Machine Learning Based Diagnosis of Lumpy Skin Disease

R Geetha[1*],T KesavaRao[2],M.E. Palanivel[3].

Sreenivasa Institute of Technology and Management Studies,Chittoor, Andhra Pradesh 517127, India.
[*]Co responding author mail: geetharctr@gmail.com

International Journal of Emerging Technologies and Innovative Research(JETIR)

***Abstract*—Lumpy skin disease is a transmissible virus contracted by cattle that has led to concern among the nations. It has a direct relation with climate as the latter plays a major role in studying the infection and the pattern of transmission followedby it. This study depicts how the various climatic factors help in determining whether the cattle in the specific region or a country has the lumpy skin disease or not by using machine learning algorithms. Machine learning algorithms employed in the present study predicted lumpy disease with accuracy and F1 score of 100% and 1.0, respectively. In the present study, four different machine learning algorithms: Adaboost, K-nearest neighbors, decision tree and random forest are employed. The present research suggests that the decision trees can be used to predict lumpy skin disease infection using the geospatial and climatic parameters. The predicting power of machine learning algorithms can help in monitoring the disease spread patterns. It will also help in the application of vaccine campaigns in regions where the spread of disease poses a great risk to health.**

Keywords*— *machine learning, lumpy skin disease, geospatial information, decision tree, random forest

I. INTRODUCTION

Lumpy skin disease or LSD is a transmissible disease [1] caused by the Neethling virus of the Poxviridae family. LSD [2] is characterized by nodules found on the skin along with fever, mucous membranes, enlarged lymph nodes, emaciation and skin oedema. Cattle are mainly prone to the virus, but the virus has been recorded in giraffes, impalas etc. Earlier, the disease was just considered as an allergic reaction or hypersensitivity to insect bites. But later, millions of cattle were found to be suffering with an unknown disease with such symptoms [1, 3]. As of now, the disease is prevalent in various European and Asian countries like Bulgaria, Greece, Russia, Bangladesh, India, Pakistan, Thailand etc. This endemic has caused significant consequences due to drop in the animal milk production, skin problems, infertility issues and the mortality of infected animals. Also there is a huge cost linked with the treatment and vaccination. The occurrence of intermittent LSD is associated with high temperatures and humidity, as the transmission of the virus is due to arthropod vector species, especially the blood feeding insects which exist near the water bodies. The outbreak of disease depends upon the movement of animals, their immune system and various climatic factors. The disease can spread through nasal discharge, saliva and infected milk to suckling calves. The disease is diagnosed using the RT-PCR test. In recent years, there has been tremendous growth in geospatial data and in technologies like big data, data mining etc. The intelligent data analysis techniques using geospatial and climatic data require computer aided diagnosis of disease [4] or machine learning [5]. The machine learning algorithms [6,7] look for similarities in the data to create a pattern that is convenient to study. Machine learning techniques have replaced traditional methods especially in big data analyses as it yields better results. The ML technique [8, 9] has been applied in various studies for predicting the incident oftransmissible diseases in humans, animals or plants. In [10], the authors developed models to predict the incidence of Coffea arabica pests and diseases using climatic factors as predictive features. The authors used various ML algorithms and concluded that the Random Forest Tree algorithm is more accurate in the predictions as compared to Multiple linear regression, K Neighbors Regressor, and Artificial Neural Networks. In [11], the authors created a ML model for malaria classification using climatic factors. The authors feature integrated the dataset to improve the performance of different proposed algorithms. The XG Boost algorithm outperformed the other comparative algorithms and K-means clustering with an average accuracy of 0.823 and 0.958 respectively. In [12], the authors used 4 different ML algorithms to predict the dengue cases. The support vector regression model outperformed in predicting the trends of dengue in five different provinces of China and attained R-squared values of 0.985 for Zhejiang province. In [13], the concept of deep learning was used to develop a model for diagnosis of COVID using image processing . In the present research, the key objective is to study the reliance of the LSD virus on the climatic and geographical factors. The study aims at the development of ML algorithms based on geospatial and meteorological features to identify the occurrence of LSDV infection in different countries having infection history. Thedata set for present research is obtained from Mendeley [14]. In the study, the research work done is as follows:

1. Lumpy Skin Disease Dataset is pre-processed by converting the values information contained in the dataset into the categorical information.
2. Four different machine learning models are employed for disease classification.
3. The applied algorithms are compared using different evaluation metrics.

The present work is organized as follows: Section 1 introduces the lumpy skin disease, its history and transmission patterns. Section 2 describes the proposed methodology used in the study. Section 3 describes the features of the dataset involved along with the various machine learning techniques applied. Section 4 includes the results of the study and the corresponding discussions. Section 5 explains the future scope of this study

II. PROPOSED METHODOLOGY

In the spread of a virus, the climatic conditions of the affected areas play the utmost importance. The spread of LSD also has a direct relation with climate as the later plays a major role in studying the infection and transmission pattern of the virus. This study aims at finding a diagnosis model for lumpy skin disease using the contribution of geographical and meteorological factors in its transmission using a machine learning algorithm. This study depicts how the various climatic factors help in determining whether the cattle in that specific region or a country has the lumpy skin disease or not by using machine learning algorithms like KNN, Random Forest Curve and Decision Tree Curve. The Fig. 1 depicts the flow chart for the proposed research methodology for the present research.

The dataset for the present research is collected from Mendeley dataset repository [14]. The dataset is available in CSV file format. The dataset is preprocessed to form the correlation among the various attributes and lumpy skin disease. The desired attributes from the dataset are further selected. After the parameter selection, a sub dataset consisting of the selected attributes is created. The four different machine learning models: KNN, Random forest, Decision Tree and Adaboost are utilized in the present research for the classification task. The hyper-parameters of the model are tuned for optimal performance. All the four designed machine learning models are tested on testing dataset. Further, the machine learning models performance is analyzed using various performance metrics. The performance of utilized models is also compared with the models already utilized in the previous research.

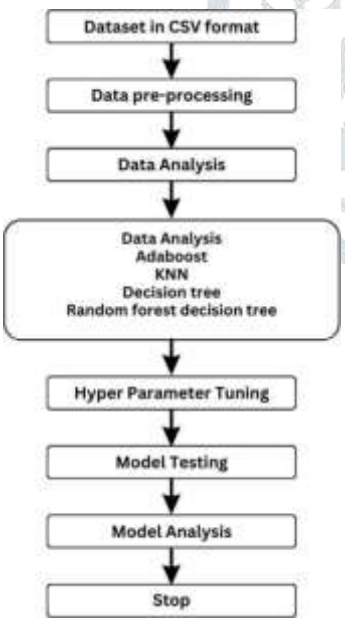

Fig. 1. Proposed Research Methodology.

III. MATERIAL AND METHODS

A. Dataset Description and pre-processing

The dataset used to conduct this research is acquired from Mendeley data repository [15]. It contains the data recorded by the countries reporting LSDV infection from the period 2011–2021.The dataset contains 24803 rows and 20 columns. The dataset columns consist of the information regarding the name of the region and its co-ordinates; meteorological data of the region; geospatial data, land cover; elevation; animal density; lumpy classification information etc. The information contained in different columns of the dataset can be divided into five different categories: meteorological data; spatial information; dominant land cover; animal density; lumpy disease classification information. The dataset includes information in 5 different categories: Meteorological data, spatial data, dominant land cover, animal density and LSDV information.

The dataset is pre-processed to make it suitable for the ML algorithm. The one-hot encoding technique is used to convert variables into numeric values. It makes utilization of data values by machine learning algorithms efficiently. The dataset is further split into two categories: train and test sets using the command from scikit-learn library. The model is developed or trained using training set and the validation and testing of the model is performed using test set. The Train:Test split of 7:3 is used in the present research.

B. Evaluation Metrics

The evaluation parameters [15, 16] used to compare the performance of ML models is based on confusion matrix which represents the category of the class values identified by ML model. Fig. 2 shows the confusion matrix.

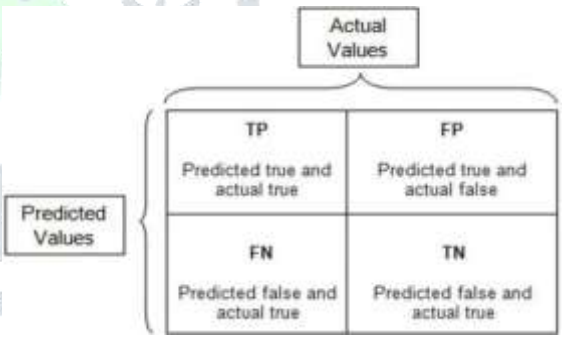

Fig. 2. Confusion Matrix.

The parameters used for model evaluation include: accuracy and F-1 score as expressed in equation 1 and 2 respectively.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$\textbf{F1 Score} = 2X\frac{\Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \quad (2)$$

Here, Precision is defined in equation 3 and Recall in equation 4,

$$\textbf{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \quad (4)$$

To calculate the accuracy of a model, accuracy_score function is imported from sklearn.metrics library into the program. The equation 1 shows the formula to calculate accuracy. f1_score function is imported from sklearn.metrics library into the program. The equation 2 shows the formula to calculate F-1 score.

C. Machine Learning Models

The present research used four different machine learning algorithm for lumpy skin disease classification:

K-Nearest Neighbor: KNN algorithm [17] is one of the simplest supervised machine learning approaches used for classification. It computes the distance between the unknown data point and the reference data point for the data records that need to be classified. Then the k nearest neighbors in the reference data are observed and a result is concluded depending on the majority class in these k data records. The steps included in the algorithm are:

Step 1: k value selection (K can be any positive integer).
Step 2: Euclidean distance calculation.
Step 3: Sorting of values based on the Euclidean distance.
Step 4: Assigning a new class to test data.

Decision Tree: A decision tree [18] classifier has tree-like structure whose main aim is to predict the data class by utilizing decision rules deduced from the training data. It is also a supervised machine learning technique. It is used for both classification and regression. Decision tree is based on tree data structure in which internal nodes represent data features, internal branches represent decision rules and each node represents the result of the decision tree. Decision tree is based on if-else conditioning and further splits the trees into sub-trees. The steps included in the algorithm are:

Step 1: Selection of best attributes or features.
Step 2: Decision tree nodes generation.
Step 3: Data classification.

Random Forest: Random Forest [19] algorithm deals with regression and classification problems. It constructs decision trees for various samples and ensembles them for better accuracy as there is low correlation between the individual models. The algorithm generates the decision trees on the data sample to get a prediction for each data. Finally, the best class is selected based on voting. The steps included in the algorithm are:

Step 1: Selection of 'n' number of datasets from 'k' number of random records.
Step 2: Creation of a decision tree.
Step 3: Output generation.
Step 4: Voting based final output.

Adaboost: It [20] is a performance boosting algorithm which combines multiple weak classifiers into a single strong classifier by reassigning weights to each case, with larger weights assigned to incorrectly classified cases. A single classifier is unable to predict a class accurately but when a combination of weak classifiers is used, it creates a strong classifying model. These weak classifiers could be decision trees, logistic regression, etc. Adaboost works with the help of decision stumps. It uses a forest of such stumps rather than trees. The steps included in the algorithm are:

Step 1: Equally weighted samples (first stumps)
Step 2: Creation of decision stump.
Step 3: Classification.

IV. RESULTS AND DISCUSSIONS

The performance of various machine learning models applied in the present research as well as models already existing in the literature are discussed in this section.

.

A. Machine Learning Models Performance

The performance of various machine learning algorithms implemented on the dataset has been compared on the basis of three evaluation metrics: accuracy score, F1 score and time taken. The Table I shows the comparison of the values of these metrics for the machine learning algorithms employed in this study.

TABLE I. EVALUATION OF MACHINE LEARNING ALGORITHMS

S. No.	ML Algorithm	Accuracy score (%)	F1 score	Time taken (sec.)
1	KNN	99.62	0.984	0.00472
2	Random forest	100	1.0	0.28109
3	Decision Tree	100	1.0	0.02026
4	Adaboost	100	1.0	0.02525

The values in Table I shows that KNN algorithm lags behind other algorithms in terms of accuracy score and F1 score. The Random forest algorithm takes significant more time to fit the classifier to the training set than all other algorithms. The Decision Tree algorithm performs best

among all the algorithms in terms of time to fit the classifier to the training set, F1 score and accuracy score. The Fig. 3 shows the confusion matrix for the various machine learning algorithms used in this study. From the confusion matrix, it can be concluded that the KNN algorithm generates the false positive and false negative cases while this value is zero for all the others.

From the study, it can be concluded that geographical and meteorological attributes of the data can be used in predicting whether an animal is suffering from lumpy skin disease via machine learning algorithms with a near perfect accuracy. Random Forest and Decision tree had an accuracy of 100%. While the accuracy can be preferred in case of a balanced dataset, it could not be said so for imbalanced data.

Fig. 3. Confusion Matrix (a) KNN algorithm (b) Random Forest algorithm (c) Decision Tree Algorithm (d) Adaboost Algorithm.

To assess the prediction capability of ML algorithms, the performance metrics like F1 score and AUC are used. The Region operating characteristic curves [21] generated for the different ML algorithms are given in Fig. 4. If the AUC value is '1' for a particular model, then the classification model performs perfectly. However if the AUC of the curve is '0' for a particular model then the model has poor classification performance. If the AUC value is in the range of 0.5 and 1, then there are more chances for the model to be predicted accurately. The ROC curve of KNN has AUC slightly less than 1. However the AUC of other machine learning models is 1.

B. Comparative Analysis of used ML Models with Models existing in literature

The Table II shows the difference in accuracy and F1 scores of the machine learning algorithms applied in the present model and the recent model published. In comparison to the present model, with maximum accuracy of 100% and f1 score 1.0, the published model procured the highest accuracy of 97% only in case of ANN and f1 score 0.94. This comparison indicates the models employed in present study are more suitable and efficient in LSD diagnosis.

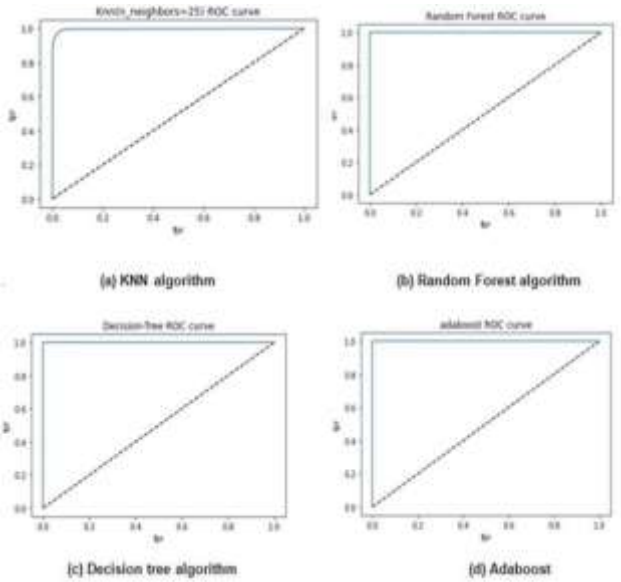

Fig. 4. ROC curves (a)KNN algorithm (b) Random Forest Algorithm (c) Decision Tree Algorithm (d) Adaboost Algorithm.

TABLE II. COMPARATIVE ANALYSIS WITH EXISTING WORK.

S.No.	Study	Algorithm used	Accuracy score (%)	F1 score
1	Present study	KNN	99.62	0.984
2	Present study	Random Forest	100	1.00
3	Present study	Decision Tree	100	1.00
4	Present study	Adaboost	100	1.00
5	[21]	Random Forest	93-96	0.69-0.79
6	[21]	Decision Tree	89-90	0.43-0.46
7	[21]	Adaboost	88-91	0.13-0.42
8	[21]	ANN	96-97	0.94
9	[21]	Support vector machine	94-96	0.71-0.80
10	[21]	Logistic regression	92-93	0.55-0.60

11	[21]	Bagging	95-96	0.74-0.81
12	[21]	XGBoost	92-94	0.54-0.65

V. CONCLUSION AND FUTURE SCOPE

It is extremely important to understand the trends and find accurate ways of disease detection to take appropriate precautionary steps. Efforts are continuously being made in improving the existing algorithms, aiming for better results. Machine learning practices are seen as key components of future scientific decision-making and monitoring systems. With the rapid increase in the disease cases, it is important to get familiar with its transmission patterns in order to at least prevent the disease from causing more harm. In order to do that, the primary step should be its detection for which this study was conducted. In the present study, the machine learning algorithms are applied on the lumpy skin disease dataset collected from the Mendeley repository and the results compared to find the best suitable algorithm for the diagnosis of disease. This study depends on the relation of the disease to the meteorological and geographical factors like vapour pressure, humidity, average temperatures, elevation data etc. Taking these as predictive factors, a LSD diagnosis model was created. It was observed that KNN lacked accuracy as compared to the other algorithms which proved to be 100% accurate. Latest work done on this dataset concluded ANN supervised machine learning algorithm to be the preferred model for diagnosis of LSDV which only provided 97% accuracy. In comparison, our proposed model provided a better accuracy and f1 score of 100% and 1.0 respectively, indicating that our proposed model is more efficient than the one suggested in the previous study [22]. With advancement in scientific technology and a deeper understanding of machine learning, an improved version of the proposed model can be created that can be based on deep learning.

REFERENCES

[1] K. Al-Salihi, "Lumpy skin disease: Review of literature," *Mirror of research in veterinary sciences and animals*, vol.3, no. 3, pp.6-23 2014.

[2] J.L. Doré, and J.I. Quin, "A skin condition in cattle probably associated with photosensitization". *Journal of the South African Veterinary Association*, vol. 14, no.1, pp.10-11, 1943.

[3] B.C. Jansen, "The Onderstepoort Veterinary Research Institute in relation to animal diseases in Southern Africa". *Africa Insight*, vol. 2, no.1, pp.47-54, 1972.

[4] P. Malhotra, S. Gupta and D. Koundal, "Computer aided diagnosis of pneumonia from chest radiographs," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no.10, pp.4202-4213, 2019.

[5] Y. Casali, N.A. Yonca, T. Comes and Y. Casali, "Machine learning for spatial analyses in urban areas: a scoping review," *Sustainable Cities and Society*, vol. 85, pp.104050, 2022.

[6] M.I Jordan and T.M Mitchell, "Machine learning: Trends, perspectives, and prospects" *Science*, vol. 349, no. 6245, pp.255-260, 2015.

[7] T. Qin, "Machine Learning Basics," in Dual Learning, Springer, Singapore, 2020, pp.11-23

[8] R. Gupta, A. Sharma, S. Gupta, M. Garg and G. Kaur, "Automatic Identification of Paddy Crop Diseases using Deep Learning Approach," in *Proc. 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2022, pp. 915-920.

[9] S. Singh, I. Gupta, S. Gupta, D. Koundal, S. Aljahdali, S. Mahajan and A.K Pandit, "Deep learning based automated detection of diseases from apple leaf images," *CMC-Computers,Materials & Continua*, vol. 71, no.1, pp.1849-1866, 2022.

[10] L.E. de Oliveira Aparecido, G. de Souza Rolim, J. da Silva Cabral De Moraes, C.T.S Costa and P.S de Souza, "Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases". *International Journal of Biometeorology*, vol.64, no.4, pp.671-688, 2020.

[11] O. Nkiruka, R. Prasad and O. Clement, "Prediction of malaria incidence using climate variability and machine learning" *Informatics in Medicine Unlocked*, vol. 22, pp.100508, 2021.

[12] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang and W. Ma, "Developing a dengue forecast model using machine learning: A case study in China" *PLoS neglected tropical diseases*, vol. 11, no.10, pp.5973, 2021.

[13] P.Kaur et al., A "Hybrid Convolutional Neural Network Model for Diagnosis of COVID-19 Using Chest X-ray Images," *Int. J. Environ. Res. Public Health*, vol.18, no.22, pp.12191, 2021.

[14] https://data.mendeley.com/datasets/7pyhbzb2n9/1

[15] J. Zhou, A.H. Gandomi, F. Chen and A. Holzinger,. "Evaluating the quality of machine learning explanations: A survey on methods and metrics" *Electronics*, vol.10, no.5, p.593, 2021.

[16] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle,"Deep neural networks for medical image segmentation," *Journal of Healthcare Engineering*, vol. 2022, pp. 2040-2295,2022.

[17] Z. Yong, L. Youwen, and X. Shixiong, "An improved KNN text classification algorithm based on clustering" *Journal of computers*, vol. 4, no.3, pp.230-237, 2009.

[18] J.R. Quinlan, "Learning decision tree classifiers" *ACM Computing Surveys (CSUR)*, vol.28, no.1, pp.71-72, 1996.

[19] A. Liaw and M. Wiener "Classification and regression by random Forest. *R news*, vol.2, no.3, pp.18-22, 2002.

[20] W. Fan, S.J. Stolfo and J. Zhang, "The application of AdaBoost for distributed, scalable and on-line learning" in *Proc. fifth ACM SIGKDD international conference on Knowledge discovery and data mining*,1999, pp. 362-366.

[21] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms" *Pattern recognition*, vol.30, no.7, pp.1145-1159, 1997.

[22] E. Afshari Safavi, "Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on meteorological and geospatial features," *Tropical Animal Health and Production*, vol 54, no. 1, pp.55, 2022.