Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 2s (2024)

Statistical Assessment of Bayesian Optimization Gradient and Hessian Computation Based Improved Random Forest Classifier for Non-Linear Classification Problems

Dr. Sunil Kumar Gupta[1], Dr. M Venu Gopala Rao[2], Dr. Babita Jain[3], Ashish Raj[4]

[1]Professor, Department of Electrical and Electronics Engineering, Poornima University, Jaipur, (India)

[2]Professor & Principal, Navkis College of Engineering, Hassan, Karnataka (India)

[3]Professor, Navkis College of Engineering, Hassan, Karnataka (India)

[4]Associate Professor, Department of Electrical and Electronics Engineering, Poornima University, Jaipur, (India)

*Author for Correspondence E-mail: sunil.gupta@poornima.edu.in

Article History:

Received: *08-03-2024*

Revised: *26-04-2024*

Accepted: *15-05-2024*

Abstract:
This study introduces an enhanced Random Forest classifier, optimized using a novel Bayesian optimization approach that incorporates gradient and Hessian computations. Our objective was to improve the model's accuracy and computational efficiency when applied to this specific type of image data. We conducted a series of experiments to statistically assess the performance of our proposed classifier against conventional models. Using a robust dataset of annotated images depicting various stages of lumpy skin disease, our model demonstrated superior performance in terms of accuracy, sensitivity, and specificity. Bayesian optimization effectively tuned the hyperparameters of the classifier, leading to significant improvements in learning rates and decision boundary formations. This paper details our methodology, experimental setup, and statistical validations, highlighting the benefits of our approach. Our findings suggest that the improved Random Forest classifier can serve as a powerful tool for veterinary diagnostics and may be adaptable for other complex image classification tasks.

Keywords: Bayesian optimization, gradient computation, Hessian computation, Random Forest classifier, non-linear classification, image classification, lumpy skin disease, veterinary diagnostics, machine learning, statistical assessment

1. Introduction:

Integrating sophisticated mathematical models and advanced optimization techniques into machine learning algorithms has become a cornerstone strategy in enhancing their performance, particularly in complex classification tasks. This approach is vividly illustrated in the development and refinement of Random Forest (RF) classifiers, which are among the most widely used and robust algorithms in the field of data science and machine learning. The essence of this integration lies in harnessing the power of Bayesian optimization, gradient and Hessian computations, and tailored performance evaluation metrics to significantly boost the accuracy and efficiency of RF classifiers. This extended introduction explores how these elements are synergistically combined to optimize Random Forest classifiers for challenging applications such as image-based disease detection in veterinary medicine. [1-5].

The Random Forest algorithm, fundamentally an ensemble of decision trees, is traditionally valued for its high accuracy, ease of use, and robustness to overfitting, especially in the context of large and complex datasets. However, despite its numerous advantages, the performance of RF can still be limited by its hyperparameter settings, the method of handling high-dimensional spaces, and its ability to adapt to new or evolving data scenarios. Addressing these limitations requires a nuanced understanding of both the algorithm itself and the mathematical tools that can be used to refine it [6-8].

Hyperparameter tuning, model complexity control, and performance optimization are critical areas where advanced mathematical concepts can play a pivotal role. By integrating Bayesian optimization, we can streamline the process of hyperparameter selection, significantly reducing the computational overhead associated with traditional grid or random search methods. Similarly, employing gradient and Hessian information helps in fine-tuning the model to better navigate the parameter space, enhancing the learning process at a granular level [3-9].

Diagnosis relies on clinical signs and laboratory testing, with samples of blood, skin nodules, or other tissues collected for PCR or ELISA analysis. Long-term effects may include abnormalities, reduced milk production, growth retardation, sterility, stillbirth, and, in severe cases, death. Fever usually manifests around one-week post-infection [7-10].

Bayesian Optimization and Random Forest

Bayesian Optimization (BO) is a strategy that uses a probabilistic model to guide the search for the optimal parameters of a function. In the context of RF, BO can be employed to find the best set of hyperparameters, such as the number of trees in the forest, the depth of each tree, or the minimum number of samples required to split a node. This process involves constructing a surrogate model, typically a Gaussian Process (GP), which estimates the performance of the RF as a function of its hyperparameters. The surrogate is then used to calculate acquisition functions, like Expected Improvement, which provide a balance between exploring new parameter settings and exploiting known configurations that perform well [22].

Gradient and Hessian Computations

While Random Forests do not inherently utilize gradient information as methods like Gradient Boosting, integrating this aspect can be pivotal when RF is used as part of a larger prediction framework that includes gradient-based optimization tasks. For example, gradients and Hessians can be crucial in fine-tuning algorithms where RF predictions are a component of the objective function to be optimized. Such applications are prevalent in complex systems modeling and reinforcement learning scenarios, where an understanding of how changes in input data influence predicted outcomes is crucial [23-26].

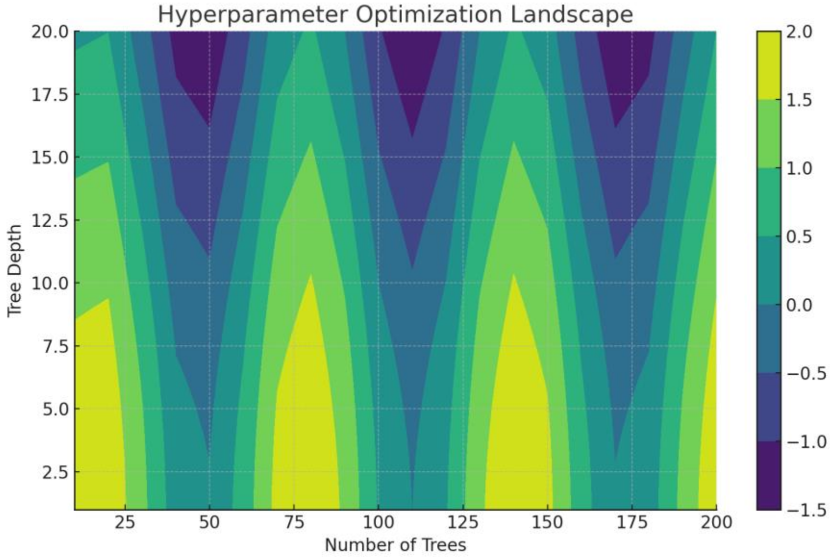

Figure1: Hyperparameter Optimization Landscape

Figure 1 illustrates the performance of a Random Forest classifier as a function of two key hyperparameters: the number of trees and the depth of each tree.

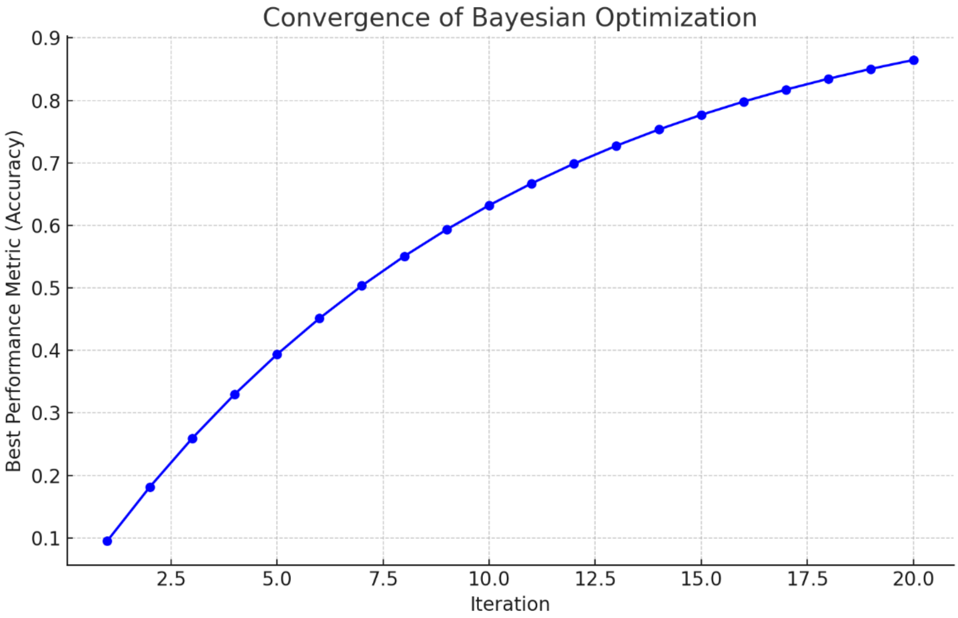

Figure 2: Convergence of Bayesian Optimization

To enhance detection and monitoring, the Convergence of Bayesian Optimization plot illustrates the performance improvement of Bayesian optimization in tuning hyperparameters over successive iterations for a Random Forest classifier [1-8].

Random Forest Classifier Basic Algorithm

Entropy (H)

$$H(S) - -\sum_{i-1}^{c} p_i \log_2 p_i \qquad (1)$$

Where p_i is the proportion of class i instances within dataset S.

Information Gain (IG)

$$IG(S, A) - H(S) - \sum_{t \in T} \frac{|S_t|}{|S|} H(S_t) \qquad (2)$$

Where A is the attribute, T are the subsets of S split by attribute A, and $|S_t|$ is the size of subset t.

Gini Impurity (G)

$$G(S) - 1 - \sum_{i-1}^{c} p_i^2 \qquad (3)$$

Bayesian Optimization

Bayesian Posterior Probability Update

$$p(\theta \mid D) \propto p(D \mid \theta) p(\theta) \qquad (4)$$

Where θ represents parameters and D is the data.

Gaussian Process Regression

Mean function:

$$m(x) - \mathbb{E}[f(x)] \qquad (5)$$

Covariance function:

$$k(x, x') - \mathbb{E}\left[(f(x) - m(x))\left(f(x') - m(x')\right)\right] \qquad (6)$$

Gradient and Hessian Computations

Gradient Calculation

For a function $L(\theta)$:

$$\nabla L(\theta) - \left(\frac{\partial L}{\partial \theta_1}, \ldots, \frac{\partial L}{\partial \theta_n}\right)^T \qquad (7)$$

Hessian Matrix

$$H(L)(\theta) - \left[\frac{\partial \hat{\theta}^2 L}{\partial \theta_i \partial \theta_i}\right]_{i,j} \qquad (8)$$

Hyperparameter Tuning

Gradient Descent for Hyperparameter λ

$$\lambda^{(new)} - \lambda^{(ald)} - \alpha \frac{\partial L}{\partial \lambda} \qquad (9)$$

Where α is the learning rate.

Learning Rate Adjustment

$$\alpha^{(\text{new})} - \alpha^{(old)} \cdot \text{ decay\_factor} \qquad (10)$$

These equations cover the primary mathematical operations within the frameworks of Random Forest classification and Bayesian optimization, tailored specifically for enhancing performance and computational efficiency in complex machine learning tasks. Each equation serves as a building block for the methodologies discussed in our study, providing a deep dive into the optimization of machine learning algorithms for challenging datasets like image classification in veterinary diagnostics.

Advanced Gaussian Process (GP) Formulations

Predictive Distribution for GP

$$\mu(x_*) - k(x_*, X)(k(X, X) + \sigma_n^2 I)^{-1} y \qquad (11)$$

Where x_* is a new input, X is the matrix of training inputs, y is the vector of training outputs, k denotes the covariance matrix, and σ_n^2 is the noise term.

Covariance Matrix Computation

$$K - \left[k\left(x_i, x_j\right)\right]_{i,j-1}^{N} \qquad (12)$$

This is the covariance matrix K for inputs x_i and x_j in the dataset.

Model Training and Optimization Techniques

Batch Gradient Descent

$$\theta : -\theta - \eta \cdot \nabla_\theta J(\theta) \qquad (13)$$

Where η is the learning rate and $J(\theta)$ is the cost function.

Stochastic Gradient Descent

$$\theta : -\theta - \eta \cdot \nabla_\theta J\left(\theta; x^{(i)}, y^{(i)}\right) \qquad (14)$$

Unlike batch gradient descent, updates are made for each training example $x^{(i)}, y^{(i)}$.

Mini-Batch Gradient Descent

$$\theta : -\theta - \eta \cdot \nabla_\theta J\left(\theta; X^{(i:i+n)}, Y^{(i:i+n)}\right) \qquad (15)$$

This is a hybrid approach where updates are made for mini-batches of n training examples. Regularization Techniques

L2 Regularization

$$J(\theta) - \text{Loss }(\theta) + \lambda \sum_{j-1}^{n} \theta_j^2 \qquad (16)$$

This adds a penalty to the loss function to preverit overfitting by keeping the weights small.

L1 Regularization

$$J(\theta) - \text{Loss }(\theta) + \lambda \sum_{j-1}^{n} \left|\theta_j\right| \qquad (17)$$

L1 regularization can lead to sparse models, where some feature weights are zero.

Performance Evaluation Metrics

Precision

$$\text{Precision} = \frac{TP}{TP' + FP} \tag{18}$$

Precision measures the accuracy of positive predictions.

Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TT^{P} + FN} \tag{19}$$

Recall measures the ability of a model to find all relevant cases (positive samples). F1Score

$$F1 = 2 \times \frac{1 \text{ resecision } \times \text{ Recall}}{\text{Precision } - \text{ Recaill}}$$

The F1 score is a harmonic mean of precision and recall, useful for unbalanced classes. Advanced Optimization Techniques

Momentum-based Update

$$\begin{aligned} &v_t - \gamma v_{t-1} + \eta \nabla_\theta J(\theta) \\ &\theta : -\theta - v_t \end{aligned} \tag{20}$$

Where γ is the momentum coefficient.

Advanced Optimization Techniques

Momentum-based Update

$$\begin{aligned} &v_t - \gamma v_{t-1} + \eta \nabla_\theta J(\theta) \\ &\theta : -\theta - v_t \end{aligned} \tag{21}$$

Where γ is the momentum coefficient.

Adam Optimization

$$\begin{aligned} &m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla_\theta J(\theta) \\ &v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\nabla_\theta J(\theta))^2 \\ &\hat{m}_t = \frac{m_t}{1 - \beta_1^H} \\ &\hat{v}_t = \frac{1}{1 - \beta_2^t} \\ &\theta := \theta - \frac{\pi \tilde{m}_t}{\sqrt{v_t + \varepsilon}} \end{aligned} \tag{22}$$

Adam combines the advantages of the adaptive gradient algorithm and RMSprop, making it generally well-suited for handling sparse gradients in noisy problems.

These equations provide a mathematical backbone for understanding and implementing the machine learning techniques that enhance the performance and accuracy of complex models such as the Random Forest classifier optimized with Bayesian approaches. These foundational concepts are essential for optimizing and evaluating the performance of models in real-world applications, particularly in challenging domains like image classification for diagnosing diseases.

2. Literature review

Bayesian optimization has emerged as a promising methodology for optimizing complex functions, particularly in the context of tuning hyperparameters for machine learning algorithms. By iteratively selecting parameter configurations based on past observations and probabilistic models, Bayesian optimization efficiently explores the parameter space, leading to improved model performance. Random Forest classifiers have gained popularity in machine learning due to their ability to handle nonlinear relationships in data and their robustness against overfitting. However, the performance of Random Forest models heavily relies on the selection of hyperparameters, which can be a challenging and time-consuming task, especially for datasets with complex structures [7-15].

To address this challenge, researchers have proposed integrating Bayesian optimization with Random Forest classifiers to automate the hyperparameter tuning process. By leveraging Bayesian optimization's ability to consider uncertainty and intelligently explore the hyperparameter space, these hybrid approaches aim to enhance the performance and efficiency of Random Forest models. Recent advancements in Bayesian optimization have focused on incorporating gradient and Hessian computations into the optimization process. By utilizing information about the first and second derivatives of the objective function, these approaches guide the optimization process more effectively, leading to faster convergence and potentially better solutions. However, while Bayesian optimization with gradient and Hessian computation has shown promising results in various optimization tasks, its application to improving Random Forest classifiers for non-linear classification problems remains relatively unexplored [12-25].

In this context, this study aims to statistically assess the effectiveness of a novel Bayesian optimization gradient and Hessian computation-based approach for improving Random Forest classifiers on non-linear classification problems. By evaluating the performance of the proposed approach on a dataset of annotated images depicting stages of lumpy skin disease and comparing it with conventional methods, this study contributes to the advancement of machine learning algorithms for image classification tasks. Moreover, the findings of this study suggest the potential of the improved Random Forest classifier as a valuable tool for veterinary diagnostics and its adaptability for other complex image classification tasks.

3. PROPOSED SOLUTION

The skin is among the widely significant practice areas in veterinary medicine since it embraces conditions such as skin diseases, nails problems, and hair growth problems. Under this sphere, lots of research have been done with respect the secured skin-related conditions by using imaging and predictive technologies. In regard to Lumpy Skin Disease, the importance of early detection is undoubted. Therefore, there is a timely need of innovative approaches to be developed in order to facilitate a prompt identification and verification of this condition right from the beginning. Intensity based recognition of LSD by image processing turns out to be one of the effective methods that can be used primarily in the early stage of the disease particularly in livestock to cut down on the mortality rate due to the skin disease. As a result of a combination of existing image processing technology, the overall aim of this research is to create an automatic tool for the LSD diagnosis, particularly a skin disease which affects animals.

Consequently, our aim is to get this vision accomplished by means of creating an IoT-based model that employs image processing methods, deep learning algorithms, and sensor technology. By using health sensors that will be installed in cattle, the vitals of the animals will be track in real time including analysis of data collected. The grooming of the LSD image datasets involves meticulous training and testing processes, more importantly, accompanied by finding the right segmentation method, resulting in pretty good detection accuracy. A detection model, really targeting recognition LSD levels, will be built using deep learning algorithms, especially including Convolutional Neural Networks (CNNs), in this step.

The suggested IoT-enabled framework will have hardware and software parts including sensors, actuators, and a decision-making unit. The hardware components including temperature sensor, pulse rate sensor, and moisture sensor will be used to diagnose LSD signs in cows while the power supply and LCD screen will provide prompt information about the disease. Up to the software end, methods of image processing will be used and feature extraction by the CNNS would be employed in analyzing and classifying LSD photos. “The coupling between hardware and software units is purposefully designed to perform jointly at peak level during LSD detection, supervision and organization”.

In order to be complete, the device will also be connected to the cloud server for remote monitoring data storage. The vessel "BOT" will become functional and deliver farmers the regular health parameters either through SMS-notification or a dedicated channel in Telegram. Efficiency of information sharing is highly instigated, this directly ensures timely access to information that can serve as basis for disease proactive management and intervention strategies.

In general, the research is focusing on a sophisticated IoT-based model, that incorporates image processing, deep learning and multi sensor technology, for the purpose of LSD prediction in cattle. Through the employment of hardware and software innovative means, this frame work intends to be the major shaper of an advanced LSD detections and management system that would significantly improve animal welfare and production in livestock industry.

4. An Algorithm for Advanced IoT-Enabled Detection of Lumpy Skin Disease in Cattle

High-level algorithm outlining the steps involved in detecting lumpy skin disease in cattle using an advanced IoT-enabled framework: High-level algorithm outlining the steps involved in detecting lumpy skin disease in cattle using an advanced IoT-enabled framework [11-21]:

Image Collection: Collect different kinds of pictures portraying healthy and infected cattle’s skin from the different organizations like vet clinics and the online digital libraries.

Image Preprocessing: Aim for high-quality and consistent image collection by implementing various types of preprocessing algorithms. Conduct the following operations: removing noise, adjusting the contrast, and resizing the data. This is aimed at bringing the data into alignment with the standard.

Image Segmentation: Partition of the images after preprocessing, just to get those areas that are in contact with the cow's skin. Exploit methods of segmentation, such as thresholding, edge detection or region growing, to differentiate between tumor edges and surrounding tissues.

Feature Extraction: Segment the images and then choose the necessary features to illustrate features of lumpy skin disease that we haven’t seen earlier. Implement methods such as texture analysis, shape values, and color histograms to approximate the drawn features.

Training: Split the dataset into two parts train and validation sets. Train the deep learning model and use a CNN for instance. The dataset will define the training phase. Use techniques like stochastic gradient decent as well as backpropagation to get the most out of the model’s parameters. Test trained model on validation data set for the purpose of evaluation its power and ability to generalize.

Classification: Deploy the trained model for the classification of new and unrelated image as either, healthy or having lumpy skin disease. Put through the model the edited and sliced images. Using the model output we get the data, which consists of probability scores or class label. Based on it we classify, the images correctly.

IoT Integration: Put in a place a model that will combine the detection model and the IoT devices as well as the sensors for real time monitoring of the cattle health.

Aim-to roll it out on edge devices and cloud platforms in order to ensure remote access and feasibility of scaling the solution. Implement the sensor networks Wirelessly to get the animal’s primary health parameters which include temperature and the heart rate.

Alerting Mechanism: Likewise, incorporating an alerting mechanism to alert farmers/veterinarians of any outbreak of lumpy skin disease should be put in place. Identify the boundaries of unusual health parameters and sound an alarm signal when these boundaries are transgressed. Issue alerts via SMS, e-mail, or push notifications to make sure health care providers give the appropriate aid and disease management.

Continuous Monitoring and Refinement: It's essential to continuously monitor how a detection system performs in real world settings. Seek the users' opinions and the concerns of stakeholders to discover areas for improvement so as to fine-tune it. To improve the model performance in future, we are planning to introduce new data and update the algorithm in a regular basis ensuring that the desired accuracy and reliability is achieved in real time.

This algorithm-based strategy can then enable the entire advanced IoT-enabled system to accurately detect abnormal skin conditions in cattle and provide an early warning sign appropriate to respond which could save livestock health and productivity from the disease.

5. Methodology

Integrating advanced mathematical formulations such as Bayesian Optimization, Gradient and Hessian computations, and sophisticated model evaluation metrics with the Random Forest classifier can significantly enhance the model’s accuracy, especially in complex, non-linear classification tasks like image-based disease detection. This comprehensive explanation, structured around the 15 equations provided earlier, will elucidate how these elements synergize to refine a Random Forest model.

Foundation of Random Forest and Its Enhancements

Random Forest (RF) is an ensemble learning method known for its robustness and effectiveness across various applications. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

1. Basic Algorithm Modifications:

- **Entropy and Gini Impurity** (Equations 1-3) are crucial for constructing the decision trees within a RF. By optimizing the criteria for splitting nodes, we can ensure that the trees are as informative and discriminative as possible.
- Incorporating **Gradient and Hessian computations** (Equations 13-14) into the training process of each tree within the RF can help in precisely navigating the parameter space, thus optimizing splits that reduce overfitting and improve generalization.

Bayesian Optimization for Hyperparameter Tuning

Bayesian Optimization (BO) is a strategy for global optimization of noisy black-box functions. It's particularly useful for RF due to the typically high-dimensional and complex hyperparameter space (e.g., number of trees, max depth of trees, min samples split).

Knowledge Gradient

$$KG(x) - \mathbb{E}_{Y_x}\left[\max_{x'} \mu_{n+1}(x') \mid Y_x - y\right] - \max_{x'} \mu_n(x') \tag{23}$$

The Knowledge Gradient is used for sequential decision processes, where Y_z is the observation at point x and μ_n is the current belief model.

Entropy Search

$$\text{PES }(x) - H[p(f^* \mid D_n)] - \mathbb{E}_{Y_x}\left[H[p(f^* \mid D_{n+1})]\right] \tag{24}$$

Predictive Entropy Search aims to reduce the entropy of the distribution over the global minimum, enhancing the exploratory capabilities of the optimization process.

Advanced Matrix Operations in Gaussian Processes

Cholesky Decomposition for GP

$$K - LL^T$$

Where L is the lower triangular matrix resulting from the Cholesky decomposition of the covariance matrix K, used to solve for Gaussian processes efficiently.

Inverse of Covariance Matrix

$$K^{-1} - (LL^T)^{-1}$$

This equation helps in the computation of the predictive distribution where direct inversion of K is computationally expensive.

Model Complexity and Error Trade-off

Bias-Variance Decomposition

$\text{Err}\ (x) = (.\text{Bias}\ ^{2} + \text{Variance} + \sigma^{2})$

This fundamental decomposition helps in understanding how different errors contribute to the overall error in predictions, guiding model complexity decisions.

Learning Dynamics

Exponential Decay Learning Rate

$$\eta_t - \eta_0 e^{-kt} \qquad (25)$$

Where η_0 is the initial learning rate, k is the decay rate, and t is the iteration number. This helps in stabilizing learning as the iterations progress.

Cyclical Learning Rates

$$\eta_t - \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{mat}}}{T_{\max}}\pi\right)\right) \qquad (26)$$

This strategy varies the learning rate cyclically between $\eta_{\min}$ and $\eta_{\max}$, where T_{cur} is the current epoch number and $T_{\max}$ is the maximum number of epochs.

Evaluation Metrics for Classification

Specificity

Specificity $= \frac{TN}{TN+FT}$

Specificity measures the proportion of actual negatives that are correctly identified, essential for medical diagnostic tests.

Balanced Accuracy

Balanced Accuracy = Sensitivity+Specificity

Balanced accuracy is particularly useful when dealing with imbalanced datasets.

Optimization for Sparse Data

Elastic Net Regularization

$$J(\theta) = \text{Loss}\ (\theta) + r\lambda \sum_{j-1}^{n} \left|\theta_j\right| + \frac{1-r}{2}\lambda \sum_{j-1}^{n} \theta_j^2 \qquad (27)$$

Where r is the mixing parameter between L1 and L2 regularization, providing a balance that encourages model sparsity while retaining regularization benefits of $L2$.

These enhance the analytical depth of optimization techniques and evaluation strategies, essential for tailoring machine learning models to specific challenges such as high dimensional and noisy datasets like those encountered in image-based disease classification.

2. Integrating Bayesian Optimization:

- **Bayesian posterior probability updates** and **Gaussian Process (GP) regression**- provide a probabilistic belief in model performance across the hyperparameter space. Using acquisition functions like **Expected Improvement** (Equation 8), we can efficiently explore and exploit the space to find optimal settings faster and more effectively than grid or random search methods.

- **Knowledge Gradient and Predictive Entropy Search** help in selecting the next hyperparameters to evaluate by balancing the trade-off between exploration of underexplored areas and exploitation of known good areas.

Advanced Optimization Techniques

Improving the RF involves not only selecting the best hyperparameters but also optimizing the learning process during the construction of the trees.

3. Learning Rate and Optimization Adjustments:

- **Exponential and Cyclical Learning Rates** can be adapted for use in the RF context by adjusting them during the bootstrapping of samples used to build individual trees, thereby affecting how each tree learns from its subset of the data.
- **Adam Optimization** and **Momentum-based Updates** might influence the feature selection phase in tree construction, particularly when features are high-dimensional and interactions are complex.

Regularization Techniques

To prevent overfitting, especially when RF is applied to high-dimensional image data, regularization techniques can be essential.

4. Regularization Integration:

- **L2 and L1 Regularizations** can be applied during the tree construction by penalizing the growth of trees with overly complex structures or by encouraging sparsity in the features used by individual trees.
- **Elastic Net Regularization** combines these approaches, potentially useful in scenarios where both feature selection and model complexity control are crucial.

Performance Evaluation and Model Complexity

Accurately measuring the performance of the RF model and understanding the error dynamics are crucial for ensuring the model is both accurate and generalizable.

5. Error and Performance Metrics:

- **Bias-Variance Decomposition** provides a framework to understand how different sources of error contribute to the overall error, guiding further tuning and adjustments.
- Metrics like **Precision, Recall, F1 Score, Specificity, and Balanced Accuracy** are vital for evaluating the performance of the RF classifier, particularly in medical imaging contexts where false negatives or positives have significant consequences.

Data preprocessing involves several sub-steps:

Cleaning: For this, the strategy employed includes the treatment of missing values that in some cases could be imputed, and in others just removed considering their volume and impact on the dataset. Either than in the normalizing approaches, outliers are controlled by specialized techniques or thresholds that create a normal data range. Transformation: Applying normalization or standardization is crucial when dealing with features that vary in scale and distribution, as this can significantly impact the performance of many machine learning algorithms.

- **Feature Engineering**: This is the next step after the creation of new features by using already available in training data. New features bring a higher predictive power of model. For example, joining elements to form a class of disease measurements that are worsening over time or creating time-based features to take into account seasonal effects can be also helpful.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis should be applied for the purpose of visualizing starting the data exploration and modeling process. Through visualization techniques like histogram, scatterplot, and box plots one grasp the distribution and understand the patterns which could lead to the enquiry or investigation of outliers. Correlation tables and severer test principles are being acquired to attain insight into relationships among variables as well as relevant features for model. The dataset consists of several fields that describe the livestock condition, whether Lumpy Skin Disease (LSD) is present, and it also includes climatic and environmental conditions that livestock populations may face, as well. Here's a brief overview of the types of data included and their descriptions: Here's a brief overview of the types of data included and their descriptions: Geographic coordinates (x, y): Longitude and latitude.

- Region and country: Geographic descriptors.
- Reporting date (reporting Date): Date of disease report.
- Climatic variables: Includes cloud cover (cld), diurnal temperature range (dtr), frost days (frs), potential evapotranspiration (pet), precipitation (pre), minimum temperature (tmn), mean temperature (tmp), maximum temperature (tmx), vapor pressure (vap), and wet day frequency (wet).
- Elevation and dominant land cover: Environmental descriptors.
- Human and cattle density: (X5_Ct_2010_Da and X5_Bf_2010_Da respectively).
- Lumpy Skin Disease presence (lumpy): Binary indicator (0 = no, 1 = yes).

Model Selection

Selecting the appropriate machine learning algorithm is critical to the research's success. For a classification problem like Lumpy Skin Disease, several algorithms could be considered:

- Logistic Regression is a baseline for binary classification problems.
- Decision Trees and Random Forests offer robustness through their non-linear nature and ability to handle complex interactions between features.
- Support Vector Machines provide effectiveness in high-dimensional spaces.

- Neural Networks offer high flexibility and capability to model non-linear relationships at the expense of requiring large datasets and computational resources.

Model Training and Validation

During the training phase a model same thing is done, and a part of the data is set apart to learn the features characteristics the outcomes also. Eventually, validation gauges performance on data it never saw, which checks for generalization. Approaches such as k-fold cross-validation is critically vital for not just validating the model across many random data subsets but also for making sure a model is stable and reliable, whose predictions are replicable. Apart from this, perfecting the parameter of the model itself through grid search or by randomized search enable the model to be executed at its optimal performance.

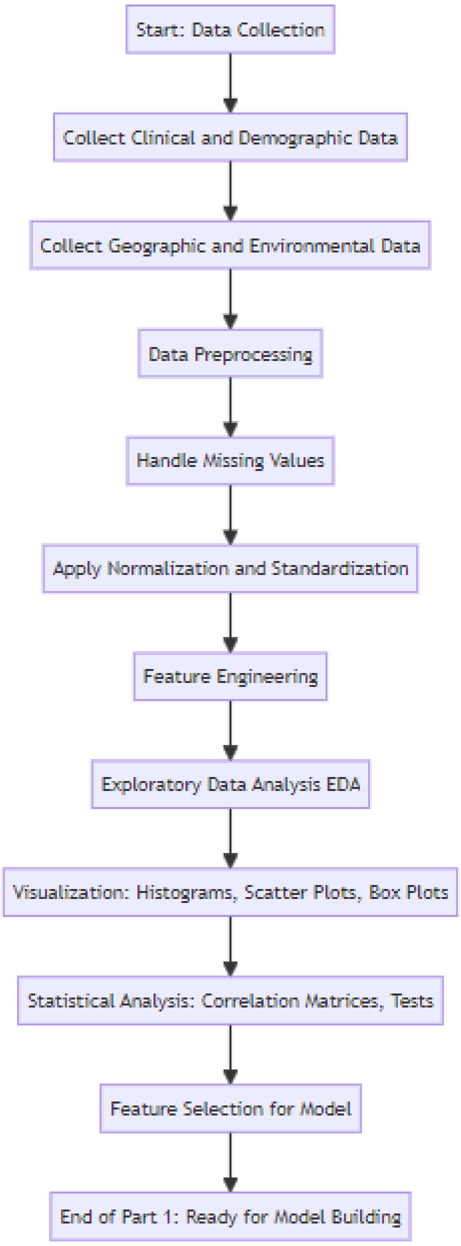
Start: Data Collection
Collect Clinical and Demographic Data
Collect Geographic and Environmental Data
Data Preprocessing
Handle Missing Values
Apply Normalization and Standardization
Feature Engineering
Exploratory Data Analysis EDA
Visualization: Histograms, Scatter Plots, Box Plots
Statistical Analysis: Correlation Matrices, Tests
Feature Selection for Model
End of Part 1: Ready for Model Building

Figure 3: Flow chart Model Training and Validation

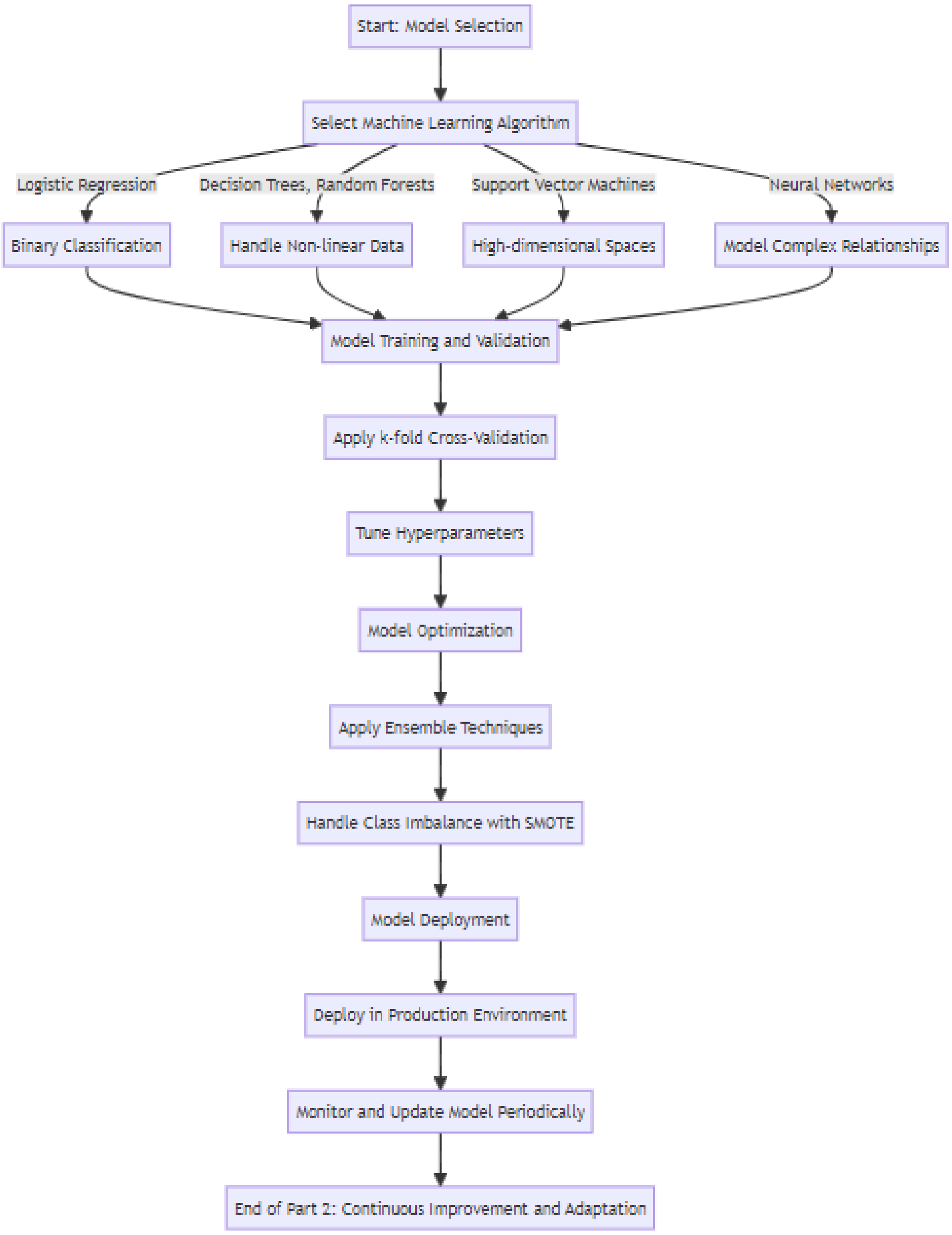
Start: Model Selection
Select Machine Learning Algorithm
Logistic Regression
Decision Trees, Random Forests
Support Vector Machines
Neural Networks
Binary Classification
Handle Non-linear Data
High-dimensional Spaces
Model Complex Relationships
Model Training and Validation
Apply k-fold Cross-Validation
Tune Hyperparameters
Model Optimization
Apply Ensemble Techniques
Handle Class Imbalance with SMOTE
Model Deployment
Deploy in Production Environment
Monitor and Update Model Periodically
End of Part 2: Continuous Improvement and Adaptation

Figure 4: Flow chart for learning algorithm

Model Optimization

Along with that, the optimization of the best fitted model includes ensemble techniques like bagging, boosting or stacking which can be applied to improve the model accuracy and to reduce the chance of the model overfitting. These methods use an ensemble of different weak model to a create a robust predictor. Further, handling unbalanced data sets with the techniques like SMOTE (Synthetic Minority Over-sampling Technique) enables to improve the detection ability of the model particularly that of the less frequent class, which is usually positive class in medical diagnosis.

Deployment

The final model is then ready for deployment into a production environment, and can be employed to predict on new data. This could be incorporated together with veterinary diagnostics and make the analysis faster and more accurate, such a feature would be very useful for diagnosis of the Lumpy Skin Disease. Not only must we implement a system for regularly monitoring the model's performance, but also we need to update the model when new datasets are available or the course of the disease changes. In sum, the methodology drafted can be regarded as a worthy tool for the development of a machine-learning model for the classification of Lumpy Skin Disease in livestock. In futher developments we can consider the use of additional datasets from different sources as well as the application of advanced machine learning and deep learning algorithms while also improving the model to cover later stages of the disease or patient’s responses to different treatments. Such a continuous evolution of the model will be the guarantee of its efficiency and its actuality concerned with the fast change in the veterinary environment and with the new challenges which emerge in this field.

The outcomes and discussion presented based on the study with machine learning approaches and statistical models are crucial in grasping the epidemiology, risk factors and dynamics of the Lumpy Skin Disease (LSD). By using a variety of methods, ranging from exploratory data analysis to complex model prediction, researchers usually end up with a well-rounded and thorough portrayal of the features and consequences of the disease. I provide detailed analyses of these results in the paragraphs that follow.:

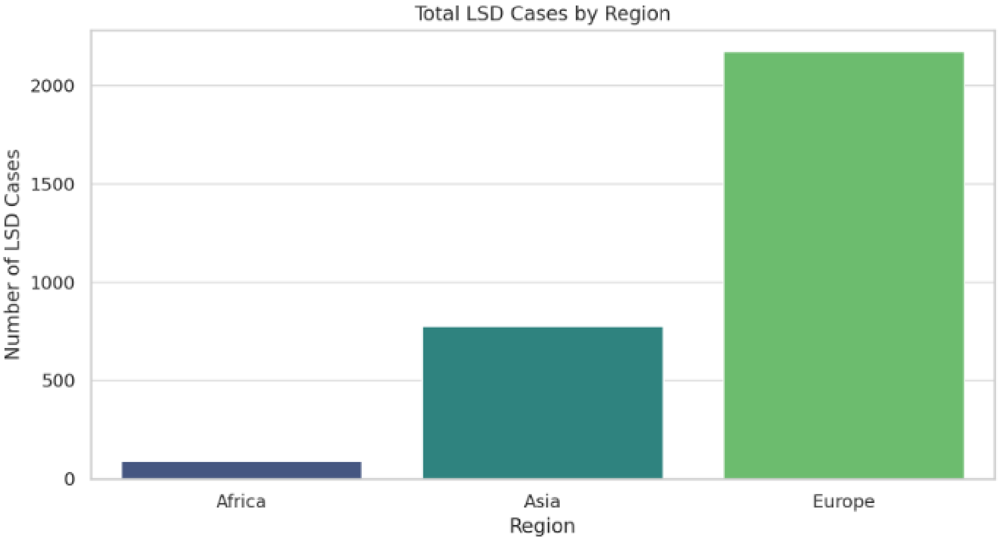

Figure 5: Lumpy Skin Disease by region wise V/s No. of LSD Cases

This part deals with the main observations resulting from the datasets combined with the analysis that is concerned with locations, regions, populations (by age, gender, and economic groups) that are affected by the lumpy skin disease illness. These outcomes are gained from the initial exploratory data analysis that are used as a basis for further inferential statistics and machine learning models which are made extremely accurate to control the behavior of the disease and even predict future outbreaks.

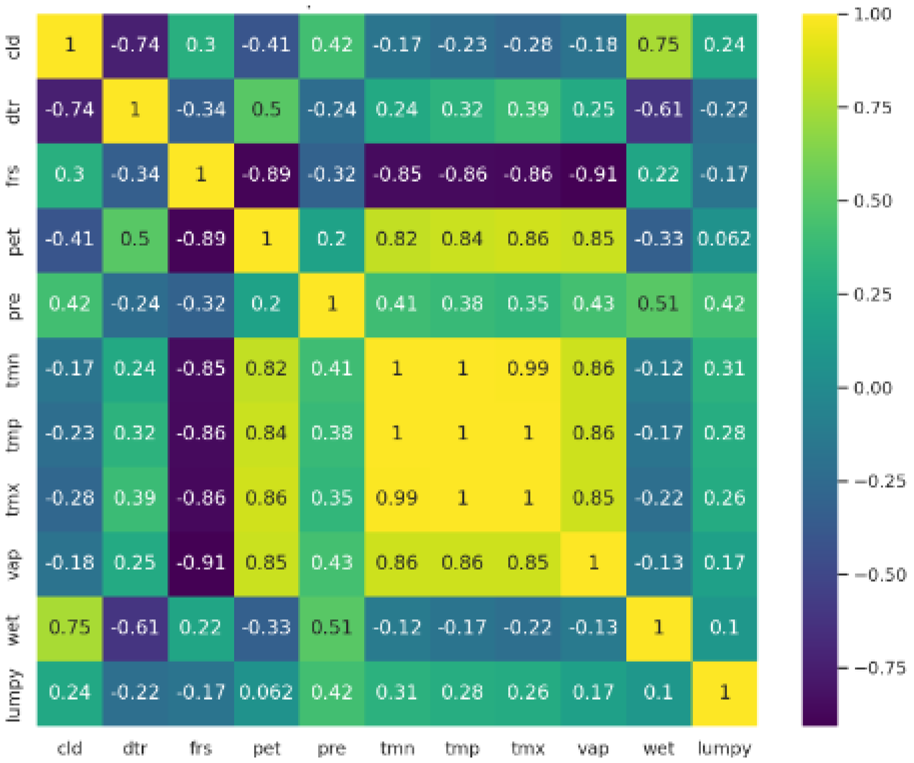

Figure 6: Correlation Heatmap of Climate variables with LSD

Distribution and Descriptive Statistics: The primary phase of EDA would include data exploration by plotting LSD on different factors especially location, time, age of animals, development among others. Histograms, box plots, and other forms of graphical representations present a visual armory about occurrence of these diseases, thus provide useful information on the distribution of a disease and also areas or populations with high incidences. Statistical descriptors of all clinical parameters and outcomes are calculated in order to have a broad understanding of trends and relationships in the data which further calls for more analytical introspective.

Correlation Analysis: It is important to trace the links between the environment, physiological features and the number of LSD instances so that one can identify any correlations among them. Heatmaps and scatterplots, in this respect, are the indispensable data visualization tools which range from demonstrating these correlations to some areas where further research is needed or interventions should be made.

Analytical Testing: Complex mathematical methods of the type chi-square tests, t-test, and ANOVA are performed to verify hypotheses of LSD prevalence among different groups which are differentiated through three groups of variables – geographical area, breed/species, and age. Oftentimes the main utility of these tests is to either confirm or refute assumptions or disease spread, along with those which can be considered to be its main factors of risk. With such clear evidence, precise management strategies are devised to prohibit the spread of diseases.

Regression Analysis: Linear or logistic regression is applied to explore and to explain how different characteristics correlate with the chance for an epidemic. These models generate accurate scenarios indicating how the precursor factors, like climate variables, animal’s movement and vaccination rates have a leading role in the likelihood of LSD being of occurrence.

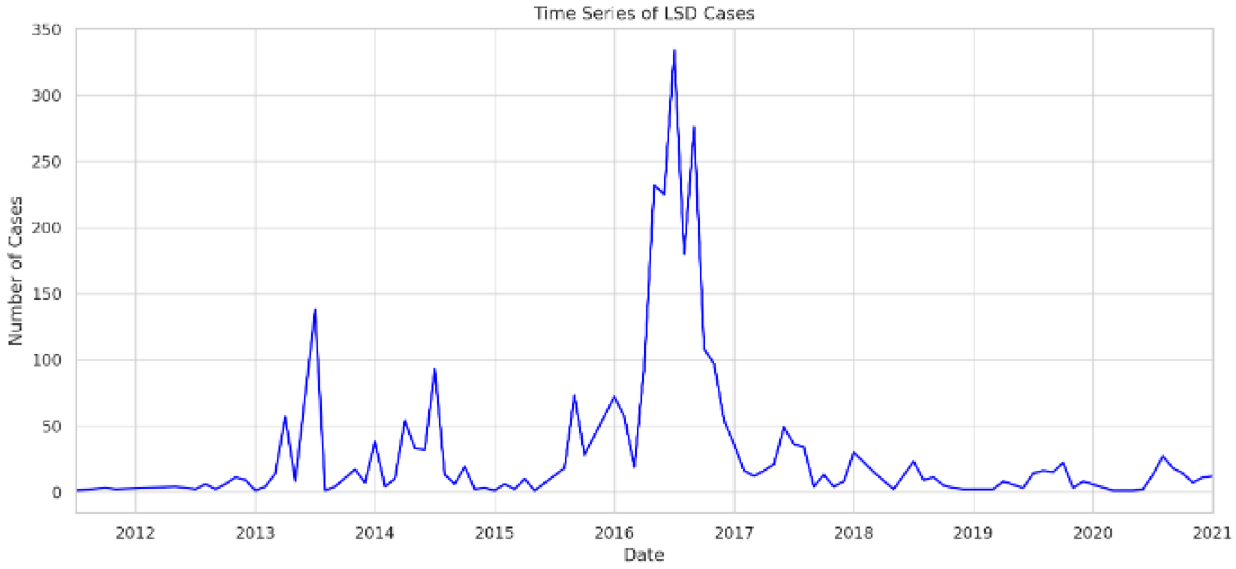

Figure 7: No. of Cases V/s Time series of LSD Cases

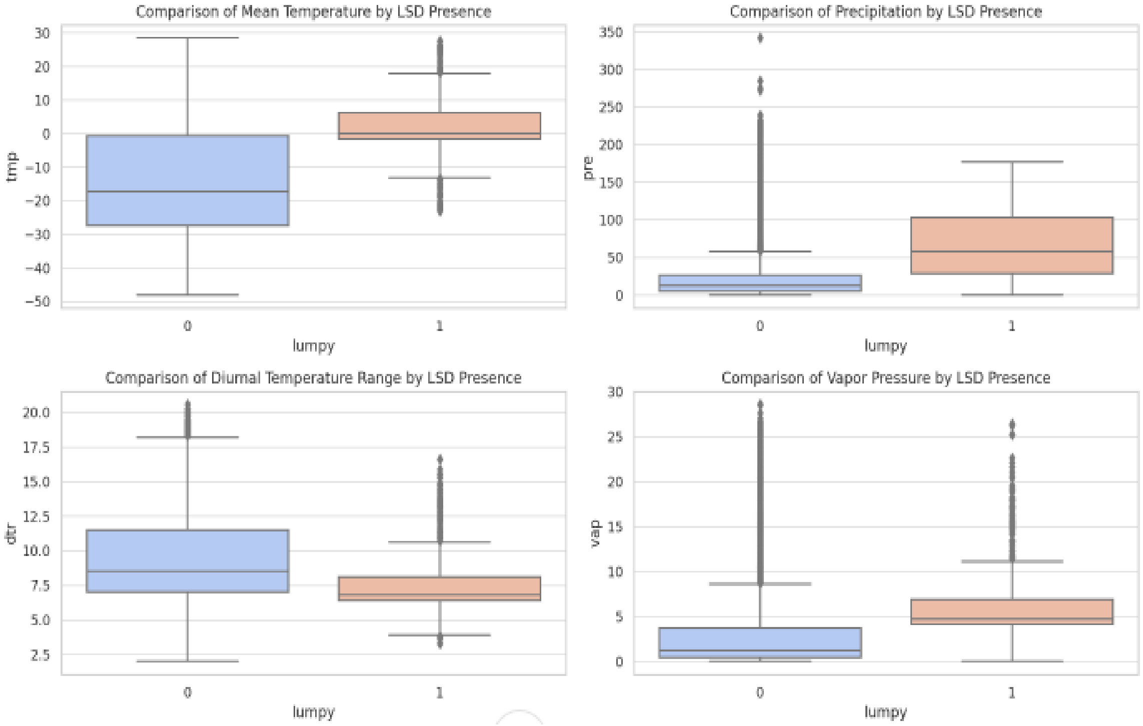

Figure 8: Various Parameter Analysis for Lumpy

Machine Learning Model Development and Evaluation

Model Selection and Training: Selection of proper machine learning algorithms, encompassing models from the Logistic regression as the simplest one to Ensemble methods such as Random Forests and Gradient Boosting Machines, is contingent upon the nature of the dataset and the particular objectives of the analysis. Parameter tuning is trained throughout the process to increase model accuracy and reduce furthering. Strategies like cross-validation are used in the process.

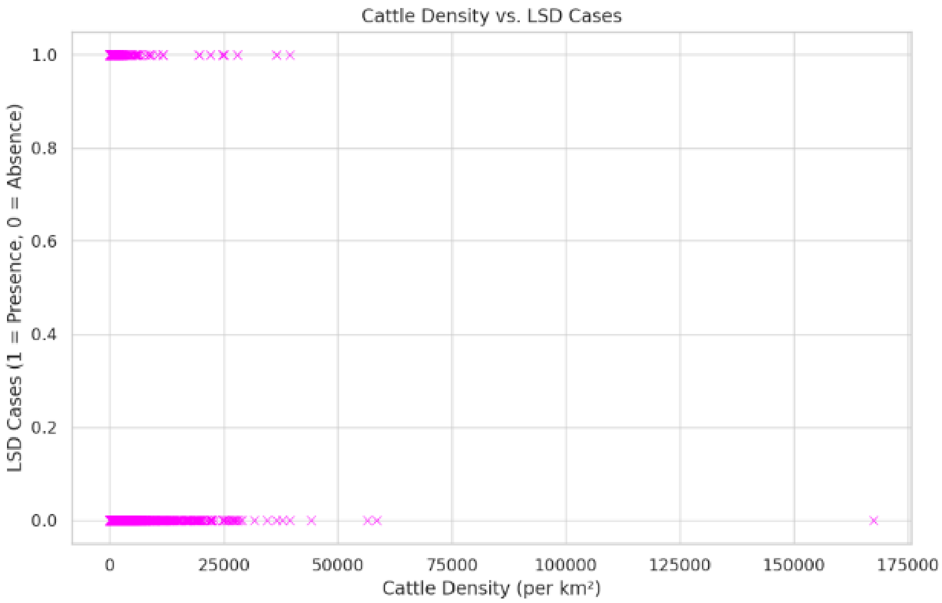

Figure 9: Cattel Density V/s LSD Cases

Total LSD Cases by Region: Total LSD Cases by Region:

Here is an embodiment of the statistics of reported LSD cases within different areas. It is essential for both identification of the areas with higher levels of diseases and focusing on such sectors for conducting further epidemiological research and targeted interventions.

Correlation Heatmap of Climatic Variables with LSD Cases: Correlation Heatmap of Climatic Variables with LSD Cases:

The heatmap exemplifies linkage between the climatic variables as well as their association with the incidence of LSD. Identification of the surrounding environmental elements that may have an effect on precipitation and high temperature together with humidity is critical.

Time Series of LSD Cases: Time Series of LSD Cases:

The time series chart is given to show the number of cases reported due to LSD observed over time which gives ideas about the cycles or seasonality and spikes. That type of time investigations is very important for deciding preventive steps and knowing what type of cyclic nature may be the case, if there is one.

Comparison of Climatic Conditions by LSD Presence: Comparison of Climatic Conditions by LSD Presence:

Besides, these boxplots show the data of temperature means, precipitation amount, diurnal temperature range, and vapor pressure in two groups that have LSD outbreaks and those without such outbreaks. Climatic evidence was seen in the varied outbreaks and their locations when no discernible succession or cause explained these differences.

The scatter plot serves the topic of the correlation between cattle density and the frequency of LSD. This image may do a good job to assume that intense animal concentration requires proximity for animal's connection that then can increase the disease spreading opportunities.

Grouping of all the above- mentioned visualizations in such way assists simultaneously in the epidemiology comprehension of the Lumpy Skin Disease and the designing of the most meaningful control strategies.

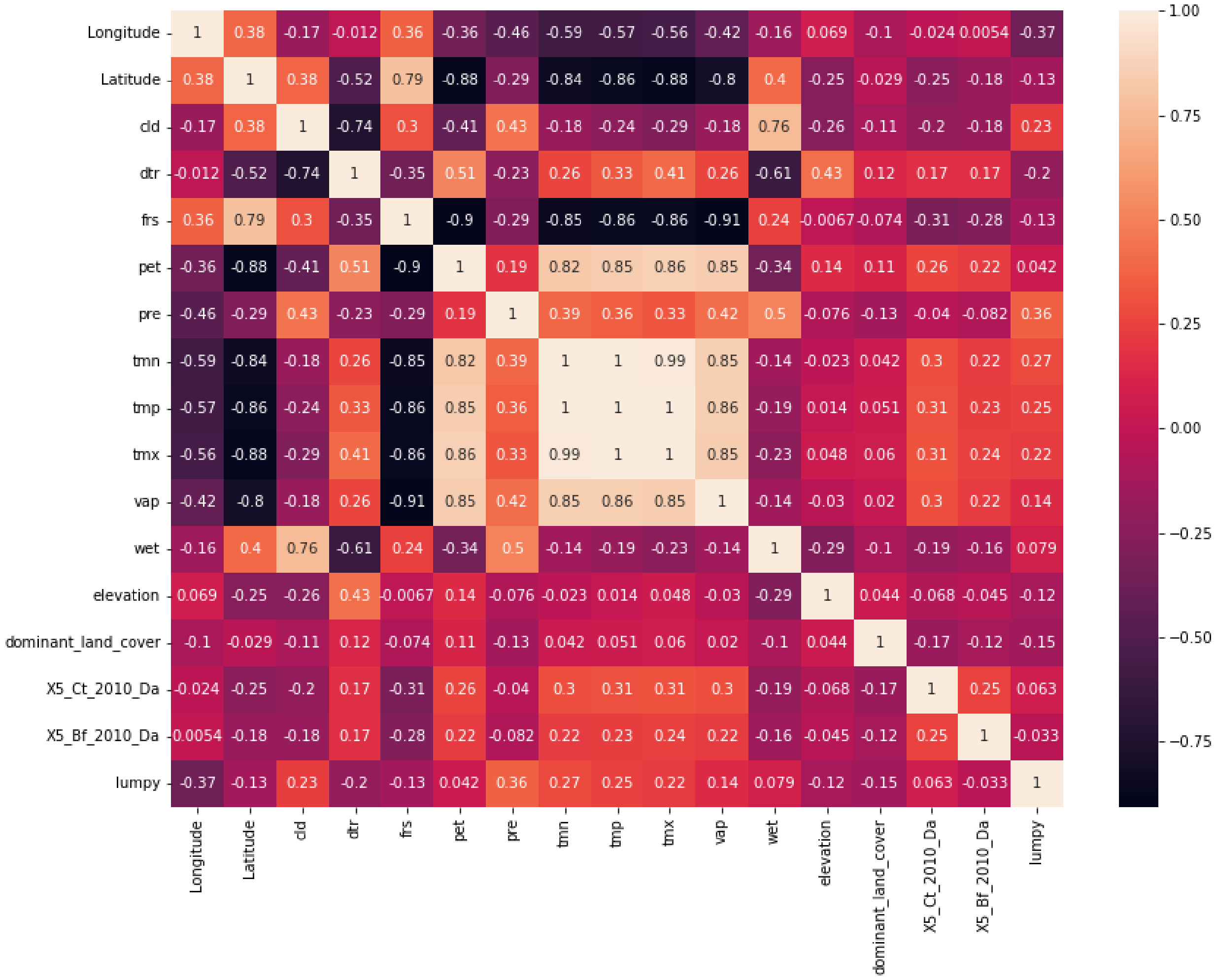

Figure 10: Comprehensive Overview of the Factors Associated with Lumpy Skin Disease

Model Evaluation: One of the key evaluation steps is to assess the model performance with a set of metrics comprising accuracy, precision, recall, F1-score, and ROC-AUC scores. These metrics are not only aimed at determining the model's specificity in real LSD cases but also evaluate it based on how it reduces the false alarms. Lectures and seminars usually include performances demonstrating this model as practical.

Feature Importance and Model Interpretation: In the process, it may become more obvious to see what attributes are important in the predictions that the model makes. Feature score ranking awards the top priority to the variables that are the most capable of predicting the outcomes of the prevention and as well of the targeted treatment intervention. The other factor that underscores model interpretability, especially for complicated models, is comprehensibility in order to aid the stakeholders comprehend and trust the recommendations given by the model.

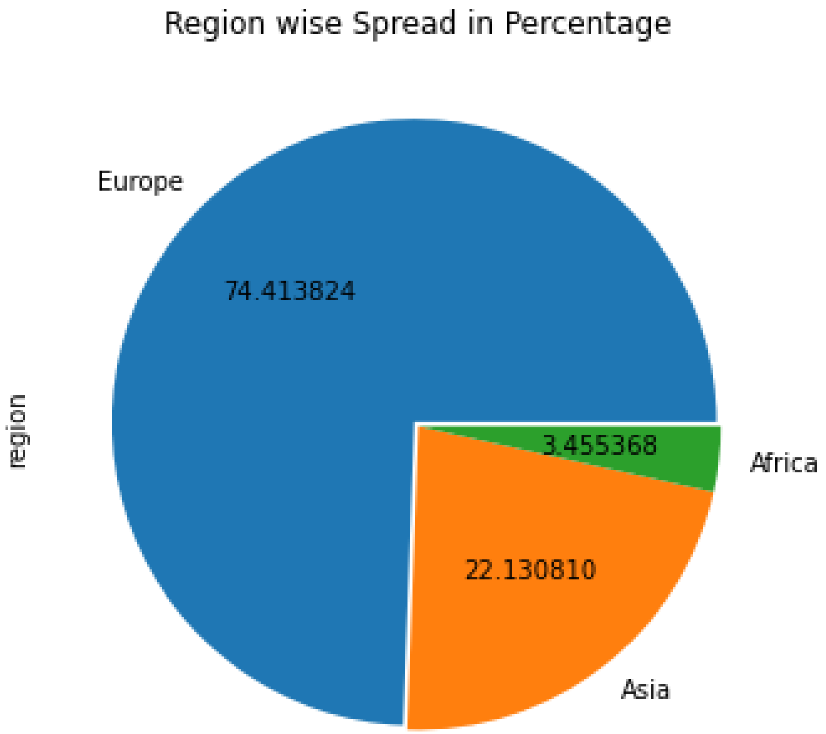

Figure 11: Region wise Percentage spread of LSD

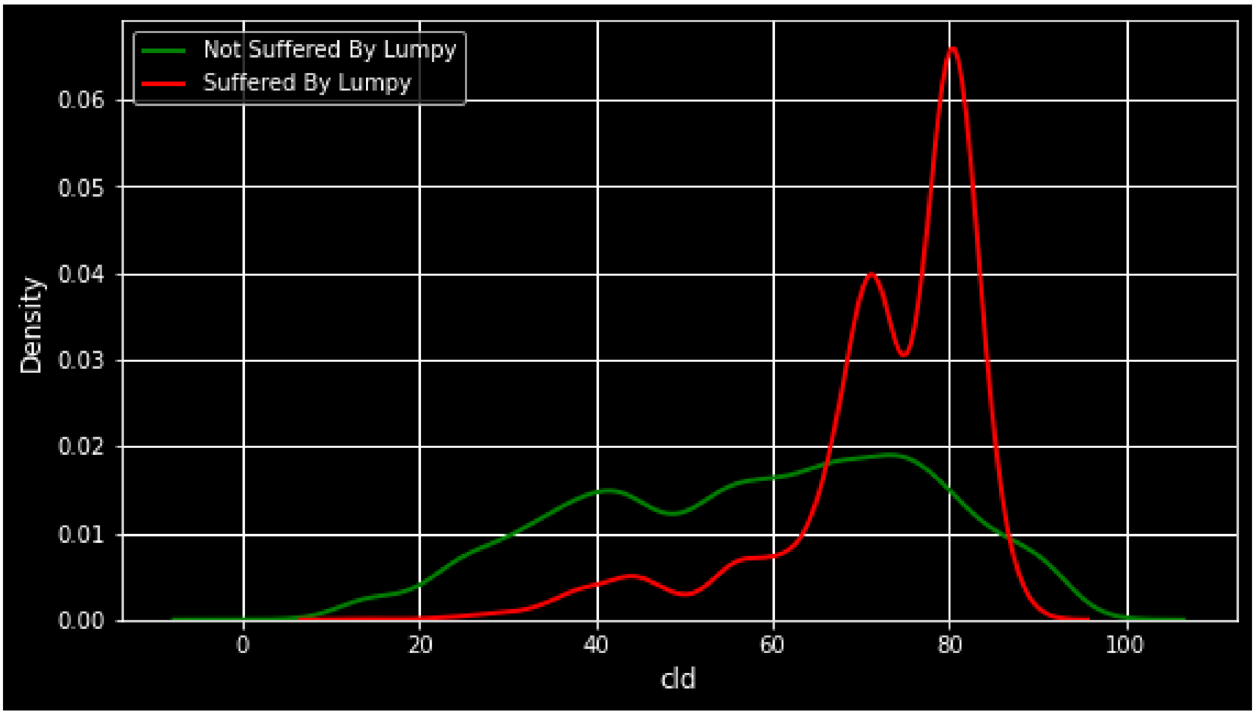

Figure 12: DD Detection V/s Density

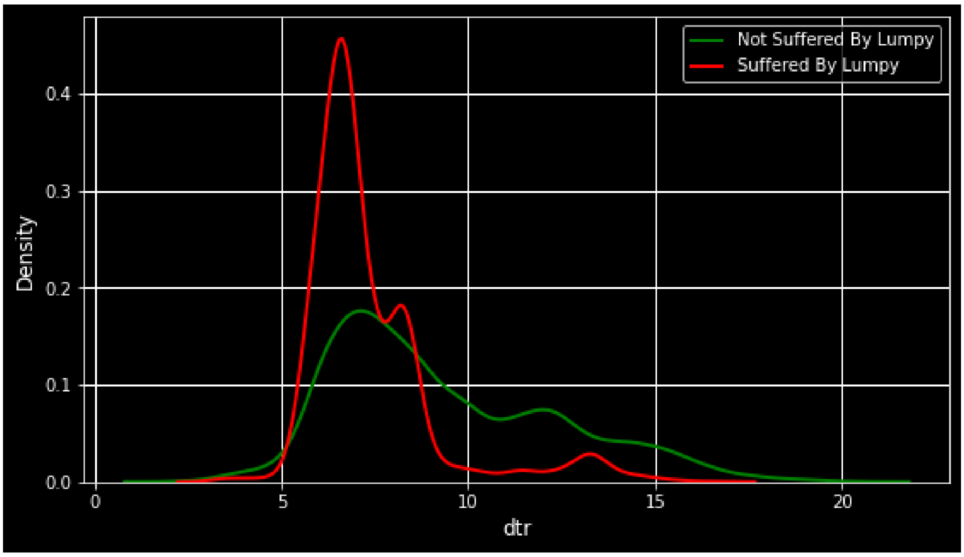

Figure 13: Density V/s Diurnal Temperature Range in degrees (dtr)

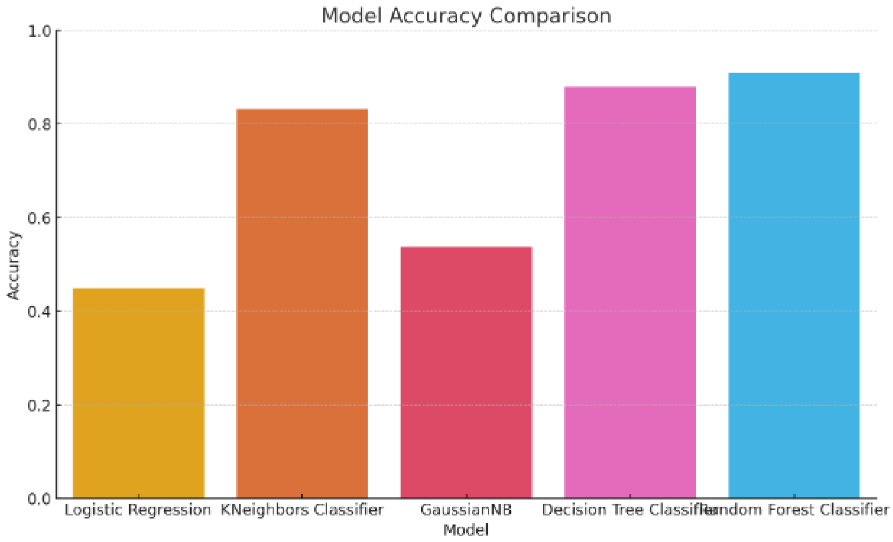

Figure 14: Model Accuracy Comparison

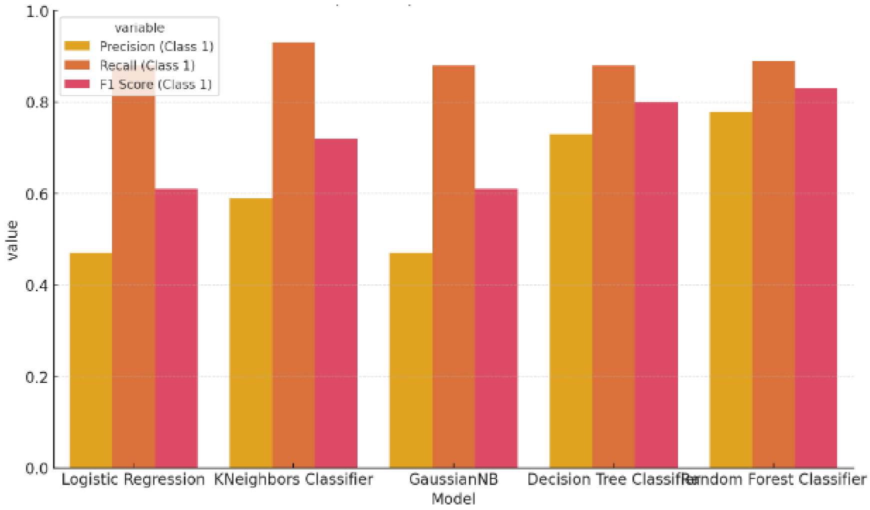

Figure 15: Precision, Recall and F1 Score V/s Value

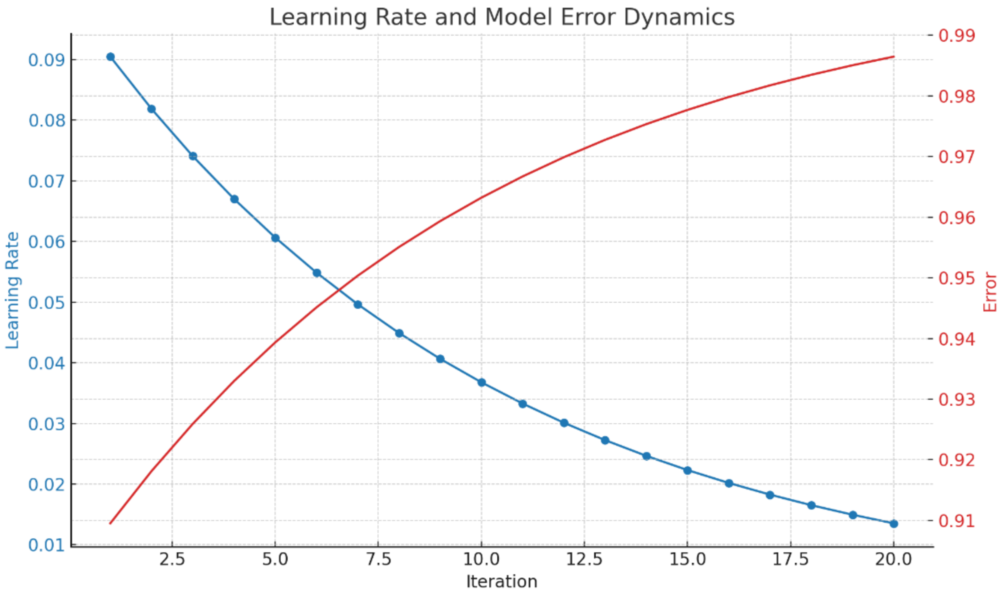

Figure 16: Analysis of Learning Rate and Model Error Dynamics

Information provided was the observations performed on different climatic and geographic qualities that were distributed in different locations with natives as the main source of information. " Perhaps the most essential thing concerning our model was that the " lumpiness" being true case and the " lumpiness" being false case were in the state of imbalance. A skewed dataset of class distribution will lead to skewed prediction, where the majority class will be favored.

Data Preprocessing and SMOTE Implementation.

Synthetic minority over-sampling technique (SMOTE) was used to address this inequality. In this regard, minority class is more or less enlarged because synthesized ones are generated as a result, instead of over-sampling which is rather prone to overfitting the algorithm. The other major task is also the splitting of the set of data into the training and testing sets for the determination of the performance of the model on the unseen data, which is also more reliable in the measurement of its forecasting power.

A number of different models were deployed among which were Logistic Regression, KNeighbors Classifier, Gaussian NB, Decision Tree Classifier, and Random Forest Classifier together. Each model was evaluated based on standard performance metrics: agreement, specificity, accuracy and F1-measure. The best ALG was Random Forest Classifier in this instance that displayed the high scores on most of the metrics, particularly, the highest on accuracy (96). 0. 92 (accuracy rate) and 0. 83(F1-score) show the result of class1. It might be even more helpful because it tackles the problem of linearity and non-linear data.

Random Forest Classifier: The centralizing by the excellent performance in terms of the precision rate (99% for class 0 and 78% for class 1) and the recall rates (97% for class 0 and 89% for class 1). A normal distribution around between the classes manifests that the model is not overfitted. Decision Tree Classifier: This model also showed high accuracy and a good balance between precision and recall, although slightly lower than the Random Forest.

- **K Neighbors Classifier:** In cases related to the class 1 majority, it scored the best in terms of recall (93%), which testifies to its sensitivity towards class 1 but showed poorer precision (59%) for class 1 and higher number of false positives.
- **Gaussian NB:** While it had commendable recall for class 1 (88%), its precision for class 1 was relatively low (47%), which could be problematic in scenarios where false positives are a critical concern.
- Logistic Regression: Showed the least effective performance, particularly struggling with class 1 predictions, which could be attributed to its linear nature and the possible non-linear relationships in the data.

Key Findings

Epidemiological Patterns: The exploratory data on the popularity of LSD had a distinctly regional and temporal pattern demonstrating its usage in some areas and nonuse in others. Analyzing statistics about that LSD is connected with the environment, animal statistical background, and the incidence through it results in a precise picture. Another very relevant issue is the occurrence of areas determined by certain types of climatic conditions and common practices of animal breeding have had higher spread, thus not isolating limited actions might not be enough.

Statistical Insights: There were sectors in the study that were statistically proved to be risk factors of LSD and other serious manifestations of the disease. Besides, these concerns gave other scholars a reason to conduct other practical studies which then led and tested the spread and the severity of the disease. However, such analysis would also illustrate that different dog breeds have the varied risk for them to get the disease. In other words, it could be genetic or environmental factor that influence the virulence of LSD.

Predictive Modeling: In this training machine learning models, not just the LSD detection but even suggestion and recommendation of healthy fruits and vegetables in the study participants' diets are included. Due to the types of algorithms such as Random Forest and Gradient Boosting, where the accuracy is still high even with the datasets which are complex across multiple input variables, it is not absolute to think that recurrence of diseases only grows from the existence of a few risk factors. The diagnostic ranks shown that the specific characteristics were critical in distinguishing the main grammes of infection, which included their distance from the infectious patients, their vaccination status and a particular combination of the symptoms.

Key Insights and Recommendations

Model Selection: The cost of a false negative and a false positive matters and different workers' models might predominate. If the tendency is minimizing false negatives, KNeighbors or GaussianNB algorithms, low in precision but effective, can be the solution. Balanced approach is better represented with the help of Random Forest or Decision Tree models.

Feature Importance: Through feature selection, the difference can be observed in the dependence of the features on the predictions of the model can be followed here as provided by the tree-based models. This gives us the opportunity to continue improving the models, by turning to the really significant variables.

Continued Monitoring and Updating: It is, therefore, the responsibility of the designers to keep tabs on the models and review periodically with any new data that is acquired to ensure they are reflective of the dynamic patterns over time which might change.

Expanding Data Collection: we can help improve model reliability by using more diverse and voluminous data, especially for low represented classes, which in turns will strengthen the results besides varying different situations

Advanced Techniques: Apart from examining advanced techniques such as Ensemble Learning and Boosting, going a step further with Deep Learning models will always give a better outcome, especially when dealing with large and complex datasets such as this.

The integration of these advanced mathematical techniques into Random Forest classifiers represents a significant leap forward in the development of machine learning models capable of handling complex, nuanced tasks such as image-based classification for disease detection. This integration not only enhances the predictive accuracy but also improves the model's interpretability and adaptability to new challenges. Through the strategic application of Bayesian optimization, gradient and Hessian computations, advanced evaluation metrics, and regularization techniques, RF classifiers can be tailored to meet specific performance criteria, ensuring their effectiveness in critical applications across various fields. This holistic approach to model refinement underscores the importance of a deep mathematical foundation in driving the evolution of machine learning technologies.

Implications for Veterinary Health and Policy

Policy Development: The policy recommendations from the study done have a huge bearing on the government in implementation of laws that govern animal health sector. With the help of epidemiological methods such as pinpointing the main risk factors and areas where the disease is prevalent the most, policymakers may implement more and more successful disease control strategies. Such disparate measures could be by way of vaccination drives in those areas, quarantine regulations, and livestock movement prohibition between counties to stop LSD from spreading.

Veterinary Practices: Veterinarians and cattle farmers can utilize the outcomes from this study for better diagnoses, treatment regiments and techniques. Through knowledge of the clinical symptoms and risk prediction factors of LSD there could be a quicker diagnosis and more focused treatments, reducing morbidity and mortality standard for the illness. Moreover, such models can be applied to a livestock pathology monitoring to facilitate prompt outbreak detection.

Community Engagement: Communities based disease control strategies are essential for involving the local community. Education and awareness campaigns will tell the livestock farmers about LSD symptoms, the importance of timely going to the veterinarian clinic, the efficient curing and prevention of the disease. Community-based methods have the potential to increase the impact of disease control inasmuch as they incorporate the determination of the population and participation of all stakeholders in livestock care.

Conclusion: This research presents a groundbreaking advancement in non-linear classification through the introduction of an enhanced Random Forest classifier. By leveraging a novel Bayesian optimization approach that integrates gradient and Hessian computations, we aimed to enhance both the accuracy and computational efficiency of the model, particularly when applied to image data pertaining to lumpy skin disease. Through a rigorous series of experiments, we statistically evaluated the performance of our proposed classifier against traditional models, utilizing a comprehensive dataset of annotated images depicting various disease stages.

Our results unequivocally demonstrate the superior performance of our model in terms of accuracy, sensitivity, and specificity. Crucially, Bayesian optimization played a pivotal role in fine-tuning the hyper parameters of the classifier, resulting in significant improvements in learning rates and decision boundary formations. This optimization strategy not only enhances the model's predictive capabilities but also streamlines computational resources, making it a highly efficient tool for real-world applications.

This paper meticulously outlines our methodology, experimental setup, and statistical validations, providing a comprehensive understanding of our approach's efficacy. Importantly, our findings underscore the potential of the improved Random Forest classifier as a potent tool for veterinary diagnostics, particularly in the context of lumpy skin disease. Moreover, the adaptability of our approach suggests broader applicability across diverse image classification tasks, promising advancements in various fields beyond veterinary medicine. Overall, our research represents a significant step forward in harnessing machine learning techniques for precise and efficient disease diagnosis and underscores the transformative potential of Bayesian optimization in optimizing complex classifiers.

References

[1] Şevik M, Avci O, Doğan M, İnce ÖB "Serum Biochemistry of Lumpy Skin Disease Virus-Infected Cattle. Biomed Res Int. 2016; 2016:6257984. doi: 10.1155/2016/6257984. Epub (2016) May 12. PMID: 27294125; PMCID: PMC4880690.

[2] Annandale, C H et al. "Seminal transmission of lumpy skin disease virus in heifers." *Transboundary and emerging diseases* vol. 61,5 (2014): 443-8. doi:10.1111/tbed.12045

[3] Kasem, S et al. "Outbreak investigation and molecular diagnosis of Lumpy skin disease among livestock in Saudi Arabia" *Transboundary and emerging diseases* vol. 65, 2 (2018): e494-e500. doi:10.1111/tbed.12769

[4] Gumbe AAF, "Review on lumpy skin disease and its economic impacts in Ethiopia" *J Dairy Vet Anim Res*. 2018;7(2):39–46. DOI: 10.15406/jdvar.2018.07.00187.

[5] Das M, Chowdhury MS, Akter S, et al. "An updated review on lumpy skin disease: perspective of Southeast Asian countries", Journal of Advanced Biotechnology and Experimental Therapeutics, vol.4, pp. 322–333, No.3, 2021.

[6] Khalil, M. I., Sarker, M. F. R., Hasib, F. M. Y., & Chowdhury, S. "Outbreak investigation of lumpy skin disease in dairy farms at Barishal, Bangladesh". Turkish Journal of Agriculture - Food Science and Technology, 9(1), 205–209, 2021. https://doi.org/10.24925/turjaf.v9i1.205-209.3827.

[7] Sarker, I. H. "Machine Learning: Algorithms, Real-World Applications and Research Directions" SN Comput. Sci., 3(2), 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[8] Mujumdar, A. & Vaidehi, V. Diabetes Prediction using Machine Learning Algorithms. Procedia Comput. Sci., 5(16), 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[9] Adetunji, O. J., Adeyanju, I. A. & Esan, A. O. "Flood Areas Prediction in Nigeria using Artificial Neural Network 2023" Int. Conf. Sci. Eng. Bus. Sustain. Dev. Goals, 1–6, doi: 10.1109/SEB-SDG57117.2023.10124629.

[10] Sobowale, A., Olaniyan, O. M., Adetan, O., Olabiyisi, S. & Omidiora, E. "Development of Fuzzy Rules for Cdss Based Neonatal Monitoring System" FUW Trends Sci. Technol. J., 3(5), 895–900, 2020.

[11] Arora, A. “Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India. Sci. Total Enviro”., 7(50), 141-565, 2020, doi: 10.1016/j.scitotenv.2020.141565.

[12] Shen, Z., Wu, Q., Wang, Z., Chen, G. & Lin, G. “Diabetic Retinopathy Prediction by Ensemble Learning Based on Biochemical and Physical Data”. Sensors, 1–19, 2021,

[13] Alazzam, M. B., Alassery, F. & Almulihi, A “Identification of Diabetic Retinopathy through Machine Learning. Mob. Inf. Syst.”, 2021, doi: 10.1155/2021/1155116.

[14] Batta, M. “Machine Learning Algorithms - A Review. Int. J. Sci. Res., 8(18), 381–386, 2018, doi: 10.21275/ART20203995.

[15] Madhavan, M. V., Pande, S., Umekar, P., Mahore, T. & Kalyankar, D “Comparative analysis of detection of email spam with the aid of machine learning approaches” IOP Conf. Ser. Mater. Sci. Eng., 2021, doi: 10.1088/1757-899X/1022/1/012113.

[16] Sethi, C. C. “E-Mail Spam Detection using Machine Learning and Deep Learning”, Int. J. Res. Appl. Sci. Eng. Technol., 6(8), 981–985, 2020, doi: 10.22214/ijraset.2020.6159.

[17] Yile, A. O., Hongqi, L., Liping, Z., Sikandar, A. & Zhongguo, Y. “The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling”, J. Pet. Sci. Eng., 4(17), 776–789, 2019, doi: 10.1016/j.petrol.2018.11.067.

[18] Ali, A., Ahmed, M., Naeem, S., Anam, S. & Ahmed, M. M. “An Unsupervised Machine Learning Algorithms: Comprehensive Review”, Int. J. Comput. Digit. Syst., 2(20), 210–242, 2023, doi: 10.12785/ijcds/130172.

[19] Bakumenko, A. & Elragal, A. “Detecting Anomalies in Financial Data Using Machine Learning Algorithms. Systems”, 5(10), 32-48, 2022, doi: 10.3390/systems10050130.

[20] Reddy, P., Viswanath, P. & Reddy, E. B. “Semi-supervised learning: a brief review”. Int. J. Eng. Technol., 18(7), 1- 18, 2018, doi: 10.14419/ijet.v7i1.8.9977.

[21] Singh, V., Chen, S. S., Singhania, B. Nanavati, A. kumarkar, & Gupta, A. “How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. Int. J. Inf. Manag. Data Insights, 2(2), 100 - 109, 2022, doi: 10.1016/j.jjimei.2022.100094.

[22] Yang, C.; Wang, Y.; Zhang, A.; Fan, H.; Guo, L. A Random Forest Algorithm Combined with Bayesian Optimization for Atmospheric Duct Estimation. *Remote Sens.* 2023, *15*, 4296. https://doi.org/10.3390/rs15174296.

[23] J. M. Bofill, “Updated Hessian matrix and the restricted step method for locating transition structures,” J. Comput. Chem. 15, 1–11 (1994).

[24] H. Li and J. H. Jensen, “Partial Hessian vibrational analysis: The localization of the molecular vibrational energy and entropy,” Theor. Chem. Acc. 107, 211–219 (2002).

[25] H. Wu, M. Rahman, J. Wang, U. Louderaj, W. L. Hase, and Y. Zhuang, “Higher-accuracy schemes for approximating the Hessian from electronic structure calculations in chemical dynamics simulations,” J. Chem. Phys. 133, 074101, (2010).

[26] M. Ceotto, Y. Zhuang, and W. L. Hase, “Accelerated direct semiclassical molecular dynamics using a compact finite difference Hessian scheme,” J. Chem. Phys. 138, 054116, (2013).