Lumpy Skin Disease Prediction Using Different Machine Learning Techniques

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

7th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

8th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Lumpy Skin Disease is a highly infectious, fatal illness that is commonly observed in cattle. The common symptoms of this disease are fever, infertility, reduced milk production and so on. Furthermore, the mortality rate of cattle infected by Lumpy Skin Disease is quite low, hence predicting the outcome of this disease earlier can reduce economic loss significantly. This research was conducted to predict if cattle are infected with Lumpy Skin Disease or not with the use of various machine learning models. A total of ten machine learning classifiers have been used and evaluation metrics were calculated for determining how well the classifiers have performed. Among all the classifiers, Random Forest Classifier and Light Gradient Boosted Machine Classifier have outperformed the other models with the F1 score of 98%.**

***Index Terms*—Lumpy Skin Disease, Classifier, SMOTE, Random Forest, Light Gradient Boosted Machine Classifier**

I. Introduction

Lumpy skin disease is a highly contagious virus that affects cattle and is indicated by fever and the formation of necrotizing skin nodules. Similar lesions can arise in the skeletal muscles and respiratory and digestive mucosae. The sickness is marked by subcutaneous edema of the limbs and ventral areas of the body, as well as widespread lymphadenitis [1]. It was first reported in Zambia in 1929 [2].

Initially, the animals possessed high fever which is followed by the development of large nodules ranging 5cm in diameter [3]. The nodules are more prevalent near head, neck, udder, scrotum, and perineum. Several African countries now have endemic cases of the disease. LSDV is transmitted by secretions and sperm by insects, arthropods that feed on blood, tainted food and drink, as well as saliva and nasal secretions in the later stages of the disease [4].

Machine Learning Models are often used for early diagnosis of diseases or prediction purposes. Ehsanallah Afshari [5] had shown the use of machine learning tools to detect the impact of meteorological and geospatial features on Lumpy Skin Disease. According to his research, the Artificial Neural Network has performed remarkably in terms of AUC and F1 scores while tested with unlabelled testing Data.

Machine learning algorithms were used to anticipate swine disease outbreaks around the world using bio-climatic parameters [6]. In the dataset that contains all the predictive aspects, Random forest performed really well showing 80.4 percent accuracy. On the other hand, SVM strategy showed 76.02 percentage accuracy in the subset dataset.

Soil and feces samples were obtained from 11 pastured poultry farms in the United States between 2014 and 2017 [7]. To estimate Listeria spp, they developed Random forest and boosting (gradient) machine predictive models which are prevalence in samples using meteorological data from the farming region. The output were 0.905 and 0.855 for Random forest and gradient boosting respectively in terms of AUC using fecal data. The research of [9] had shown the use of machine learning in the diagnosis of breast cancer. It showed a comparative analysis between three well-known machine learning classifiers that are KNN, Naive Bayes, and Random Forest. They have used the Wisconsin Diagnosis Breast Cancer dataset for their purpose.

In this research, the authors have used a Lumpy Skin

Disease Dataset [8] to train ten machine Learning Classifiers and then compare the outcomes in terms of F1 score, recall, and precision. Synthetic Minority Oversampling Technique (SMOTE) has been used to resolve the imbalanced nature of data before fitting the classifier. The classifiers include Decision Tree Classifier, Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier, SGD Classifier, LGBM classifier, K- Nearest Neighbour, Gradient Boosting Classifier, Gaussian Naive Bayes Classifier, and XGD Classifier. Random Forest Classifier and LGBM classifier performed equally well with macro average of Recall and Precision is 98%.

II. Dataset

A. Data Collection & Description

The dataset used in this research was collected from Lumpy Skin disease dataset from Mendeley Data [8] which was compiled from various resources [11]–[13]. The aim of using this particular dataset is to use machine learning techniques to predict LSDV related diseases based on meteorological and geospatial variables. This dataset consists 24803 instances and 19 different features along with the class label. Geographic coordinates of LSDV outbreaks, as well as country, region, and reporting date; meteorological characteristics such as cloud cover percentage (monthly), diurnal temperature, wet day frequency, frost day frequency, monthly precipitation, mean temperature (daily), potential evapotranspiration (per day), average maximum and minimum temperature (monthly), vapor pressure; global geospatial elevation; spatial information on land cover; and lastly buffalo and cattle population density data are included in the features. The binary class label defines whether or not a specific instance is associated with Lumpy Skin Disease.

B. Data Preprocessing

In the preprocessing steps, initially the columns of region, country and reporting date were dropped because those features had almost 88% missing values among total instances. The rest of the features did not contain any missing value.
Next, the frequencies or value counts of the class labels were checked. It was found that the class labels are imbalanced with 21,764 instances labeled as 0 with only 3039 samples labeled as 1. To deal with this problem, the two possible options are doing undersampling or oversampling. But the issue with undersampling is that a lot of data get lost in the process. That is why in this research, the authors preferred oversampling technique. Out of several techniques, SMOTE was used for this work. This method generates synthetic data for the minority group. The significant advantage of SMOTE rather than other oversampling techniques is that this process does not generate exact duplicates, but it creates data points that are slightly different from the original. This helps the final model with less overfitting than other approaches. Fig. 1 shows the visualization of value counts before and after SMOTE oversampling on the dataset. It can be observed that after oversampling both 0 and 1 classes have 21764 equal instances.
After oversampling, the data was standardized using standard scaler so that the mean of the values of each feature becomes 0 and standard deviation becomes 1. Next, the standarized data was randomly splitted into 80-20 train-test ratio. After that, this preprocessed data was used to train and evaluate the classifiers.

III. Methodology

A. Algorithm Description

1) Logistic Regression: Logistic Regression (LR) is an approach for predictive modeling which examines the link between multiple independent variables and predicts the values of dependent factors in which the range is in between 0 and 1 with the use of sigmoid function. If the value is smaller than 0.5, it is labeled as 0, and otherwise, it is labeled as 1 It is generally used for categorizing a large set of information.

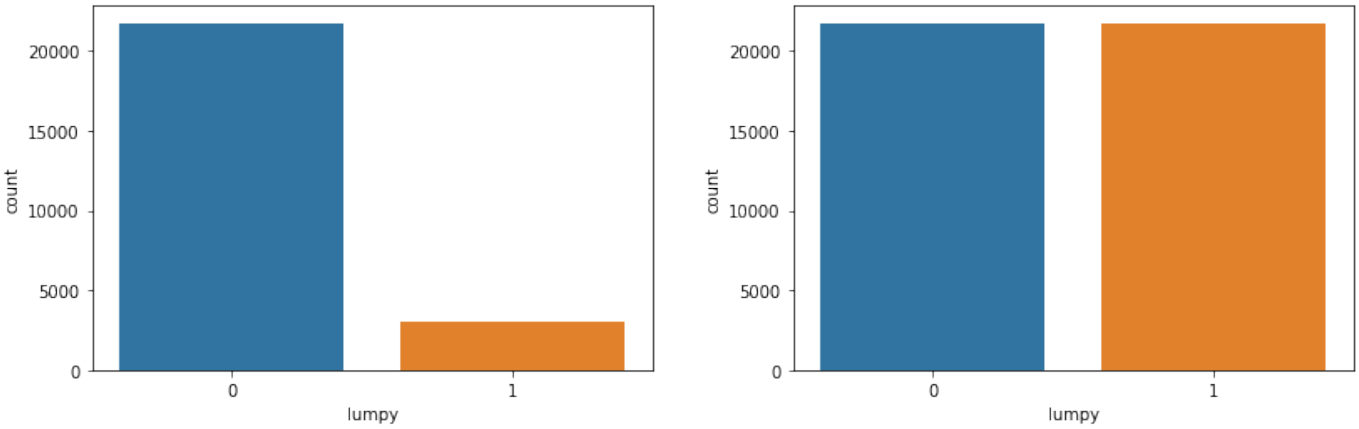
2) Decision Tree Classifier: The training data is used to teach the decision tree classifier simple decision rules and predict the value or the class of targets. It has a structure of tree, in which the branches refer to decision rules, internal nodes refer to the features of the dataset and the leaf or terminal nodes mean the outcome. It is a commonly used supervised learning technique since it is easy to comprehend as it can be visualized as a tree, and it also does not require much data processing.

3) Random Forest Classifier: The random forest classifier is a group of decision tree classifiers where each of the single trees predicts a class and the one with the most votes is taken as the final prediction of the model. This is powerful because compared to individual decision tree based models, a group of them produces much better performance and the reason behind it is that the correlation between those models is low. To make sure of the low correlation, two methods such as bagging and feature randomness are applied.

4) Support Vector Classification: The support vector classifier finds hyperplanes, which can be called as the best decision boundary, are created that separates the n-dimensional space, distinctly classifying the data points. Each data point in the SVM (Support Vector Machine) algorithm is plotted in a multidimensional space where each feature's value is the value of a specific coordinate. In order to create the hyperplane, the extreme cases or the support vectors, the closest data points to the hyperplane that can influence the position of the hyperplane are chosen by the SVM.

5) k-Neighbors Classifier: K nearest neighbors algorithm considers the similarity between the available and new data and categorizes them accordingly. The similarity means the closeness or the distance between those points, so the euclidean distance between the K number of neighbors is calculated.

6) Gaussian Naive Bayes: The features in Naive Bayes Classifier, a probabilistic ML technique based on the Bayes theorem, are considered to be independent, and the contribution of each feature to the output is presumed to be equal. The Gaussian Naive Bayes assumes that continuous values



(a) Frequency Before Oversampling

(b) Frequency After Oversampling

Fig. 1: Value Counts Before and After SMOTE Oversampling Technique

corresponding to the features are distributed as Gaussian or Normal distribution.

7) SGD Classifier: Gradient Descent is an optimization process where the goal is minimizing the cost function. In Stochastic Gradient Descent (SGD), instead of using the whole observations, the gradient is calculated using a random little part of those observations. By doing this, the computational time can be reduced.

8) Gradient Boosting Classifier: Gradient Boosting Classifier is based on boosting technique where weak learners like decision trees are ensembled so that it can learn from them iteratively, resulting in building a strong predictive model that provides enhanced prediction accuracy. It is used for minimizing the bias error of the model. For predicting categorical target variables, the log loss function is used as cost function.

9) Light Gradient Boosted Machine Classifier: Light GBM is an enhancement of the Gradient Boosting technique that includes a sort of automated feature extraction where the boosting examples with larger gradients are focused upon. As it is based on the concept of decision tree algorithms, the tree is split leaf node wise, resulting in reducing more loss compared to depth or level wise algorithms. It is highly efficient with its faster training speed and better accuracy.

10) XGBoost Classifier: XGBoost classifier or eXtreme Gradient Boosting is an upgraded version of gradient boosting decision tree algorithm, that can perform more accurately in less amount of time. It is an ensemble learning method, which bases the final result on the findings of numerous models.

B. Evaluation Metrics

Evaluation criteria such as Accuracy, Precision, Recall, and F1 score were measured to assess how well the various classifiers performed. Accuracy score is determined by dividing the correctly predicted sample by total number of input samples. It should be used for the data which has no imbalance in it, otherwise the score could be misleading. The proportion of true positives that were successfully identified is shown by recall. Precision denotes the proportion of correct outcomes that were predicted positive. Because calculating merely the accuracy does not guarantee the right evaluation of a particular model, the F1 score, which is the combined average of the precision and recall values, was also calculated.

IV. Results

A total of 10 models were used for the dataset namely Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Support Vector Classification, k-Neighbors Classifier, Gaussian Naive Bayes, SGD Classifier, Gradient Boosting Classifier, Light Gradient Boosted Machine Classifier and XGBoost Classifier. From table-1, we can observe that the Logistic Regression model achieved 89% Precision, 89% Recall and F1 score of 89% for Normal class and 89% Precision, 89% Recall and F1 score of 89% for the Lumpy class. For the Decision Tree Classifier, the Precision, Recall and F1 scores of the Normal class are 97%, 96% and 97% whereas for the Lumpy class those are 96%, 97% and 97% respectively. Random Forest Classifier shows an increase in all the metrics for both Normal and Lumpy class. This model showed a Precision of 98% along with 98% Recall and 98% F1 score for Normal class and percentages for the Lumpy class. Next, Support Vector Classification was applied. For the Normal class, the Precision was 95%, Recall was 94% and the F1 score for this model was 94%. For the Lumpy class, the precision was 94%, the recall was 95% and the F1 score was 94%. The next model was k-Neighbors Classifier with 0.99% precision 0.95% recall and 0.97% f1-score for the normal class and 0.96% precision 0.99% recall and 0.97% f1-score for the Lumpy class.

The 6th model was Gaussian Naive Bayes with 87% precision, 90% recall and 89% for Normal class and 90% precision, 87% recall and 88% for Lumpy class. The next one used by the authors was SGD Classifier. For the Normal class, the Precision was 89%, Recall and F1 score both were 90% for this model. For the Lumpy class, the precision was 90%, the recall and F1 score both were 89%. Gradient

TABLE I: Accuracy, Precision, Recall and F1-score of Normal and Lumpy Class for the 10 Different Models

	Normal Class				Lumpy Class			
Model Name	Accuracy Score	Precision	Recall	F1 Score	Accuracy Score	Precision	Recall	F1 Score
Logistic Regression	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
Decision Tree Classifier	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97
Random Forest Classifier	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**
Support Vector Classification	0.94	0.95	0.94	0.94	0.94	0.94	0.95	0.94
k-Neighbors Classifier	0.97	0.99	0.95	0.97	0.97	0.96	0.99	0.97
Gaussian Naive Bayes	0.88	0.87	0.90	0.89	0.88	0.90	0.87	0.88
SGD Classifier	0.90	0.89	0.90	0.90	0.90	0.90	0.89	0.89
Gradient Boosting Classifier	0.96	0.97	0.94	0.96	0.96	0.94	0.97	0.96
Light Gradient Boosted Machine Classifier	**0.98**	**0.98**	**0.97**	**0.98**	**0.98**	**0.97**	**0.98**	**0.98**
XGBoost Classifier	0.95	0.96	0.94	0.95	0.95	0.94	0.96	0.95

Boosting Classifiers showed Precision, Recall and F1 scores for the Normal class 97%, 94% and 96% respectively whereas for the Lumpy class the scores were 94%, 97% and 96%. Light Gradient Boosted Machine Classifier was the second best model applied to the dataset. For Normal class, 98% Precision, 97% Recall and 98% F1 score were achieved and for the Lumpy class, the Precision, Recall and F1 score were 97%, 98% and 98% respectively. The last model applied was XGBoost Classifier with 96% Precision, 94% Recall and 95% F1 score for the Normal class and 94% Precision, 96% Recall and 95% F1 score for the Lumpy class.

TABLE II: Performance Metrics Comparison Between Prior Work [5] and Proposed Work

	Lumpy Class			
Model Name	Accuracy Score	Precision	Recall	F1 Score
Artificial Neural Networks (prior work)	0.96	0.88	1	0.94
Random Forest Classifier (proposed work)	**0.98**	**0.98**	**0.98**	**0.98**
Light Gradient Boosted Machine Classifier (proposed work)	**0.98**	**0.97**	**0.98**	**0.98**

V. Analysis & Discussion

The outcomes of our experiments suggest that LSDV can be predicted with great accuracy in previously unreported test data-sets using machine learning approaches that use climatic and geospatial information as predictive variables. Random Forest Classifier and Light Gradient Boosted Machine Classifier, for example, both reported 98% accuracy. However, where one of the several classes is more frequent than others such as our data-set, accuracy cannot be only reliable metric [10]. As a result, when evaluating the predictions of machine learning algorithms, other metrics like F1 score, precision as well as recall should also be given equal importance. When all predictive variables were considered and performance criteria other than accuracy were used, the maximum performance was related with the Random Forest Classifier and the Light Gradient Boosted Machine Classifier (Table 1).

The explanation for the improved performance of the Random Forest Classifier might be connected to the fact that various nominally correlated models yield substantially better performance in the Random Forest Classifier which can classify with high degree of accuracy. Furthermore, the Light Gradient Boosted Machine Classifier generates the same performance since it is based on the notion of decision tree methods. It eliminates more loss than other level-based decision tree algorithms, demonstrating that it is very efficient in training and has higher accuracy. In this situation, however, it returned the same results as the Random Forest Classifier.

Another key reason for the enhanced performance compared to prior works [5] on this dataset [8] is that in this work (table 2), we made the dataset balanced. In prior research [5], machine learning models were applied to this dataset [8] with more than 87% of the data categorized as "regular class" and the rest as "lumpy class". In the current work, all the classifiers performed better after oversampling the data set using the SMOTE approach and standardizing with standard scaler.

VI. Conclusion & Future Works

In conclusion, machine learning techniques such as the Random Forest Classifier and the LGBM Classifier have the ability to reliably forecast the occurrence of the Lumpy skin disease virus using climatic and geospatial parameters. Using this method in locations where LSDV infection is a high risk, monitoring and awareness programs, as well as preventive measures such as vaccine campaigns, could be immensely beneficial.

The lack of sufficient data and predictor variables is one of this study's significant limitations. Initially, there was just a limited amount of data classed as "lumpy class". Furthermore, there is a chance that regions with various geolocations and clamatic conditions are LSDV predictors that the researchers are not aware of. In the future efforts of this work, it will be a must to overcome these limitations.

References

[1] Weiss, K. E. (1968). Lumpy skin disease virus. In Cytomegaloviruses. Rinderpest Virus. Lumpy Skin Disease Virus (pp. 111-131). Springer, Berlin, Heidelberg.

[2] Coetzer, J. A. W., & Tuppurainen, E. (2004). Lumpy skin disease. Infectious diseases of livestock, 2, 1268-1276.

[3] Department of Jobs, Precincts and Regions. (2022, March 8). Lumpy skin disease. Agriculture Victoria. https://agriculture.vic.gov.au/biosecurity/animal-diseases/cattle-diseases/lumpy-skin-disease

[4] Sprygin, A., Artyuchova, E., Babin, Y., Prutnikov, P., Kostrova, E., Byadovskaya, O., & Kononov, A. (2018). Epidemiological characterization of lumpy skin disease outbreaks in Russia in 2016. Transboundary and emerging diseases, 65(6), 1514-1521. https://doi.org/10.1111/tbed.12889

[5] Afshari Safavi, E. Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features. Trop Anim Health Prod 54, 55 (2022). https://doi.org/10.1007/s11250-022-03073-2

[6] Liang, R., Lu, Y., Qu, X., Su, Q., Li, C., Xia, S., . . . Chen, Q. (2020). Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data. Transboundary and emerging diseases, 67(2), 935-946.https://doi.org/10.1111/tbed.13424

[7] Golden, C. E., Rothrock Jr, M. J., & Mishra, A. (2019). Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms. Food research international, 122, 47-55.

[8] Afshari Safavi, Ehsanallah (2021), "Lumpy Skin disease dataset", Mendeley Data, V1, doi: 10.17632/7pyhbzb2n9.1

[9] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

[10] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems: O'Reilly Media.

[11] Latham, J., Cumani, R., Rosati, I., & Bloise, M. (2014). Global land cover share (GLC-SHARE) database beta-release version 1.0– 2014. FAO: Rome, Italy.

[12] Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Scientifc data, 7(1), 1-18. https://doi.org/10.6084/m9. fgshare.11980500

[13] Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T. P., Vanwam- beke, S. O., Wint, G. W., & Robinson, T. P. (2018). Global distribution data for cattle, bufaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Scientifc data, 5(1), 1-11. https:// doi.org/10.1038/sdata.2018.227