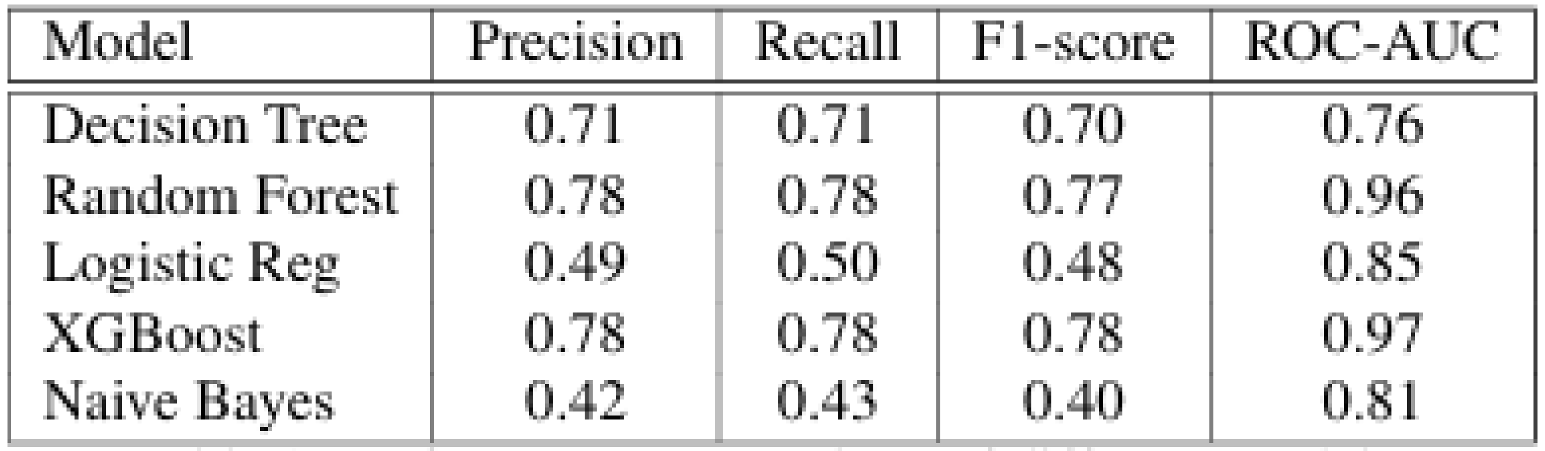
XGBoost:

- A powerful gradient-boosting algorithm that builds an ensemble of decision trees sequentially.
- Each tree corrects errors from previous ones, making XGBoost highly efficient for classification, regression, and ranking tasks.
- The confusion matrix shows how well the XGBoost model predicts various vector-borne diseases, with correct predictions along the diagonal and misclassifications in off-diagonal cells, highlighting the model's strengths (e.g., Lyme disease, Zika) and weaknesses (e.g., Rift Valley Fever).


Confusion Matrix for XGBoost

True Label \ Predicted Label	Chikungunya	Dengue	Japanese_encephalitis	Lyme_disease	Malaria	Plague	Rift_Valley_fever	Tungiasis	West_Nile_fever	Yellow_Fever	Zika
Chikungunya	55	5	1	0	1	1	0	0	1	0	0
Dengue	5	42	1	1	1	0	3	3	1	0	1
Japanese_encephalitis	1	0	47	0	3	0	6	3	4	1	4
Lyme_disease	0	1	0	60	0	0	0	0	2	0	1
Malaria	0	0	2	0	46	2	0	0	0	2	0
Plague	2	1	0	4	1	42	1	0	0	3	1
Rift_Valley_fever	1	4	5	1	0	0	41	7	3	3	2
Tungiasis	0	0	5	0	0	0	4	51	1	0	1
West_Nile_fever	0	1	2	3	1	1	0	0	37	3	3
Yellow_Fever	0	1	2	5	0	2	0	1	1	52	6
Zika	1	0	1	2	1	1	0	0	0	2	47

Results & Analysis(1)

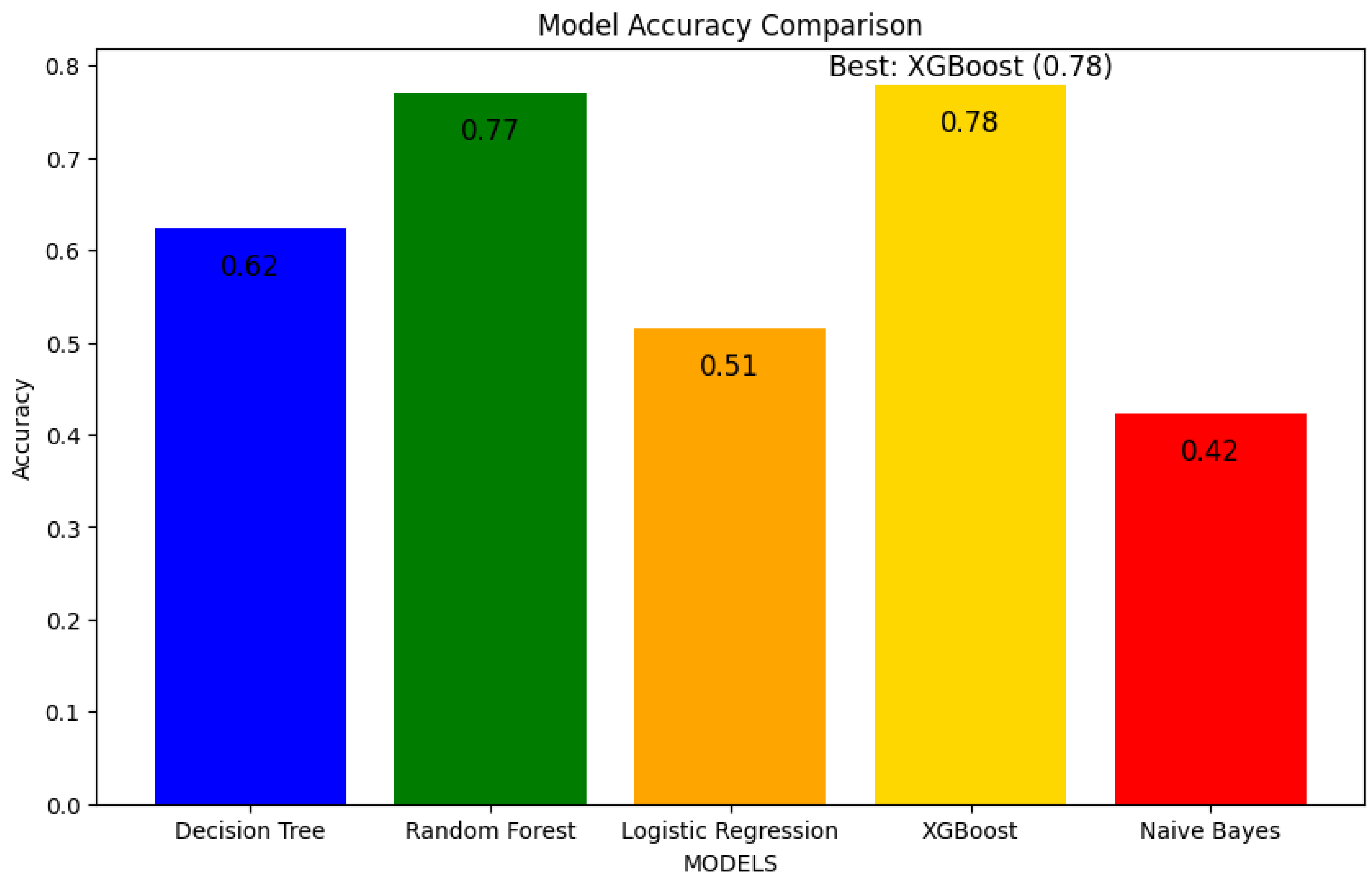


Model	Precision	Recall	F1-score	ROC-AUC
Decision Tree	0.71	0.71	0.70	0.76
Random Forest	0.78	0.78	0.77	0.96
Logistic Reg	0.49	0.50	0.48	0.85
XGBoost	0.78	0.78	0.78	0.97
Naive Bayes	0.42	0.43	0.40	0.81

Table 1. Performance comparison of different models.

- Precision: XGBoost and Random Forest achieved the highest precision (0.78), while Naive Bayes had the lowest (0.42), indicating more false positives.
- Recall: XGBoost and Random Forest also excelled in recall (0.78), capturing the most true positives, whereas Naive Bayes performed poorly (0.43).
- F1-Score: XGBoost showed the best balance with an F1-score of 0.78, while Naive Bayes had the lowest (0.40).
- ROC-AUC: XGBoost (0.97) and Random Forest (0.96) demonstrated strong class separation, while Decision Tree had a lower score (0.76).

Results & Analysis(2)



- XGBoost achieved the highest accuracy at 0.78, making it the top-performing model.
- Random Forest closely followed, performing well but slightly below XGBoost.
- Logistic Regression and Decision Tree showed moderate accuracy, indicating their limited performance on this dataset.
- Naive Bayes had the lowest accuracy among all models, suggesting it struggled with this data.
- The results highlight that ensemble-based methods like XGBoost and Random Forest are more effective compared to simpler models like Naive Bayes and Logistic Regression.

Conclusion

- Best Precision: The Random Forest model achieved the highest precision score (0.78), making it the top choice for minimizing false positives, crucial when avoiding misclassification of non-disease cases.

- Best Recall: Both Decision Tree and XGBoost models had the highest recall scores (0.78), meaning they are effective in correctly identifying true disease cases, essential for early diagnosis.

- Best Overall Performance: XGBoost had the highest ROC-AUC score (0.97), demonstrating the best ability to distinguish between disease and non-disease cases.

Random Forest excels in precision, ideal for reducing false alarms, while Decision Tree and XGBoost strike a balance between precision and recall. XGBoost, with the highest ROC-AUC, provides the most reliable overall performance, making it the best model for accurate vector-borne disease predictions.

Timeline

- Data Pre-processing: 28th Aug 2024 to 3rd Sep 2024
- Feature Selection and Extraction: 5th Sep 2024 to 12th Sep 2024.
- Model Selection and Training: 13th Sep 2024 to 30th Sep 2024.
- Model Testing and Evaluation: 15th Oct 2024 to 22nd Oct 2024.

Future Work:

- Focus on exploring boosting methods like AdaBoost and XGBoost – 4th Nov 2024 to 12th Nov 2024.
- Evaluate the effectiveness of SVM and MLP - 15th Nov 2024 - 26th Nov 2024.

Individual Team members’ contributions

- Yusuf Jamal: Literature review, Data preprocessing and visualization, XGBoost, Report, PPT
- Rajat Kumar: Decision Trees, Random Forest, Naive Bayes, Code, Data preprocessing
- Rajat Vatwani: Logistic Regression, XGBoost, Literature review, Report, PPT
- Ramit Nag: Naive Bayes, Literature review, Logistic Regression, PPT

Thank You