Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 4 (2022) 555 - 562 ISSN Media Elektronik: 2580-0760

Applying Different Resampling Strategies In Random Forest Algorithm To Predict Lumpy Skin Disease

Suparyati[1], Emma Utami[2], Alva Hendi Muhammad[3]
[1,2,3]Program Pasca Sarjana, Universitas Amikom Yogyakarta
[1]suparyati@students.amikom.ac.id, ema.u@amikom.ac.id, [3]alva@amikom.ac.id

Abstract

The spread of Lumpy Skin Disease (LSD) that infects livestock is increasingly widespread in various parts of the world. Early detection of the disease's spread is necessary so that the economic losses caused by LSD are not higher. The use of machine learning algorithms to predict the presence of a disease has been carried out, including in the field of animal health. The study aims to predict the presence of LSD in an area by utilizing the LSD dataset obtained from Mendeley Data. The number of lumpy infected cases is so low that it creates imbalanced data, posing a challenge in training machine learning models. Handling the unbalanced data is performed by sampling technique using the Random Under-sampling technique and Synthetic Minority Oversampling Technique (SMOTE). The Random Forest classification model was trained on sample data to predict cases of lumpy infection. The Random Forest classifier performs very well on both under-sampling and oversampling data. Measurement of performance metrics shows that SMOTE has a superior score of 1-2% compared to the use of Random Undersampling. Furthermore, Re-call rate, which is the metric we want to maximize in identifying lumpy cases, is superior when using SMOTE and has slightly better precision than Random Undersampling. This research only focuses on how to balance unbalanced data classes so that the optimization of the model has not been implemented, which creates opportunities for further research in the future.

Keywords: *Lumpy Skin Disease, Machine Learning, Oversampling, Random Forest, Random Undersampling*

1. Introduction

Lumpy Skin Disease (LSD) is a disease in cattle caused by a virus from the family Poxviridae, genus Capripox which is characterized by the appearance of nodules in the body of livestock, fever, decreased appetite, causing emaciation in livestock [1],[2]. The spread of LSD originated in Africa, the Middle East, and Asia and has become endemic in various countries [3]. LSD entered Indonesia in early 2022 in Riau Province [4]. The tropical climate accompanied by wet and humid areas in Indonesia is a catalyst that accelerates the development of mechanical vectors carrying LSD viruses, such as Aedes sp. This worrying condition has provided its own motivation for research to carry out early detection to prevent its wider spread [5].

In the field of computer science, machine learning has long been used to assist the classification process of various existing problems. It can also be used in the field of animal health such as the detection of LSD in cattle [6],[7], prediction of mastitis in cattle [8],[9], skin disease in cattle [10], respiratory disease in cattle [11], penyakit pascapersalinan sapi pe dairy cow postpartum disease [12], *African Swine Fever* [13], and lameness in goats [14]. Although it has great potential, the problem of uneven distribution of data is still a classic problem in the classification process [15]. This class imbalance usually occurs when the class distribution between the majority and minority classes is not the same [16]. Data on unbalanced classes can vary from mild to severe. The effect of high class imbalance can affect the overall classification accuracy because the model most likely predicts most of the data that belongs to the majority class. Such a model will yield biased results, and performance predictions for the minority class often have no impact on the model.

A general approach to overcome the problem of unbalanced data classes by using resampling techniques that can be either undersampling or oversampling [17]. Bagui stated that oversampling increases training time and undersampling reduces training time, if the data is very unbalanced then both oversampling and undersampling increase recall significantly but if the data is not too balanced then resampling will have little impact and with oversampling more minority data will

Accepted: 16-06-2022 | Received in revised: 08-08-2022 | Published: 22-08-2022

be detected [18]. Undersampling is a process of reducing the number of data in the majority class that can be done randomly. One technique that can be used is Random Undersampling, which in the process can result in the loss of some important information [15]. Handling class imbalance with underSampling to predict software defects has the highest AUC score (= 95.6%), maximum accuracy value (= 96.9%) and the ROC curve closest to the upper left corner [19].

In contrast, the oversampling technique uses a new sample added to the minority data class to balance the dataset. The effectiveness of applying the oversampling method to unbalanced data before the modeling stage shows that all oversampling methods help improve the overall performance of the classification model [20]. One of the oversampling techniques is the Synthetic Minority Oversampling Technique (SMOTE). This technique works by duplicating data and measuring the similarity between neighboring minority class samples, where each data will be reproduced based on the nearest neighbor line [21]. Previous research has shown that the application of SMOTE in detecting credit card fraud with Ensemble shows much better recall results than when not using SMOTE [22]. In another study, the SMOTE-based data point oversampling approach to solve the problem of credit card data imbalance in detecting financial fraud can significantly improve the ability to predict positive class [23]. SMOTE has also been used in the health sector, one of which is to balance data in predicting cervical cancer from various risk factors [24].

Furthermore, the utilization of the Random Forest (RF) classifier has been shown to be effective in high dimensional spaces and applied to unbalanced tasks [25]. Random forest is also the best model with the highest accuracy in predicting pharmacodynamic interactions [26]. Wang (2021) in his research states that when the distance between classes and the sample variance of the expanded data is closer to the original data, the random forest classification is the best in the experiment designed [27]. One of the uses of Random Forest in the Health sector is for a coronary heart disease prediction system, where the use of Random Forest shows 98% accuracy, 99% sensitivity and 95.8% precision [28]. The results of the Parkinson's disease prediction study imply that the Random Forest Classifier with SMOTE can produce a model with higher accuracy than the Bagging Classifier with SMOTE or the Boosting Classifier with SMOTE when analyzing unbalanced data [29]. In addition, the use of the SMOTE method on the credit dataset coupled with the use of the Random Forest model can increase the predictive value so that the results are more accurate [30]. Random forest with SMOTE is also the best model in the classification of HB vaccination status where the accuracy of the identification of non-vaccinated Hepatitis-B status increases by 30.08% [31]. In another study, Kishor (2021) stated that the initial prediction of diabetics using SMOTE to balance data and the use of Random Forest could achieve maximum accuracy (97.81%), sensitivity (99.32%), specificity (98.86%), and AUC (99.35%) [32]. In research on Cervical Cancer Prediction using Outlier deduction and Over sampling methods, it is concluded that Random Forest is the best among several popular machine learning classifiers [33]. From the background and literature described above, in this study a sampling method will be used to help balance the LSD dataset. Furthermore, the Random Forest classification method will be used to predict the data class according to the previously resampled dataset. This paper consists of four chapters, after which the research method will be described followed by a discussion of research results. At the end, it will be closed with the conclusion of this research.

2. Research Methods

This research has a workflow that starts from data collection, data processing, training data sharing and validation data, data resampling, data modeling and performance evaluation of the model. The flow chart of the research carried out is shown in Figure 1.

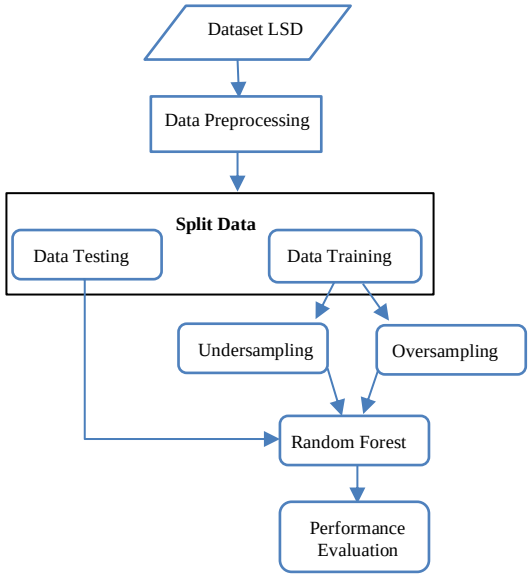


Figure 1: Research Flow

2.1. Data retrieval

Lumpy Skin Disease (LSD) dataset derived from Mendeley data will be used in this study. The dataset consists of 21,803 data which is divided into two data classes, namely lumpy class (1) with 3,039 data and no lumpy class (0) with 21,764 data.

The percentage of lumpy class of 12.25% and no lumpy class of 87.75% of the total data shows that the LSD dataset that will be used for modeling is not balanced. An illustration of the comparison of the lumpy class and the no lumpy class can be seen in Figure 2.

DOI: https://doi.org/10.29207/resti.v6i4.4147

Figure 2. Class Distribution of the LSD Dataset

The LSD dataset has twenty data attributes which are meteorological and geospatial features which can be described in Table 1 below [6].

2.2. Data Preprocessing

Before the dataset is used in classification modeling, it must first process the data. The first step is to delete five attributes that are not used in the modeling, namely 'region', 'country', 'reportingDate', 'X5_Ct_2010_Da', and 'X5_Bf_2010_Da'. Then the distribution of training data and validation data is carried out using*train_test_split*from*scikit learn library*with a ratio of 80:20. Before resampling the dataset, it is necessary to separate the original dataframe for testing purposes, so that model testing can be carried out on the original test set, not on the test set created by the resampling technique. The main goal is to fit the model into undersample and oversample dataframes so that the model detects patterns and tests them on the original test set.

2.3. Random Forest Classifier

The algorithm that will be used in predicting modeling is Random Forest which is a bagging ensemble of decision tree classifiers, where each tree chooses a class and the majority classification is taken in the end [34]. Bagging is a technique where each training sample is taken and replaced, so that on average each tree has 2/3 unique examples and 1/3 duplicates to reduce variance in the model [35]. Random Forest uses $\sqrt{p}$ random features for each tree (is the total number of features), which means the model is less prone to overfitting if some features are missing from the trees.

Information retrieval determines the most effective features to use when splitting the decision tree nodes. Metrics measure the entropy decrease by using certain features [36]. Given a classification c a feature vector x with components x_k, the information obtained (IG) by entering the feature x_k is defined as equation (1).

$$IG(c|x_k) = H\,(c) - H\,(c|x_k) \qquad (1)$$

$$H(x) = -\sum_i^{nc} p_{ci} log_2(p_{ci}) \qquad (2)$$

In equation (2) it can be explained that H that is the entropy of the data set with class nc, p_{ci} , is the probability of choosing that class. The greater the value of information acquisition, the more effective the feature in classifying data. The decision tree is divided into nodes based on which feature yields the most information gain.

Table 1. LSD Dataset Attributes

Attribute		Information
cld	:	Monthly Cloud Cover in percent
dtr	:	Diurnal Temperature Range in degrees Celsius
frs	:	frost day frequency in a month
pet	:	potential evapotranspiration in millimeters per day
pre	:	Precipitation is any product of the condensation of water vapor in the atmosphere in millimeters per month
tmn	:	daily mean temperature in degrees Celsius
tmp	:	Temperature in degrees Celsius
tmx	:	monthly average maximum and minimum temperature in degrees Celsius
vape	:	vapor pressureis in hectopascals
wet	:	wet day frequency in days
x	:	latitude x axis spatial coordinates
y	:	longitude y-axis spatial coordinate
Region	:	the continent of the outbreak
country	:	country of outbreak
Reporting date	:	reporting date of outbreak
elevation	:	altitude of geographic location in meters
dominant_ landcover	:	dominant land cover
X5_Ct_ 2010_Da	:	quick view GIS file of dasymetric cattle
X5_Bf_ 2010_Da	:	quick view of GIS files from dasymetric buffalo
lumpy	:	classification of whether infected with LSD code: 1, not infected with code: 0

2.4. Model Performance Evaluation

Evaluation of model performance will be carried out using precision values, recall, F1-Score and ROC AUC . curve [37]. The accuracy value is calculated by dividing the number of correct predictions by the total number of predictions. While the precision value is the accuracy of positive predictions as shown in equation (3)

$$Precision = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \qquad (3)$$

While recall itself is a positive instance that is correctly detected by the classifier shown in equation (4).

DOI: https://doi.org/10.29207/resti.v6i4.4147

$$Recall = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad (4)$$

In equation (5), F1-score is an evaluation metric that combines precision and recall numbers because the two values can have different weights. Thus, F1 is the harmonic average obtained from the results of precision and recall, the value ranges from 0 to 1.

$$F1 = 2\,X\,\frac{\text{precision x recall}}{precision+recall} \quad (5)$$

Receiver Operating Characteristics (ROC) is made based on the value obtained from calculations using a confusion matrix, namely between False Positive Rate and True Positive Rate where the classification performance is said to be good if it is close to the point (0,1). Area Under Curve (AUC) is the area under the ROC curve which is integral to the ROC function [26].

3. Results and Discussions

The LSD dataset used in this study requires handling the unbalance of data classes first to ensure that the prediction model produced is unbiased. This can be overcome by applying the resampling method to the LSD dataset. The first method used in this study is to undersampling the data, which aims to have the same number of classes, both for the lumpy class and the no lumpy class. The new class distribution after the random undersampling process can be seen in Figure 3 and Figure 4.

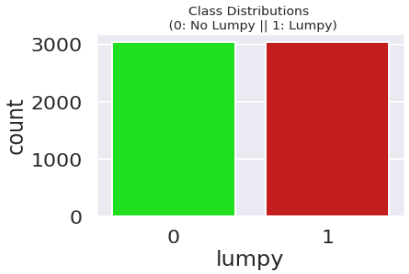


Figure 3: Class distribution after the undersampling process

In Figure 3, it can be seen that the distribution of lumpy and no lumpy classes is currently balanced.

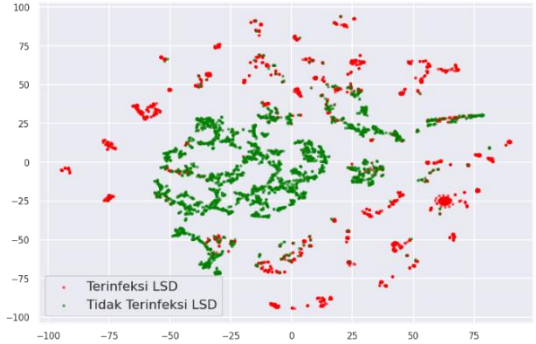
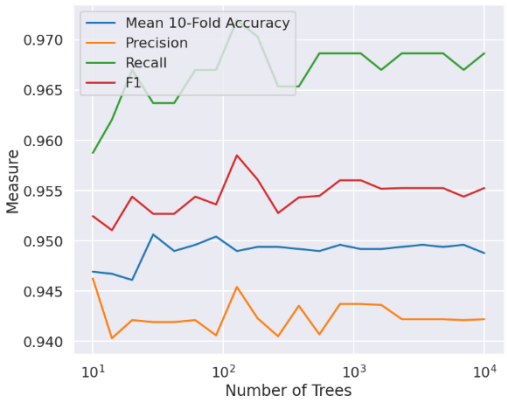


Figure 4: Visualization of data distribution after the undersampling process

Figure 4 shows the presence of certain distinct regions of points of the same class, in particular the lumpy cluster which contains most of the lumpy cases and1 the no lumpy cluster with some mixed lumpy samples. It can be said that the data classes are mostly separable and the classifiers done should perform well. However, there are some lumpy data that are closely mixed with no lumpy data, after the reduction of tSNE which may make it more difficult to classify these data.

After this, the model will be trained using a varying number of trees, then examine the performance metrics at each iteration and use them to determine the optimal number of trees to use in RF. This is done so that False Negative has the lowest possible value so that the lumpy case is not missed and can maximize the Recall value and F1 score, as visualized in Figure 5.



Gambar 5: Grafik metrik kinerja setiap iterasi

In Figure 5 it can be seen that all metrics did not increase or decrease consistently when adding more than approximately 1,000 trees to the Ensemble.

One of the most common metrics for evaluating how many trees to use in a model is Out of Bag Error, in which the mean predictive error rate in the sample is excluded from the sample tree bag. This metric can measure errors due to the varying number of trees in the ensemble [38].

The results in Figure 6 show that the number of trees will be optimal after reaching thousands, so in this study the number of trees was chosen to be 1,000. OOB Error indicates how well the model will generalize so use this as an indicator for the number of trees to use.

Furthermore, modeling is carried out using the Random Forest classification, where the final prediction results are shown using the confusion matrix in Figure 7.

Based on Figure 7 above, it can be concluded that undersampling Random Forest modeling has a recall of 96%, precision 94%, f1-score 94.9% and false negative 4%. This means that 4% of lumpy cases were incorrectly identified as no lumpy cases by the model.

DOI: https://doi.org/10.29207/resti.v6i4.4147

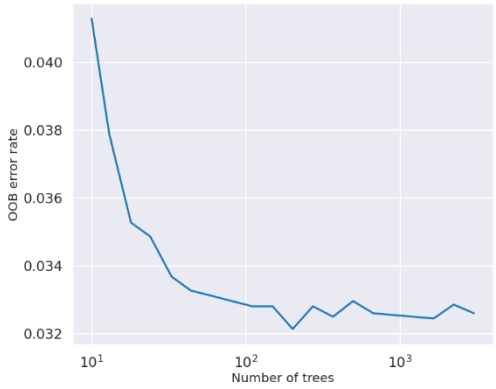


Figure 6: Out of Bag Error Rate Graph

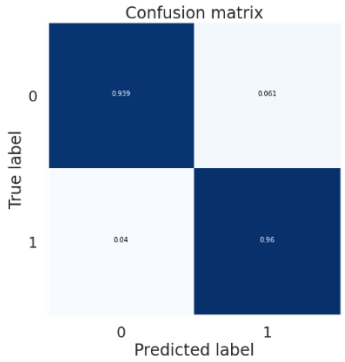


Figure 7: Confusion matrix of a model with undersample

The model uses a threshold of 50%, where if 50% or more trees choose one class, then that class will be the final classification. To increase the variation of the probability threshold output with the model in calculating the varying levels of True Positive and False Positive, the Receiver Operator Characteristic (ROC) curve is used [39].

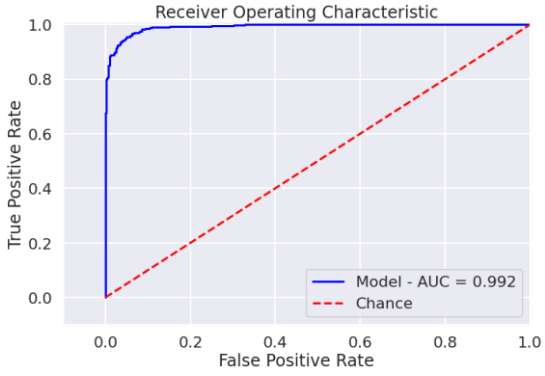


Figure 8: ROC curve of a model with undersample

On the ROC curve, the perfect model will be between (0,1) where the closer to this point, the more ideal the model's performance will be. The area under the ROC - AUC curve is another metric used to determine model performance; where the closer to 1 this number, the better the model. Based on this, in Figure 8 it can be seen that the ROC curve is close to 1. This indicates that the model's performance is ideal and the area under the ROC-AUC curve of 0.992 is a very good model.

In addition to using the ROC-AUC curve, the Precision and Recall plot can be used, where the perfect model will be if it is at (1,1) at the top right of the plot.

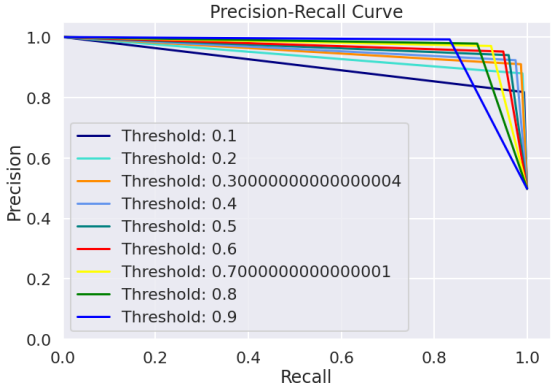


Figure 9: Precision-Recall curve of a model with an undersample

This visualization in Figure 9 shows that changing the threshold too low will make all data classified as lumpy. This means the model will identify all lumpy cases but misclassify no lumpy cases. If the set threshold is too high, the model will lose many lumpy cases.

The lower precision is most likely due to the lumpy case being close to the no lumpy case, in the sense that the Euclidean feature space which is the red dot mixed with the green cluster in the tSNE plot of Figure 4 above.

The 50% threshold can be changed to a lower value which will increase Recall, but it won't do so for now. Instead, in this study generate synthetic data using SMOTE by utilizing the imblearn library. After implementing SMOTE, the lumpy and no lumpy data classes are balanced.

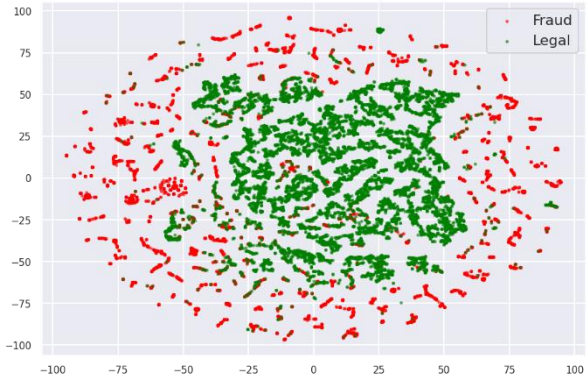


Figure 10: Visualization of data distribution after the SMOTE process

The tSNE plot in Figure 10 presents the data oversampled by SMOTE showing a very well-defined region, where the lumpy sample surrounds the no lumpy case sample.

DOI: https://doi.org/10.29207/resti.v6i4.4147

Classification of data to predict the existence of lumpy cases evaluated using a confusion matrix has the results as shown in Figure 11.

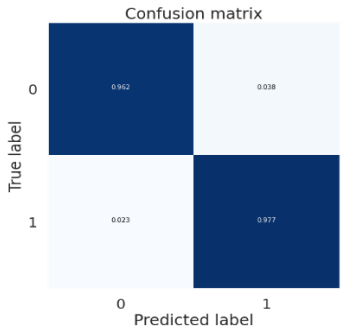


Figure 11: Confusion matrix of the model with SMOTE

Random Forest modeling with SMOTE has a recall of 97.7%, precision 96.2%, f1-score 96.9% and 2.3% false negative, which means 2.3% of lumpy cases were identified incorrectly as no lumpy cases by the model. as shown in Figure 11.

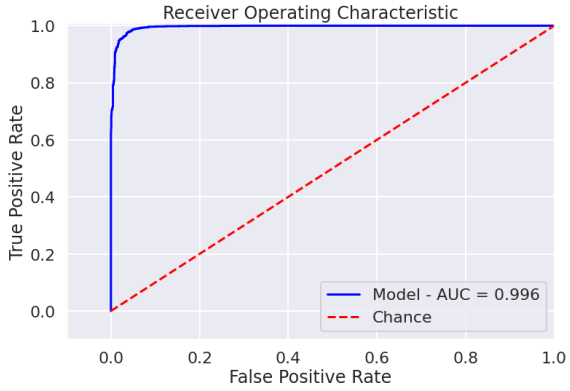


Figure 12: ROC curve of the model with SMOTE

The ROC curve in Figure 12 above is close to 1, so the performance of the model can be said to be ideal and the area under the ROC-AUC curve of 0.996 is a very good model. The results also show that all performance metrics are more improved, where AUROC, Recall, and Precision are all higher than the same metric from less sampled data.

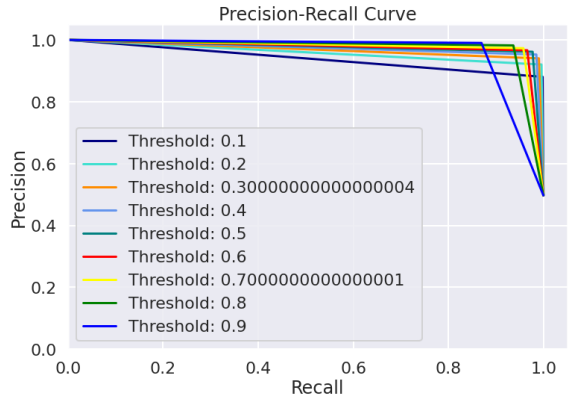


Figure 13: Precision-Recall curve of a model with an undersample

The Precision-Recall curve in Figure 13 above also shows that for all the same thresholds, the curve is much closer to (1,1) than the unsampled data.

From the experimental results above, it is known that the use of the SMOTE oversampling technique can minimize False Negatives with a difference of 1.7% compared to the use of the undersampling technique in classifying LSD. The AUC metric with the oversampling technique has increased by 0.4% from the undersampling technique. The increase in the scores of the recall, precision, f1-score and AUC metrics with SMOTE indicates that the oversampling technique is more suitable to be used to balance the data in this LSD classification.

4. Conclusion

Undersampling and oversampling techniques used to balance the data on the prediction of the presence of LSD are very useful to avoid classification of biased models. The use of tSNE to visualize the data shows that most of the lumpy and no lumpy cases are separable and distinct from each other. This allows any classifier to work properly. The Random Forest classifier performs very well on data undersampling, but has a superior score with the oversampling technique using SMOTE. All performance metrics scored higher between 1-2% using the SMOTE method for data resampling. Next, the Recall rate, which is the metric we want to maximize in identifying lumpy cases,

This research only focuses on how to balance unbalanced data classes, so that the optimization of the model has not been implemented which creates opportunities for further research in the future.

Reference

[1] I. Lojkic, "Complete Genome Sequence of a Lumpy Skin Disease Virus Strain Isolated from the Skin of a Vaccinated Animal," *Am. Soc. Microbiol.*, pp. 1–2, 2018, doi: https://doi.org/10.1128/genomeA .00482-18.

[2] R. Garcia, "'One Wefare': a framework to support the implementation of OIE animal wefare standards," *OIE*, pp. 3–13, 2017, doi: 10.20506/bull.2017.1.2589.

[3] A. Sprygin, Y. Pestova, D. B. Wallace, E. Tuppurainen, and A. V Kononov, "Transmission of lumpy skin disease virus : A short review," *Virus Res.*, vol. 269, no. May, p. 197637, 2019, doi: 10.1016/j.virusres.2019.05.015.

[4] K. Pertanian, "Kementan Siapkan Sumberdaya Tangani Lumpy Skin Disease Pada Sapi Di Riau," 2022. http://ditjenpkh.pertanian.go.id/kementan-siapkan-sumberdaya-tangani-lumpy-skin-disease-pada-sapi-di-riau (accessed Apr. 07, 2022).

[5] W. Molla, K. Frankena, G. Gari, M. Kidane, D. Shegu, and M. C. M. De Jong, "Seroprevalence and risk factors of lumpy skin disease in Ethiopia," *Prev. Vet. Med.*, 2018, doi: 10.1016/j.prevetmed.2018.09.029.

[6] E. A. Safavi, "Assessing machine learning techniques in forecasting lumpy skin disease occurrence based on

DOI: https://doi.org/10.29207/resti.v6i4.4147

meteorological and geospatial features," *Trop. Anim. Health Prod.*, pp. 1–11, 2022, doi: 10.1007/s11250-022-03073-2.

[7] G. Rai *et al.*, "A Deep Learning Approach to Detect Lumpy Skin Disease in Cows," *EasyChair Prepr.*, 2020.

[8] N. A. Ghafoor, "MasPA : A Machine Learning Application to Predict Risk of Mastitis in Cattle from AMS Sensor Data," *AgriEngineering*, vol. 3, pp. 575–583, 2021.

[9] R. M. Hyde *et al.*, "Automated prediction of mastitis infection patterns in dairy herds using machine learning," *Sci. Rep.*, vol. 10, no. 4289, pp. 1–8, 2020, doi: 10.1038/s41598-020-61126-8.

[10] G. Workee, "Cattle skin diseases identification model using machine learning approach," Bahir Dar University, 2021.

[11] H. A. Rojas, B. J. White, D. E. Amrine, and R. L. Larson, "Predicting Bovine Respiratory Disease Risk in Feedlot Cattle in the First 45 Days Post Arrival," *MDPI*, 2022.

[12] S. Van Der Beek, G. Layer, and G. Gmbh, "Prediction of postpartum diseases of dairy cattle using machine learning," in *Proceedings of the World Congress on Genetics Applied to Livestock Production*, 2018, no. 11.104.

[13] R. Liang *et al.*, "Prediction for global African swine fever outbreaks based on a combination of random forest algorithms and meteorological data," *Transbound. Emerg. Dis.*, vol. 67, no. September 2019, pp. 935–946, 2020, doi: 10.1111/tbed.13424.

[14] J. Kaler, J. Mitsch, J. A. Vázquez-diosdado, N. Bollard, T. Dottorini, and K. A. Ellis, "Automated detection of lameness in sheep using machine learning approaches : novel insights into behavioural differences among lame and non-lame sheep," *R. Soc. Open Sci.*, vol. 7, no. 190824, 2020, doi: http://dx.doi.org/10.1098/rsos.190824.

[15] M. S. Shelke, P. R. Deshmukh, and P. V. K. Shandilya, "A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique," *Int. J. Recent Trends Eng. Res. Res.*, vol. 3, no. 4, pp. 444–449, 2017.

[16] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Inf. Softw. Technol.*, vol. 139, no. August 2020, p. 106662, 2021, doi: 10.1016/j.infsof.2021.106662.

[17] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques : Overview Study and Experimental Results," *Int. Conf. Inf. Commun. Syst. Fig.*, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[18] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00390-x.

[19] S. Goyal, *Handling Class-Imbalance with KNN (Neighbourhood) Under-Sampling for Software Defect Prediction*, vol. 55, no. 3. Springer Netherlands, 2022.

[20] S. Demir and E. K. Şahin, "Evaluation of Oversampling Methods (OVER , SMOTE , and ROSE) in Classifying Soil Liquefaction Dataset based on SVM , RF , and Naïve Bayes SVM , RF ve Naive Bayes ' e Dayalı Olarak Zemin Sıvılaşma Veri Setinin Sınıflandırılmasında Aşırı Örnekleme Yönteml," no. 34, pp. 142–147, 2022, doi: 10.31590/ejosat.1077867.

[21] P. Wibowo and C. Fatichah, "An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," vol. 7, no. January, pp. 63–71, 2021.

[22] Y. Xiao and J. Lian, "Credit Card Fraud Detection using SMOTE and Ensemble Methods Credit Card Fraud Detection using SMOTE and Ensemble," *Int. J. Eng. Res. Sci.*, vol. 7, no. 8, 2021.

[23] N. Mqadi, N. Naicker, and T. Adeliyi, "A SMOTe based oversampling data-point approach to solving the credit card data imbalance problem in financial fraud detection," *Int. J. Comput. Digit. Syst.*, vol. 10, no. 1, pp. 277–286, 2021, doi: 10.12785/IJCDS/100128.

[24] C. H. Bhavani and A. Govardhan, "Materials Today : Proceedings Cervical cancer prediction using stacked ensemble algorithm with SMOTE and RFERF," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2021.07.269.

[25] B. Pes, "Learning from High-Dimensional and Class-Imbalanced Datasets Using Random Forests," *Information*, vol. 12, no. 286, 2021, doi: 10.3390/info12080286.

[26] N. A. Farhana, F. M. Afendi, A. Fitrianto, and S. H. Wijaya, "Classification modeling of support vector machine (SVM) and random forest in predicting pharmacodynamics interactions," *J. Phys. Conf. Ser. Pap.*, 2021, doi: 10.1088/1742-6596/1863/1/012067.

[27] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[28] A. Abdul, R. M. Isiaka, R. S. Babatunde, and J. F. Ajao, "An Improved Coronary Heart Disease Predictive System Using Random Forest," *Asian J. Res. Comput. Sci.*, vol. 11, no. 1, pp. 17–27, 2021, doi: 10.9734/AJRCOS/2021/v11i130253.

[29] H. Byeon, "Exploring Parkinson ' s Disease Predictors based on Basic Intelligence Quotient and Executive Intelligence Quotient," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 106–111, 2021.

[30] V. M. Putri, M. Masjkur, and C. Suhaeni, "Handling Problems of Credit Data for Imbalanced Classes using SMOTEXGBoost Handling Problems of Credit Data for Imbalanced Classes using SMOTEXGBoost," *J. Phys. Conf. Ser. Pap.*, vol. 1830, 2021, doi: 10.1088/1742-6596/1830/1/012011.

[31] V. M. Putri, M. Masjkur, and C. Suhaeni, "Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status," *J. Phys. Conf. Ser. Pap.*, 2021, doi: 10.1088/1742-6596/1863/1/012073.

[32] A. Kishor and C. Chakraborty, "Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE," *Int. J. Syst. Assur. Eng. Manag.*, 2021, doi: 10.1007/s13198-021-01174-z.

[33] K. Gowri and M. Saranya, "Cervical Cancer Prediction using Outlier deduction and Over sampling methods," *Int. J. Innov. Res. Eng.*, vol. 3, no. 3, pp. 186–190, 2022.

[34] L. E. O. Breiman, "Random Forest," in *Machine*

DOI: https://doi.org/10.29207/resti.v6i4.4147

Learning, 2001, pp. 5–32.
[35] S. M. Learning, "Bagging and Random Forest for Imbalanced Classification," 2021.
[36] S. Chiu and D. Tavella, "Introduction to Data Mining," *Data Min. Mark. Intell. Optim. Mark. Returns*, pp. 137–192, 2008, doi: 10.1016/b978-0-7506-8234-3.00007-1.
[37] R. Fekadu, "Machine Learning Models Evaluation and Feature Importance Analysis on NPL Dataset," no. NeurIPS 2021.
[38] J. Miao and W. Zhu, "Precision-recall curve (PRC) classification trees," *Evol. Intell.*, 2021, doi: 10.1007/s12065-021-00565-2.
[39] S. Pérez, F. Pablo, M. Camblor, and P. Filzmoser, "Visualizing the decision rules behind the ROC curves : understanding the classification process," *AStA Adv. Stat. Anal.*, no. 0123456789, 2020, doi: 10.1007/s10182-020-00385-2.

DOI: https://doi.org/10.29207/resti.v6i4.4147