IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS 2025)
23-24 October 2025, Kushtia, Bangladesh

Explainable Machine Learning Framework for Detecting Lumpy Skin Disease with Environmental and Climate Factors

Md. Rifat Hossen
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
rifat.220614@s.pust.ac.bd

Md. Uzzal Mia
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
uzzal.220605@s.pust.ac.bd

Md. Noman Bhuiyan
Dept. of ICE, PUST,
Pabna, Bangladesh
noman112543@gmail.com

Manik Kumar Saha
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
manik.saha@pust.ac.bd

Rinku Islam
Dept. of CSE
Varendra University
Rajshahi, Bangladesh
rinku.vu.cse41@gmail.com

Md. Sarwar Hosain
Information and Communication Engineering
Pabna University of Science and Technology
Pabna, Bangladesh
sarwar.ice@pust.ac.bd

***Abstract*—Lumpy Skin Disease (LSD) is a viral illness affecting cattle, marked by rapid transmission and significant economic losses. Management challenges and diagnosis issues make LSD a major health and financial risk for the livestock sector. However, this study presents a comprehensive comparative analysis of eleven machine learning algorithms specifically aimed at binary classification within an environmental context. The evaluated algorithms comprise traditional classifiers, ensemble techniques, and advanced boosting methods, including support vector machine, k-nearest neighbors, random forest, decision tree, CatBoost, XGBoost, AdaBoost, linear classification, as well as stacking and voting classifiers. Evaluation was based on performance metrics like accuracy, precision, recall, and F1-score, with CatBoost achieving the highest results (98% accuracy, 91% precision, 91% recall, and 91% F1-score). Typically, ensemble methods surpassed standalone models in effectiveness. An assessment of feature importance, using CatBoost and SHAP (SHapley Additive exPlanations), identified cloud cover, diurnal temperature range, and frost days as vital predictors, corroborated by historical climate data from 2010 that demonstrated notable predictive relevance. The application of interpretability tools ensured that top-performing models remained transparent and actionable. This research offers a robust, interpretable, and flexible classification framework appropriate for environmental modeling and wider machine learning applications, underscoring the value of algorithm comparison and varied feature utilization for effective predictive modeling.**

***Index Terms*—Machine Learning, Binary Classification, CatBoost, Ensemble Methods, Explainable AI, Lumpy Skin Disease.**

I. INTRODUCTION

Lumpy Skin Disease (LSD) is a highly contagious viral infection in cattle, caused by a virus similar to sheeppox. It spreads through insect bites or direct contact with lesions, saliva, nasal secretions, milk, or semen of infected animals. Symptoms include skin nodules, fever, nasal discharge, swollen lymph nodes, excessive tearing, and reduced milk production, with potential for secondary bacterial infections, and severe cases can be fatal. LSD was first discovered in Egypt, initially in southern and eastern Africa. In the 1970s, it spread to northwest and sub-Saharan West Africa, reaching the Middle East by 2000, and by 2013, it had reached Turkey and the Balkans. Recent outbreaks in China, Bangladesh, Georgia, and Russia have raised global concerns, though Australia, New Zealand, and the Western Hemisphere have no reported cases.

Currently, no targeted antiviral therapy for LSD exists. Management involves supportive care: treating skin lesions, providing wound care, and using antibiotics to prevent secondary infections like pneumonia. Also, insect control strategies, including sprays, reduce transmission risk. Due to rapid spread and containment challenges, LSD poses a significant threat to cattle health globally.

Livestock plays a crucial role in agriculture and significantly supports the economies of various nations, including Bangladesh. It has a substantial impact on the nation's gross domestic product and agricultural sector. A study in Bangladesh revealed that the prevalence of LSD in cattle stands at about 26.5%, with a mortality rate of 0.26% and a case fatality rate of 0.97% [1]. Cattle and livestock provide not only food, such as dairy products, but also generate employment opportunities and a steady income source for remote urban areas. Poor handling and ineffective management of livestock can result in their scarcity, which leads to significant food shortages and adversely affects livelihoods. Additionally, awareness of cattle health remains alarmingly low, contributing to thousands of cattle deaths annually. In 2022, more than 155,366 cattle were affected or perished due to LSD [2]. LSD first emerged in Zambia, an African nation, and is highly contagious and perilous [3]. The Lumpy Skin Disease

Virus (LSDV), part of the Capripoxvirus genus, primarily impacts cows and water buffalos [4]. The Lumpy Skin Disease Virus is highly contagious and can persist for long durations in diverse environments, making early detection crucial for effective eradication [5]. LSD can significantly impact cattle in several ways. It may cause infertility, result in abortions, and severely reduce calf survival rates [6]. Additionally, it may lead to decreased milk production, causing significant economic losses for farmers. The affected cattle may also become more susceptible to other illnesses, such as Foot and Mouth disease and tuberculosis. Once the disease has peaked, no specific cure exists; however, early detection can aid in managing it through appropriate treatment. LSD symptoms comprise high fever, skin lesions, and body nodules. Photographic images and video footage can capture symptoms such as elevated nodules, watery eyes, scabs on the skin, and lesions in the oral and nasal cavities [7]. Machine learning is shown to be an efficient and non-invasive approach for the early detection of LSD [8]. Convolutional neural networks (CNN) enhance detection capabilities. These models identify subtle variations in images, enabling early detection of underlying LSD symptoms [9]. Early identification can halt disease transmission and may facilitate treatment. Additionally, image analysis can monitor the disease's progression in cattle. Several authors have focused on the early identification of LSD. Sivamurugan and Uthayan [10] evaluated various models, such as Xception, InceptionV3, VGG19, DenseNet121, ResNet50, and MobileNetV2, using a dataset of 6000 images categorized into three groups. The MobileNetV2 model demonstrated the highest accuracy at 0.9182. Karthikevan et al. [11] evaluated various ensemble-learning techniques, including random forest, extra trees classifier, sequential CNN, and pre-trained models (Dense121, Resnet50), on a dataset comprising 700 images to classify healthy and LSD cattle, achieving a peak accuracy of 96.85% with Resnet50.

Ujjwal et al. [12] utilized gradient boosting, KNN, decision trees, random forests, SVM, naive Bayes, Adaboost, and CNN to detect occurrences of lumpy skin disease. They examined these techniques using a dataset with 18603 instances and 16 features. Their analysis demonstrated that the random forest algorithm outperformed the other methods mentioned. Saqib et al. [13] implemented the MobileRMSNET model alongside the RMSprop optimizer. Their hybrid model achieved an accuracy of 95% on a dataset that includes 464 images of healthy cows and 329 images of infected cows, aiding in the management of LSD in dairy herds. The analysis done by the various authors has shown significant performance by the various individual models. However, the application of hybrid and ensemble-based models combined with explainable artificial intelligence (XAI) and LIME techniques remains an open area for research. The contributions of this paper are:

- This study employs ensemble models such as CatBoost, XGBoost, AdaBoost, stacking, and voting classifiers to accurately predict LSD, a topic seldom addressed in veterinary disease modeling. Incorporating geo-environmental factors like temperature, precipitation, vapor pressure, land cover, and elevation improves the accuracy of predictions.
- Explainable AI (XAI) techniques, like SHAP and LIME, are used to clarify model decisions, enhancing transparency and trust among domain experts. This incorporation of interpretable machine learning promotes informed decision-making in animal healthcare.

The rest of the paper is divided into three sections. Section II details the methodology and explains the proposed model comprehensively. Section III follows with the results and discussions, offering a comparative analysis of various models. Finally, Section IV concludes with a summary of the paper's findings.

II. Methodology

This section details the steps taken to create a machine learning classification framework aimed at detecting LSD in cattle, utilizing environmental and geospatial features. The process encompasses data preprocessing, feature selection, model training with various machine learning algorithms, and interpretability through SHAP and LIME techniques. Fig. 1 illustrates the model diagram.

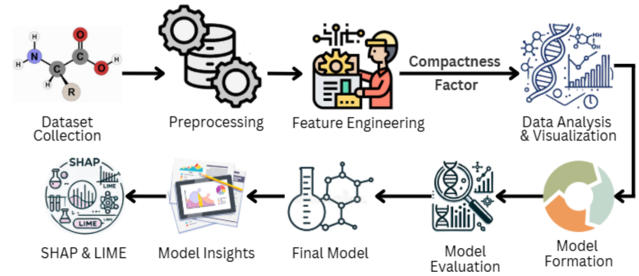

Fig. 1: Block Diagram of Proposed Model.

A. Data Collection and Pre-processing

The Mendeley Lumpy Skin Disease archive dataset [14] includes environmental, climatic, and geographical variables related to LSD prevalence. Each record includes temperature metrics (mean, max, min), precipitation, cloud cover, vapor pressure, land cover types, elevation, and other spatial features. The target variable denotes LSD presence or absence. After cleaning for missing values, the dataset was structured for supervised classification, normalizing continuous variables and encoding categorical ones. Exploratory data analysis identified 14 relevant features within 24,803 entries, including geographical coordinates, climatic variables (cloud cover, precipitation, temperature), elevation, land cover types, and binary labels for LSD. Data was sourced from "Lumpy skin disease data.csv" and loaded into a Pandas DataFrame. Preliminary inspection revealed missing values in categorical features (region, country, reportingDate), which were resolved during preprocessing, while numerical features remained complete and standardized.

B. Feature Selection and Feature Engineering

The dataset consists of both numerical and categorical variables. The key features examined include:

1) Climatic Variables: Cloud cover (cld), diurnal temperature range (dtr), frost frequency (frs), potential evapotranspiration (pet), precipitation (pre), and minimum (tmn), mean (tmp), and maximum (tmx) temperatures, along with vapor pressure (vap) and wet-day frequency (wet).

2) Geographical and Environment Variables: Elevation, the primary land cover type, and the density of cattle and buffalo.

3) Target Variable: Binary label (lumpy) indicating presence (1) or absence (0) of LSD.

Categorical variables were encoded, and numerical features were scaled to standardize their ranges. The significance of features was assessed through correlation analysis and tree-based methods, pinpointing the predictors that most significantly influence LSD occurrence.

C. Exploratory Data Analysis & Visualization

Categorical variables were encoded, and numerical features were scaled to standardize their ranges. The significance of features was assessed through correlation analysis and tree-based methods, pinpointing the predictors that most significantly influence LSD occurrence.

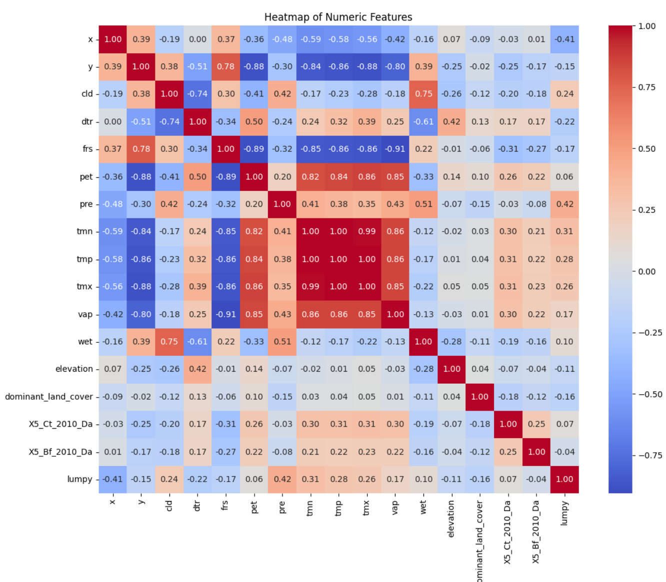

Fig. 2: Heat map depicting the strength of relationships between features.

Fig. 2 shows that the correlation heatmap displays significant intercorrelations among temperature variables ($r > 0.95$) and moderate correlations between moisture indicators. The lumpy status variable positively correlates with temperature (r ≈ 0.3-0.4) and moisture variables (r ≈ 0.1-0.4), indicating climate dependencies linked to disease occurrence.

D. Model Classification

1) Linear Classification: Linear classifiers distinguish data points by utilizing a linear decision boundary. An example of this is Logistic Regression, which models the probability of class membership through the sigmoid function applied to a linear combination of features. The fundamental equation is as Eq. 1:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \tag{1}$$

2) K-Nearest Neighbors Classification: KNN classifies a data point according to the most common label among its k nearest neighbors, using a distance metric such as Euclidean distance [15]. The distance between points $x_i x_i$ and $x_j x_j$ as Eq. 2:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{2}$$

3) Random Forest Classification: The random forest algorithm is a meta-learning machine learning technique [16]. It uses various random tree classifications to establish an overall classification for a given set of inputs. The algorithm's advantages come from its strong learning ability, reliability, and practical hypothesis space [17]. It is defined as Eq. 3

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{3}$$

4) Decision Tree Classification: Decision Trees divide data according to feature thresholds to enhance class purity, utilizing metrics such as Gini impurity Eq. 4:

$$\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

5) GBoost Classification: Gradient Boosting sequentially constructs additive models, with each new model addressing the residuals of its predecessors. The update for the model is as Eq. 5:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5}$$

6) XGBoost Classification: : Gradient boosting machines (GBMs) are regression strategies similar to boost-ing [18]. These machine-learning systems are effective in various applications [19]. In GBMs, new models are fitted sequentially to enhance response variable estimation accuracy. This technique constructs new base learners aligned with the negative gradient of the loss function for the entire ensemble. Essentially, we combine weak learners to create a robust model for specific problems. This is calculated using the following formula [20] as Eq. 6.

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

7) Support Vector Classification: SVM identifies the hyperplane that optimizes the margin between classes, framed as a constrained optimization problem as Eq. 7:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{7}$$

8) Adaboost Classification: AdaBoost incrementally modifies weights on training samples to emphasize more difficult cases and integrates weak classifiers based on their accuracy. The weight of the classifier as Eq. 8 and 9:

$$\alpha_m = \frac{1}{2} \ln\left(\frac{e_m}{1-e_m}\right) \tag{8}$$

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right) \tag{9}$$

9) CatBoost Classification: CatBoost is a gradient boosting algorithm that naively manages categorical features and employs ordered boosting to minimize overfitting. Its model update process mirrors that of gradient boosting as Eq. 10:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{10}$$

10) Stacking and Voting Classification: Stacking involves training a meta-classifier on the predictions generated by several base classifiers. Voting consolidates the outputs of base classifiers, utilizing majority rules (hard voting) as Eq. 11:

$$\hat{y} = \arg\max_{c} \sum_{t=1}^{T} P_t(y = c \mid x) \tag{11}$$

E. Final Model & Evaluation

To evaluate the models, various metrics were used. Accuracy, the ratio of correct classifications to total samples, was primary. However, since accuracy can be misleading with imbalanced data, metrics like precision, recall, and F1-score were also calculated.

The confusion matrix illustrated true positives, false positives, true negatives, and false negatives, providing insights into error types. This multi-metric evaluation ensured a comprehensive and reliable comparison of the models. The suggested CatBoost model outperformed traditional algorithms, demonstrating its ability to capture complex nonlinear relationships in the LSD dataset.

F. Model Insights

In this study, we employ Explainable AI (XAI) techniques such as SHAP and LIME to enhance the interpretability of our CatBoost model used for classifying Lumpy Skin Disease. SHAP (SHapley Additive exPlanations) provides a robust and theoretically sound method for assessing the contribution of each input feature to model classification, offering a comprehensive perspective on feature importance [21]. On the other hand, LIME (Local Interpretable Model-agnostic Explanations) focuses on producing interpretable models that accurately explain individual predictions, thereby clarifying the behavior of complex models for human understanding [19]. By integrating both global and local interpretability tools, we validate our model's decisions, increase transparency, and improve trust and usability in real-world applications within the livestock industry.

III. Results and Discussion

Table II and Fig. 3 provide a thorough comparison of eleven classification algorithms assessed using the dataset. The performance metrics consist of accuracy, precision, recall, and F1-score, offering a complete perspective on each model's predictive abilities.

TABLE I: Comparison of LSD Detection Studies

Ref	Models Used	Dataset	Accuracy	XAI
[10]	Xception, VGG19, ResNet50, MobileNetV2	6000 img (3 cls)	91.82%	No
[11]	RF, Extra Trees, CNN, ResNet50	700 img (2 cls)	96.85%	No
[12]	GB, KNN, DT, RF, SVM, NB, CNN	18603 rows (tabular)	RF best	No
[13]	MobileRMSNET + RMSprop	793 img (329+464)	95%	No
Our	**Multiple's Machine Learning Models**	**Climate (tabular)**	**98% (catboost)**	**Yes (SHAP, LIME)**

TABLE II: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Linear Classification	0.95	0.78	0.77	0.78
KNN Classification	0.96	0.83	0.89	0.86
Random Forest Classification	0.97	0.84	0.91	0.87
Decision Tree Classification	0.97	0.84	0.92	0.88
GBoost Classification	0.97	0.91	0.85	0.88
XGBoost Classification	0.97	0.85	0.91	0.88
SVM Classification	0.96	0.85	0.80	0.82
Adaboost Classification	0.96	0.84	0.82	0.83
Catboost Classification	**0.98**	**0.91**	**0.91**	**0.91**
Stacking Classification	0.97	0.88	0.89	0.88
Voting Classification	0.97	0.87	0.89	0.88

The CatBoost Classification algorithm displayed excellent performance across all evaluation metrics, attaining the highest accuracy at 98%, a precision of 91%, recall of 91%, and an F1-score of 91%. This exceptional performance makes CatBoost the best choice for this classification task. The Extra Tree Classification algorithm trailed slightly behind, reporting an accuracy of 96.61%, precision of 82.71%, recall of 91.43%, and an F1-score of 86.85%. Other ensemble methods showed remarkable outcomes, with XGBoost Classification achieving an accuracy of 97.00% and Random Forest Classification reaching 96.75%. In contrast, traditional algorithms like SVM Classification, which has an accuracy of 95.75%, and KNN Classification, at 96.37%, performed sufficiently but fell short compared to the ensemble methods' performance.

Fig. 4 presents the confusion matrix for the best-performing model, showcasing outstanding classification performance characterized by high rates of true positives and true negatives. The model accurately identified 4,296 instances as class 0 (true negatives) and 555 instances as class 1 (true positives). The misclassification rates were impressively low, with only 52 false negatives and 58 false positives, reflecting a strong ability to differentiate between the two classes.

Fig. 5 of the CatBoost feature importance analysis highlighted the key variables impacting the classification process. Climate-related factors, namely X5_Bf_2010_Da and X5_Ct_2010_Da, showed considerable predictive strength, indicating that temporal climate trends are vital for precise classification. Additionally, environmental elements like 'dominant land cover', elevation, and several meteorological variables (wet, vap, tmx, tmp, tmn, pre, pet) played a role in the model's decision-making process, each with differing levels of significance.

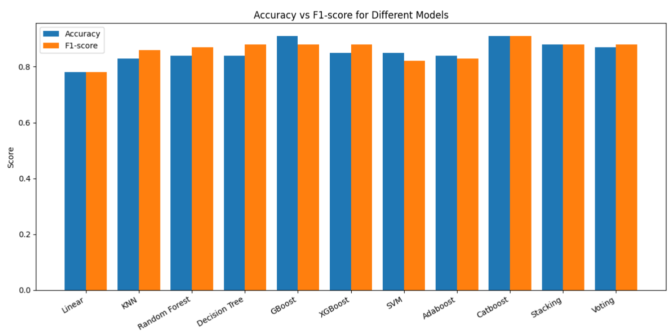

Fig. 3: Accuracy & F1-score Based Performance Model.

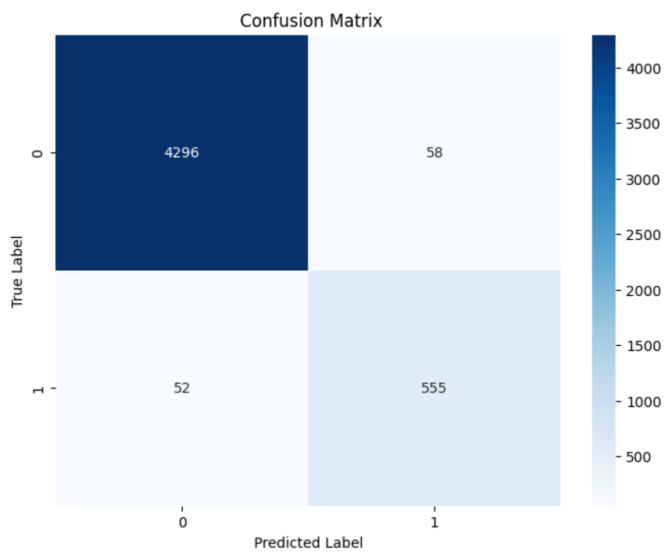

Fig. 4: Confusion Matrix of Based Performance Model.

The LIME analysis gave in-depth insights into the interpretability of individual predictions. Fig. 6 illustrates a case where the model predicted a 98% probability for class 0 and 2% for class 1. The decision tree visualization displays the step-by-step decision-making process, highlighting important thresholds such as cld is greater than or equal to -0.78, X5_Bf_2010_Da is greater than -0.17, and tmx is greater than 0.77. The SHAP summary plot, Fig. 7, shows how feature values affect predictions. High SHAP values (red/pink) tend to push predictions toward one class, while low values (blue) steer them toward the opposite class. This plot indicates that X5_Bf_2010_Da has varying effects, both positive and negative, depending on its value, whereas cld consistently presents significant negative SHAP values in most cases. To evaluate our approach's effectiveness, we conducted

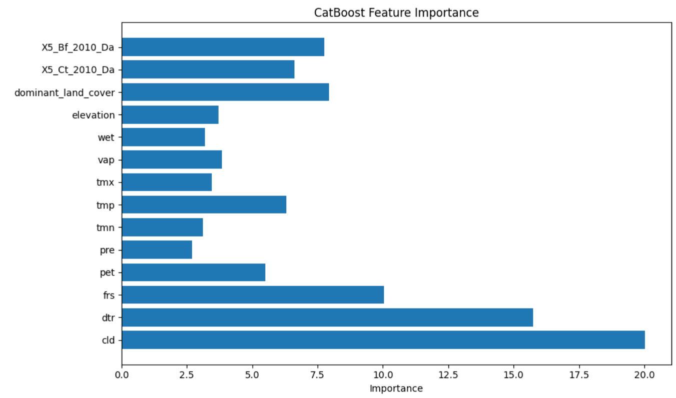

Fig. 5: Feature Importance of Based Performance Model.

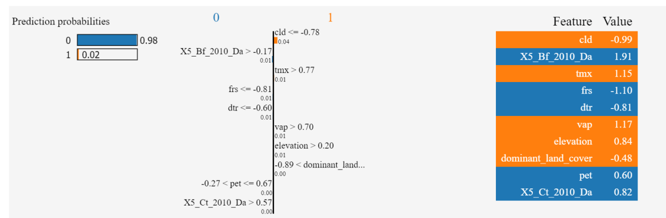

Fig. 6: LIME analysis showing local feature contributions to model predictions.

a comparative analysis with recent studies on LSD. Most methods, as shown in Table I, used deep learning architectures like ResNet50, DenseNet121, and MobileNetV2, achieving accuracies of 91% to 96.85%. However, these studies lack explainable AI (XAI) techniques, limiting their transparency in veterinary contexts. In contrast, our model achieved 98%

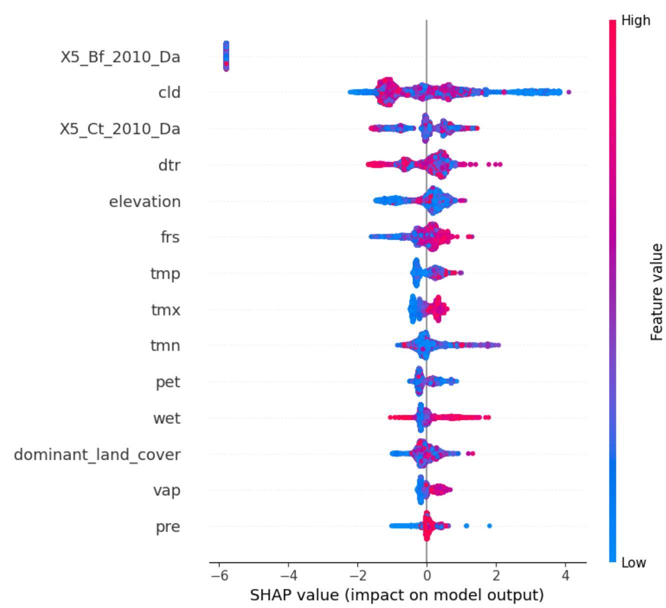

Fig. 7: The influence of different features on the model's predictions.

accuracy using CatBoost and integrated SHAP and LIME for post-hoc interpretability, enhancing its suitability for informed decision-making in animal health diagnostics.

Although the model demonstrates high accuracy, there is a concern regarding overfitting and restricted temporal generalizability. Its significant dependence on climate-related features could impact its performance if climate patterns change. Future efforts should incorporate a variety of features and conduct cross-temporal evaluations to improve its robustness.

IV. Conclusion

This study compares eleven machine learning classification algorithms for binary tasks with environmental data. The models include traditional algorithms (Logistic Regression, SVM), ensemble methods (Random Forest, XGBoost, AdaBoost, Voting, Stacking), and advanced gradient boosting techniques (CatBoost). The models were evaluated for accuracy, precision, recall, and F1-score to ensure robustness and generalizability. The CatBoost algorithm achieved the highest performance with 98% accuracy and exceptional precision, recall, and F1-score. Ensemble models like XGBoost also showed strong performance, surpassing individual classifiers. We used SHAP to interpret the CatBoost model and identify the most influential features, with cloud cover (cld), diurnal temperature range (dtr), and frost days (frs) as key predictors.

Our study identifies effective classification models and emphasizes the importance of interpretability in practical applications. By integrating explainable AI with efficient classifiers, we link predictive accuracy to better decision-making. This framework adapts to similar challenges, providing valuable insights for future research.

References

[1] M. A. Uddin, M. A. Islam, A. K. M. A. Rahman, M. M. Rahman, A. K. M. Khasruzzaman, M. P. Ward, and M. T. Hossain, "Epidemiological investigation of lumpy skin disease outbreaks in bangladeshi cattle during 2019-2020," *Transboundary and Emerging Diseases*, vol. 69, no. 6, pp. 3397–3404, Nov. 2022, epub 2022 Sep 12.

[2] Open Government Data (OGD) Platform India, "State/uts-wise details of the death of cattle due to lumpy skin disease (lsd) in the country during 2022," https://data.gov.in, 2022, answers Data of Rajya Sabha Questions for Session 258.

[3] L. Khan and M. Gul, "A comprehensive review on lumpy skin disease," *Pure and Applied Biology*, vol. 13, no. 3, Sep. 2024.

[4] G. Moudgil, J. Chadha, L. Khullar, S. Chhibber, and K. Harjai, "Lumpy skin disease: Insights into current status and geographical expansion of a transboundary viral disease," *Microbial Pathogenesis*, vol. 186, p. 106485, Jan. 2024.

[5] M. Akther, S. H. Akter, and S. Sarker, "Global burden of lumpy skin disease, outbreaks, and future challenges," *Viruses*, vol. 15, no. 9, p. 1861, Aug. 2023.

[6] Y. R. Khan, A. Ali, and K. Hussain, "A review: Surveillance of lumpy skin disease (lsd) a growing problem in asia," *Microbial Pathogenesis*, vol. 158, p. 105050, Sep. 2021.

[7] E. Mulatu and A. Feyisa, "Review: Lumpy skin disease," *Journal of Veterinary Science & Technology*, vol. 9, no. 3, 2018.

[8] M. A. Niloy, T. Bhowmik, J. Abedin, S. J. Ferdous, and I. Jahan, "Exploring machine learning techniques for symptom-based detection of livestock diseases," Ph.D. dissertation, Brac University, 2024.

[9] D. K. Saha, "An extensive investigation of convolutional neural network designs for the diagnosis of lumpy skin disease in dairy cows," *Heliyon*, vol. 10, no. 14, 2024.

[10] V. Sivamurugan, K. R. Uthayan, and V. Thanikachalam, "Cow disease (lsd) classification system for predicting different severity levels," *International Journal of Early Childhood Special Education*, vol. 14, no. 8, 2022.

[11] S. Madhanakumar *et al.*, "Lumpy skin disease prediction using machine learning," in *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*. IEEE, 2024, pp. 887–895.

[12] N. Ujjwal, A. Singh, A. K. Jain, and R. G. Tiwari, "Exploiting machine learning for lumpy skin disease occurrence detection," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2022, pp. 1–6.

[13] S. Muhammad Saqib, M. Iqbal, M. Tahar Ben Othman, T. Shahazad, Y. Yasin Ghadi, S. Al-Amro, and T. Mazhar, "Lumpy skin disease diagnosis in cattle: A deep learning approach optimized with rmsprop and mobilenetv2," *Plos one*, vol. 19, no. 8, p. e0302862, 2024.

[14] E. Afshari Safavi, "Lumpy skin disease dataset," 2021. [Online]. Available: https://doi.org/10.17632/7pyhbzb2n9.1

[15] M. R. Hossen, M. U. Mia, R. Islam, M. S. Hosain, M. K. Hasan, and T. Shimamura, "Facial expression recognition: A machine learning approach with svm, random forest, knn, and decision tree using grid search method."

[16] M. Z. H. Zamil, M. R. Islam, S. Debnath, M. T. Mia, M. A. Rahman, and A. K. Biswas, "Stroke prediction on healthcare data using smote and explainable machine learning," in *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2025, pp. 1–6.

[17] Y. Ao, H. Li, L. Zhu, S. Ali, and Z. Yang, "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling," *Journal of Petroleum Science and Engineering*, vol. 174, pp. 776–789, 2019.

[18] T. Bikmukhametov and J. Jäschke, "Oil production monitoring using gradient boosting machine learning algorithm," *Ifac-Papersonline*, vol. 52, no. 1, pp. 514–519, 2019.

[19] J. Al-Faruk, M. R. Hossen, J. Das, E. Hossain, Y. Safat, and O. J. Akhi, "Evaluating targeted productivity in bangladesh's garment sector using machine learning and deep learning with explainable AI: A data-driven method for enhanced production planning," in *Proc. 3rd Int. Conf. on Big Data, IoT and Machine Learning (BIM)*. Taylor & Francis Group, 2025, pp. 1–20.

[20] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[21] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," *arXiv preprint arXiv:1706.06060*, 2017.