Journal of Theoretical and Applied Information Technology
31st January 2024. Vol.102. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMBALANCED CLASS LEARNING IN VISION BASED CLASSIFICATION OF VECTOR MOSQUITO SPECIES

RESHMA PISE [1], KAILAS PATIL [2]

[1]Assistant Professor, Department of Computer Engineering, Vishwakarma University, India

[2] Professor, Department of Computer Engineering, Vishwakarma University, India

E-mail: [1] reshma.pise@vupune.ac.in, [2] kailas.patil@vupune.ac.in

ABSTRACT

Vector-borne diseases, primarily transmitted by mosquitoes, remain a significant global public health concern. Accurate and timely identification of mosquito species is crucial for comprehending disease transmission patterns and implementing effective vector control measures. In recent years, vision-based deep learning techniques have shown promising results in the classification of mosquito species. However, the natural class imbalance present in real-world mosquito species datasets can negatively impact the predictive performance of CNN based classifiers. This paper presents three popular class imbalanced learning strategies: Oversampling, Under-sampling, and Synthetic Minority Oversampling Technique (SMOTE) to address the skewed class distribution. We investigate the effectiveness of these solutions on the imbalanced images dataset of vector mosquito species with the aim of enhancing the performance of CNN-based classifier. The classification outcomes demonstrate that SMOTE based CNN outperforms other techniques in terms of evaluation metrics: sensitivity, specificity, F-score and accuracy. Class imbalance learning techniques are vital in vector control applications where the accurate classification of rare class i.e., harmful species is crucial for effective monitoring and control of disease vectors.

Keywords: *Class Imbalance, Deep Learning, Mosquito Species, Image Classification, Sampling Techniques.*

1. INTRODUCTION

Vision based automatic classification of insect genera, species and sex has become a great research interest owing to the advancement in Machine Learning (ML) and Deep Learning (DL) coupled with image processing techniques [1-2]. Identifying the presence of diverse species within a region can provide valuable insights about ecological system, serving as a metric for tracking the existence of invasive species and environmental shifts. Deep learning models like Convolutional Neural Networks (CNNs) are finding growing usage in the field of ecological research. CNNs are specifically designed to handle image-related tasks, making them well-suited for image classification problems like insect species identification. The CNN classifier is trained with a dataset of images of different species. The network extracts and learns morphological feature representations unique to each species from the dataset. The trained model can be employed for predicting the species of a new image [3].

CNNs have demonstrated very high accuracy in automatic genera, species and sex classification tasks. However one of the major challenges for developing accurate CNN model is the class imbalance which is inherent for small insect datasets like mosquitoes. For some of the insect classes, adequate number of samples may not be available and these classes are underrepresented [4-6]. The imbalanced dataset contains significantly less number of samples of certain rare species than other species [7].

There may be several reasons for this:

(i) Geographic Variation: Certain mosquito species may be rare or difficult to find in some geographical regions, limiting the availability of samples.

(ii) Seasonal Variation: Some species are more abundant during certain times of the year.

(iii) Mosquito Control Efforts: Due to aggressive vector programs, populations of species which act as disease vectors can be reduced leading to an imbalance in the dataset.

1.1 Class-Imbalance Problem

Class imbalance is an extremely common scenario in computer vision tasks, where the dataset has skewed distribution i.e., the training set does not have an equal distribution of images across all classes. It occurs when some of the classes (minority) contain significantly fewer samples as compared to other classes (majority). The predictive capability of classification algorithm is affected by skewed class distribution. Class imbalance can lead to biased classifiers that perform poorly on underrepresented classes [8-9]. Skewed data exists in numerous real-life applications including medical diagnosis, fraud detection, intrusion detection and species identification among others [10]. In such applications, precise identification of the rare class instance holds greater significance than that of the majority class instance. In the typical machine learning task of predicting cancer, the frequency of disease cases (cancer patients) can be 1000 times lower than that of normal cases (healthy individuals). Classification goal is to detect individuals with disease i.e., the accurate identification of cancer patient is more crucial than the healthy patients. We would be more interested in correctly predicting the minority class instances. As there are a huge number of majority class examples, the classifier might exhibit high classification accuracy but it may struggle to learn and make accurate predictions for minority class instances. The performance of the classifier leans towards the majority class and the model will perform poorly in identifying the minority class [11-13].

Despite advances in deep neural networks (DNNs), the domain of deep learning with class imbalanced data still remains relatively unexplored. There has been a notable lack of systematic research regarding the impact of imbalanced dataset on CNN classifier's performance when applied to mosquito species datasets.

In our research work, we evaluate data-level solutions to mitigate the class imbalance within a dataset comprising of images of adult mosquitoes belonging to two vector species.

The key contributions of this study can be summarized as follows.

1. We examine three class imbalance learning methods: i) Oversampling ii) Under-sampling and iii) SMOTE for addressing the class imbalance present in a dataset comprising images of two vector mosquito species: Aedes Aegypti and Culex Quinquefasciatus with Aedes representing the minority class.

2. We apply the imbalance learning techniques on the imbalanced dataset and transfer learn a pre-trained VGG-16 CNN model to distinguish between the two species.

3. We evaluate and determine the most effective class imbalance learning solution in terms of four evaluations metrics: sensitivity, specificity, F-score and accuracy.

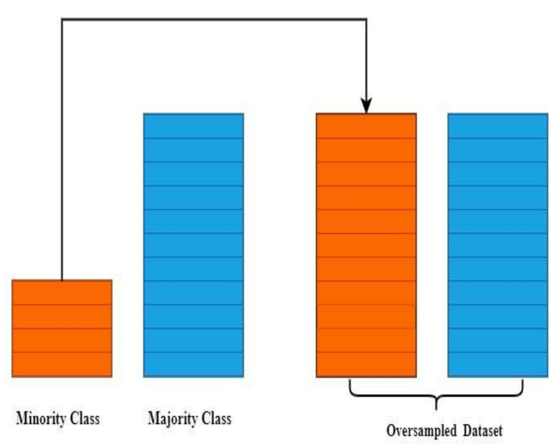
The rest of the paper is outlined as follows. Section 2 describes popular techniques for addressing the class imbalance problem and provides an overview of existing related studies. Section 3 presents our proposed methodology, which includes details on the dataset, model implementation and the evaluation metrics employed. Section 4 provides discussion on experimental results. Finally, the conclusion is presented in section 5.

2. CLASS-IMBALANCE LEARNING AND RELATED WORK

There have been several solutions proposed to address the issue of class imbalance. They can be broadly classified into two categories: Data Level and Algorithm level. Data Level sampling methods remain a popular choice for tackling the class-imbalance because of their simplicity and apparent effectiveness [14-15]. Data level approaches involve resampling the instances in the training dataset to balance the class distribution. Data level techniques can be further categorized into: (i) Oversampling (OS) (ii) Under-sampling (US) and (iii) Hybrid methods.

2.1 Oversampling

This is the most commonly adopted sampling method in deep learning. A simple version of oversampling is Random minority sampling technique (ROS) which involves duplicating randomly selected minority class samples until their count matches that of the majority classes as shown in Fig. 1.



ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 1. Oversampling

It is established from several research studies that OS significantly boosts the performance of ML and DL classifiers when dealing with imbalanced training data [16]. In a comparative study involving seven data sampling methods, ROS emerged as the most effective method when evaluating the performance of three pre-trained CNN architectures (LeNet-5, ResNet-10, and All CNN) across three benchmark multi-class image datasets: MNIST, CIFAR-10, and ImageNet [11]. The primary evaluation metric employed in this study is the Area Under the Curve of the Receiver Operating Characteristic Curve (ROC-AUC). The findings revealed that OS technique led to a significant performance enhancement as compared to baseline model i.e., the CNN trained on the original imbalanced data, without causing over fitting issues. In contrast, under sampling resulted in poorer performance compared to the baseline model.

Hensman et al. analyzed the impact of class imbalance and ROS on the classification performance of ALexnet CNN. The original CIFAR-10 dataset, consisting of 10 classes with 6000 images per class, was employed to create 10 imbalanced datasets for testing. The authors report that OS significantly enhanced F1-scores from baseline model for all distributions [17].

ROS has been successfully employed to tackle class skew ness in numerous medical image classification studies. Oversampling of malignant (minority class) images was effective to mitigate the imbalance when training a Resnet model to identify micro-calcifications within mammography images [18]. Application of ROS on the training data improved AUC-ROC score of ResNet-22 CNN in discerning malignant (cancerous) and benign (healthy) samples from three digital mammography image datasets [15]. In another study, a CNN was built to classify cancerous and noncancerous specimens using two heavily imbalanced datasets containing histopathology images of breast cancer specimens [19]. The authors applied OS, US and hybrid techniques: SMOTE and ADASYN to tackle the imbalance. The classification results revealed that OS methods improved the balanced accuracy up to 3-4% as compared to the baseline model, on both the datasets.

It has been observed that while oversampling is effective, it may lead to over- fitting due to the creation of duplicates of minority class instances. Over-fitting is caused when the model tightly fits the training data, making it difficult to generalize to new data [20-21].

2.2 Under Sampling

This is another commonly employed technique. A basic form of under-sampling is Random under-sampling (RUS). As shown in Fig. 2, majority class examples are randomly discarded to obtain equal number of examples in each class [22].

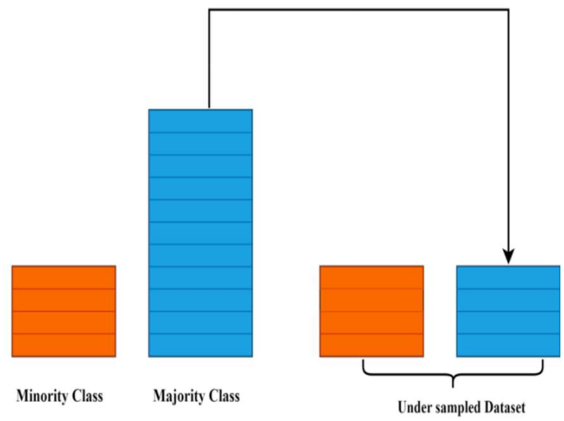


Figure 2. Under-sampling

RUS emerged as the most effective sampling technique when compared to the other five sampling methods in training three classifiers: Gradient-Boosted Trees, Logistic Regression and Random Forest on Slowloris attacks data set in terms of AUC and GM metric [23]. Employing an US approach that entails augmenting the size of the minority class by selecting a specific number of misclassified minority class instances (false positives) proved to be effective in enhancing the conditional precision of CNN in classifying imbalanced cloud images data [24]. In another study, application of US enhanced the classification accuracy of the pre-trained CNNs for Plankton image classification task [25]. In their study on classifying chest X-rays, Qu et al. determined that both OS and US were effective strategies in mitigating the class imbalance issue [26]. Since RUS creates models with substantially reduced number of examples, it leads to a saving in computational resources and training time [23-24].

The drawback of US techniques is that there is a possibility of dropping out informative samples and features forcing the model to learn from a lesser amount of information, which can adversely impact model's performance [15, 27-28].

Journal of Theoretical and Applied Information Technology
31st January 2024. Vol.102. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3. Hybrid Sampling Methods

Hybrid sampling algorithms have been developed to overcome the challenges associated with both US and OS techniques [11]. These algorithms focus on producing artificial examples by interpolating neighboring data points from the minority class rather than merely duplicating instances. The synthetic instances are created in the feature space as opposed to the data space. By adding these synthetic instances to the original dataset, valuable information is preserved, resulting in an augmented dataset that can be used for training classification models. Two well-known algorithms of this type encompass the Synthetic Minority Over-sampling Technique (SMOTE) and the Adaptive Synthetic (ADASYN). However, SMOTE remains highly preferred, as it has demonstrated its effectiveness in achieving superior classifier performance compared to solely oversampling and under-sampling in imbalanced datasets.

SMOTE produces artificial examples of the minority class by interpolating the feature values of closely positioned existing minority class instances (Fig. 3). These synthetic instances are subsequently incorporated into the original dataset [29].

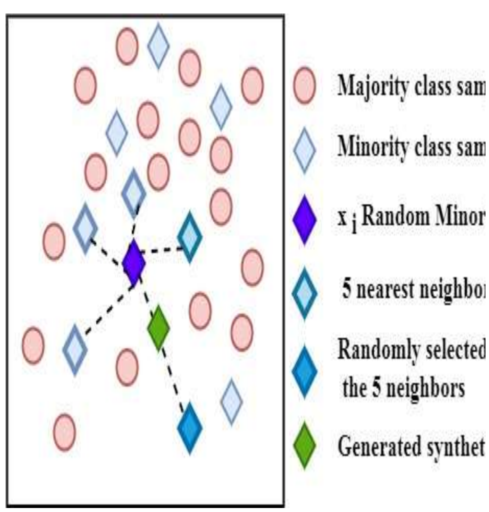


Figure 3. SMOTE Oversampling

Here is a high-level overview of how SMOTE works for image classification:

Algorithm : SMOTE Algorithm

Input: M_org : Dataset consisting of images and their corresponding labels.

Output: M_syn : Augmented dataset

1. M ← Minority class

   N ← Number of synthetic samples to generate i.e., sampling rate

   k ← Number of nearest neighbors

2. For each instance x belonging to M

   a) Identify the k nearest neighbors of x using a distance metric such as Euclidean distance between x and all other instances of M.

   b) Select N random samples (i.e., x_1, x_2 … x_N) from x's k-nearest neighbors.

   c) For each sample x_k belonging to M' (k = 1, 2, 3…N) generate a new sample using the formula:

   $x' \leftarrow x + \text{random}(0,1) * (x - x_k)$

   where the function random(0, 1) generates a random number within the range of 0 to 1.

3 Add x', the synthetic sample with the new feature values to the dataset M_syn.

4 Return the augmented dataset M_syn.

5 End

In a comparative study, Joloudari et al. evaluated the impact of various data imbalance solutions on the performance of DNNs and CNNs using random data distributions of three datasets: KEEL, breast cancer, and Z-Alizadeh Sani [30]. The proposed SMOTE based CNN achieved 99.08% accuracy and outperformed the other approaches in terms of eight evaluation metrics: accuracy,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

precision, recall, F1-score, G-Mean, specificity, AUC, and Kappa. An efficient deep oversampling method called "DeepSMOTE", based on SMOTE technique improved the loss function [31]. The technique demonstrated its significant effectiveness, particularly in settings involving oversampling with GANs.

Reza et al. trained a CNN classifier for classifying histopathological breast cancer images using imbalanced training datasets. They assessed the effectiveness of OS, US, SMOTE, and ADASYN [19]. The SMOTE CNN exhibited superior performance over other techniques concerning sensitivity, specificity, F-Score, accuracy, and balanced accuracy (BAC).

SMOTE technique was used in combination with different feature extraction methods to train the ML classifiers. The results indicated that both oversampling and transforming binary features into numerical values significantly enhanced the performance of the base model and Random forest outperformed the other models in terms of accuracy, sensitivity, specificity, precision, and F1-score [32]. SMOTE was employed to tackle imbalanced human activity datasets that included activities like walking, jogging, and jumping which are naturally skewed. The technique led to a significant improvement in classifiers' learning for underrepresented activities i.e., minority class instances [14, 33].

3. PROPOSED METHODOLOGY

3.1. Dataset Description

The dataset comprises of mobile-captured images of adult mosquitoes belonging to vector species [34]. As shown in Fig. 4, the original dataset was utilized to generate an imbalanced distribution, with 302 images of Aedes Aegypti and 1208 images of Culex Quinquefasciatus. In this dataset, the Culex class is approximately four times more prevalent than the Aedes class, resulting in a class imbalance ratio of 1:4. The dataset was randomly split into partitions of 70% for model training, 20% for validation, while the remaining 10% portion of the data was reserved for testing.

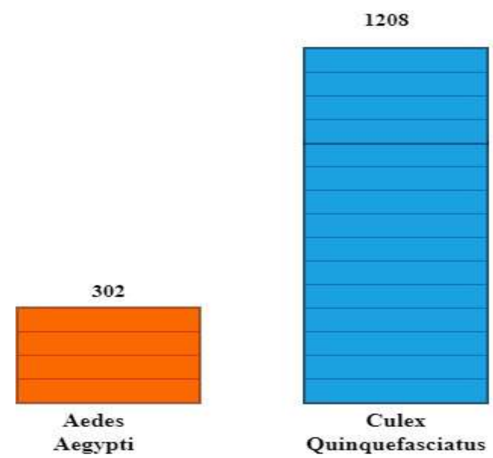


Figure 4. Imbalanced Dataset

Fig. 5 depicts the important phases involved in the proposed system.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

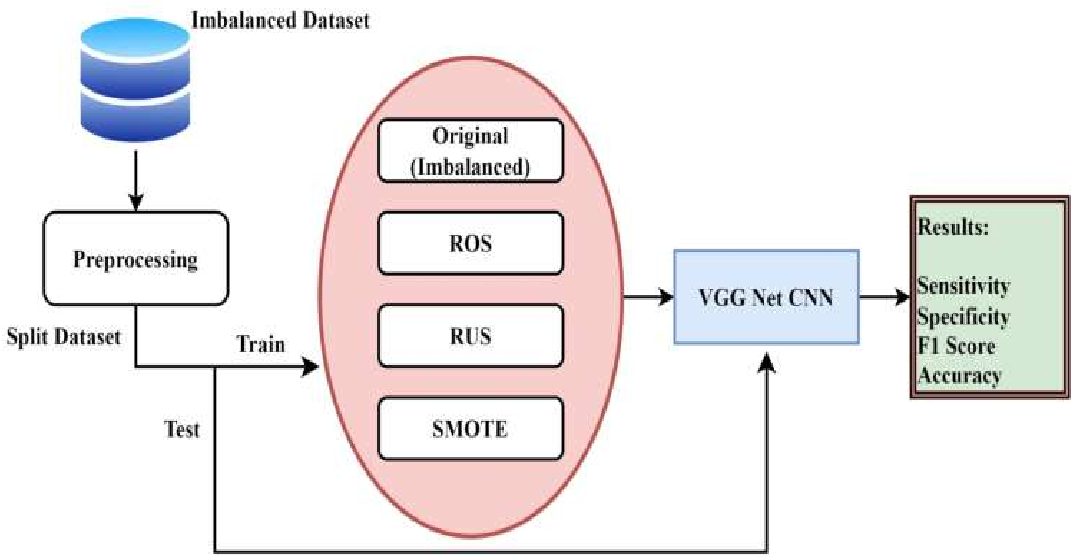


Figure 5. Proposed Methodology

3.2 Model Training

In the current work, a popular pre-trained deep CNN architecture: VGG-16, was adopted for the image based mosquito species classification. The experiments were conducted using open-source DL framework TensorFlow and Keras API , written in Python. To compile and train the models, we chose the Adam algorithm to optimize the model parameters, utilized 'binary_crossentropy' as the loss function with a learning rate of 0.001, and set the number of epochs to 50. The network's input consists of images with dimensions 224 x 224 x 3. The network is trained on the imbalanced dataset to analyze the impact of class imbalance on model's performance (Baseline model). The training data is balanced by applying the three sampling techniques discussed under section 2. To achieve this, we employed the ImageDataGenerator and the Keras imblearn package. The models were trained on the balanced datasets obtained.

3.3 Model Evaluation

Evaluation metrics play a vital role in every classification technique as they are instrumental in assessing the effectiveness of the classifier in learning the data. The most widely used metrics for evaluating classification performance of CNNs are accuracy and its inverse i.e., error rate. Accuracy quantifies the overall correctness of the classifier's predictions by calculating the ratio of correct predictions, which includes both true positives and true negatives, to the total number of predictions made by the classifier.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

True Positives (TP): These are the instances from the positive class that the classifier correctly identifies as positive.

True Negatives (TN): These are the instances from the negative class that the classifier correctly identifies as negative.

Similarly False Positives (FP) and False Negatives (FN) are the instances from the negative (or positive) class that the classifier incorrectly predicts as positive (or negative).

Though accuracy is an intuitive metric, it can be misleading in scenarios with imbalanced datasets because the prediction result shows a bias towards the majority class. For example, let's consider a dataset where only 1% of the instances are labeled as positive (minority class). In such a scenario, the classifier may predict all observations as majority class most of the time and achieve 99% accuracy whereas it fails to correctly predict a single example from the minority class. The emphasis is usually on correctly identifying the minority class, as it is the class of interest and misclassifying those instances can have severe consequences in real-world applications. In such cases, other evaluation metrics like precision, recall, F1- score, AUC-ROC are used to correctly identify the minority class instances [8, 30, 32].

Recall (Sensitivity or True Positive Rate): Recall quantifies the fraction of accurately predicted positive instances (true positives) in relation to all the actual positive instances (comprising both true positives and false negatives). This metric is important when the goal is to monitor and minimize false negatives. Recall is calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

Specificity (True Negative Rate): Specificity measures the fraction of correctly predicted negative instances (true negatives) among all actual negative instances (true negative and false positives). It is a measure of classifier's ability to correctly identify negative instances.

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

Precision: It is the ratio of correctly predicted positive instances (true positives) among total number of positive predictions (true positives and false positives). It is particularly valuable when the primary goal is to reduce the occurrence of false positives.

$$Precision = \frac{TP}{TP + FP}$$

F-Score: Also known as the F1-score, is a metric that proves particularly advantageous in the context of imbalanced datasets. It gives a balanced assessment of the model's efficiency by combining both precision and recall.

$$F1 - Score = \frac{2 * TP}{2 * TP + FP + FN}$$

4. RESULTS AND DISCUSSION

In this section we present the classification reports and evaluate the effectiveness of imbalance learning approaches. The classifier's performance is measured in terms of sensitivity, specificity, F1-score and accuracy. Table 1 presents the statistics for the baseline model and the three models resulting from the application of sampling solutions.

Table 1: Classification results

Data Sampling Approach	Sensitivity	Specificity	F-score	Accuracy
Baseline model (Imbalanced data)	0.6512	0.9201	0.6540	0.8583
RUS	0.7302	0.7521	0.5241	0.7461
ROS	0.7710	0.8273	0.6687	0.8610
SMOTE	0.8203	0.8875	0.7049	0.8685

As indicated in the table, the presence of imbalanced data (where 80% of the training samples represent Culex species), adversely affects the classification performance of the CNN model. The model exhibits bias in favor of the Culex class i.e., every sample is assigned a low probability of Aedes class during testing. This accounts for lower sensitivity (0.6512) compared to specificity (0.9201) though the overall accuracy is high (0.8583).

Application of RUS on the majority class instances resulted in an improvement in sensitivity in the trained model because it addressed the imbalance in the minority class instances. However, this improvement has been accompanied by a significant drop in specificity, F-score, and accuracy. This will result from the loss of valuable information during under sampling and the classifier has to learn with limited number of samples.

Employing oversampling techniques: ROS and SMOTE resulted in notable performance improvement as compared to the baseline model. The SMOTE approach outperformed the others in terms of all the evaluation metrics. These findings suggest that the techniques used to address class imbalance tend to decrease the gap between sensitivity and specificity values.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The performance of RUS can be enhanced by combining it with techniques like Near-Miss or One-sided selection. These methods are designed to retain crucial information essential for learning. Similarly several variations of SMOTE such as Borderline-SMOTE, Deep SMOTE and Safe-Level-SMOTE can be employed to improve upon the results obtained. These algorithms generate new examples in close proximity to the easily misclassified data points, aiming to boost the classification accuracy of the minority class.

5. CONCLUSIONS

In this study, we investigated the effect of class imbalance on CNN classifier's performance for the vision based classification of mosquito species. We applied three popular techniques: RUS, ROS and SMOTE on imbalanced images dataset of Aedes Aegypti and Culex Quinquefasciatus , where Aedes Aegypti species is underrepresented. The effectiveness of the imbalance learning techniques on the CNN classifier's performance was compared in terms of sensitivity, specificity, F-score and accuracy. The findings of the experiments indicate that class imbalance leads to a stronger bias toward the majority i.e., Culex class and poor performance on Aedes species (rare class). Oversampling techniques exhibited significant improvement of performance over the baseline whereas under sampling resulted in poor performance. The results reveal that the SMOTE-based approach demonstrates robustness and outperforms other methods achieving 87% accuracy. The findings suggest that oversampling techniques, particularly those based on synthetic data, provide a reliable solution to alleviate the challenges posed by imbalanced datasets in CNN-based image classification.

Class imbalance techniques are vital in applications where misclassification of rare class is costly, as they help overcome the inherent challenges posed by imbalanced datasets, leading to more accurate and reliable classification results.

REFERENCES:

[1] Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., Venturini, G., Martineau, M., Conte, D., Raveaux, R., Arnault, I., & Munier, D. (2017) "A survey on image-based insect classification Pattern Recognition", 65. 273-284. 10.1016/j.patcog.2016.12.020.

[2] Tuda, M., & Luna-Maldonado, A. I. (2020). "Image-based insect species and gender classification by trained supervised machine learning algorithms", Ecological Informatics, 60, 101135. https://doi.org/10.1016/J.ECOINF.2020.101135

[3] Pise, R., & Patil, K. (2023) "A Deep Transfer Learning Framework for the Multi-Class Classification of Vector Mosquito Species", Journal of Ecological Engineering, 24(9), 183–191. https://doi.org/10.12911/22998993/168501

[4] Bjerge, K., Alison, J., Dyrmann, M., Frigaard, C. E., Mann, H. M. R., & Høye, T. T. (2023) "Accurate detection and identification of insects from camera trap images with deep learning" PLOS Sustainability and Transformation, 2(3), e0000051. https://doi.org/10.1371/journal.pstr.0000051

[5] Johnson, R. A., Chawla, N. V., & Hellmann, J. J. (2012) "Species distribution modeling and prediction: A class imbalance problem", Proceedings - 2012 Conference on Intelligent Data Understanding, CIDU 2012, 1, 9–16. https://doi.org/10.1109/CIDU.2012.6382186

[6] Ong, S. Q., & Ahmad, H. (2022) "An annotated image dataset for training mosquito species recognition system on human skin", Scientific Data, 9(1), 1–6. https://doi.org/10.1038/s41597-022-01541-w

[7] Høye, T. T., Ärje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., Mann, H. M. R., Meissner, K., Melvad, C., & Raitoharju, J. (2021) " Deep learning and computer vision will transform entomology", Proceedings of the National Academy of Sciences of the United States of America, 118(2), 1–10. https://doi.org/10.1073/PNAS.2002545117

[8] Johnson, J. M., & Khoshgoftaar, T. M. (2019). "Survey on deep learning with class imbalance", Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0192-5

[9] Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009), "Classification of imbalanced data: A review", International Journal of Pattern Recognition and Artificial Intelligence, 23(4), 687–719. https://doi.org/10.1142/S0218001409007326

[10] Haque, M. M., Skinner, M. K., & Holder, L. B. (2014), " Imbalanced class learning in epigenetics", Journal of Computational Biology, 21(7), 492–507. https://doi.org/10.1089/cmb.2014.0008

[11] Buda, M., Maki, A., & Mazurowski, M. A. (2018) "A systematic study of the class

imbalance problem in convolutional neural networks", Neural Networks, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

[12] Li, Q., Mao, Y. (2014) "A review of boosting methods for imbalanced data classification", Pattern Analysis and Applications, 17, 679–693. https://doi.org/10.1007/s10044-014-0392-8

[13] Valova, I., Harris, C., Mai, T., & Gueorguieva, N. (2020), "Optimization of convolutional neural networks for imbalanced set classification", Procedia Computer Science, 176, 660–669. https://doi.org/10.1016/j.procs.2020.09.038

[14] Alharbi, F., Ouarbya, L., & Ward, J. A. (2022), "Comparing Sampling Strategies for Tackling Imbalanced Data in Human Activity Recognition", Sensors, 22(4), 1–20. https://doi.org/10.3390/s22041373

[15]Walsh, R., & Tardy, M. (2023), "A Comparison of Techniques for Class Imbalance in Deep Learning Classification of Breast Cancer", Diagnostics, 13(1), 1–19. https://doi.org/10.3390/diagnostics13010067

[16] Shorten, C., & Khoshgoftaar, T. M. (2019), " A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0197-0

[17] Masko, D., & Hensman, P. (2015), " The Impact of Imbalanced Training Data for Convolutional Neural Networks", KTH, Skolan för datavetenskap och kommunikation (CSC).

[18] Bria A, Marrocco C, T. F. (2020), " Addressing class imbalance in deep learning for small lesion detection on medical images", Comput Biol Med., 120:103735. https://doi.org/10.1016/j.compbiomed.2020.103735

[19] Reza, M. S., & Ma, J. (2019) , "Imbalanced histopathological breast cancer image classification with convolutional neural network", International Conference on Signal Processing Proceedings, ICSP, 2018, 619–624. https://doi.org/10.1109/ICSP.2018.8652304

[20] Cai, Y., Li, Y., & Zheng, Z. (2015), "Oversampling Method for Imbalanced Classification", Computing and Informatics, 34(5), 1017–1037

[21] Li, Z., Kamnitsas, K., & Glocker, B. (2021), " Analyzing Overfitting under Class Imbalance in Neural Networks for Image Segmentation", IEEE Transactions on Medical Imaging, 40(3), 1065–1077. https://doi.org/10.1109/TMI.2020.3046692

[22] Yen, S. J., & Lee, Y. S. (2006), "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset", Lecture Notes in Control and Information Sciences, 344, 731–740. https://doi.org/10.1007/978-3-540-37256-1_89

[23] Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019), " Severely imbalanced Big Data challenges: investigating data sampling approaches", Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0274-4

[24] Matsuoka, D. (2021), "Classification of imbalanced cloud image data using deep neural networks: performance improvement through a data science competition", Progress in Earth and Planetary Science, 8(1). https://doi.org/10.1186/s40645-021-00459-y

[25] Kim J, Lee H, Park M (2016), "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning", 2016 IEEE International Conference on Image Processing (ICIP), 3713–7. https://doi.org/10.1109/ICIP.2016.7533053

[26] Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J., & Tyrrell, P. N. (2020), " Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging", International Journal of Computer Assisted Radiology and Surgery, 15(12), 2041–2048. https://doi.org/10.1007/s11548-020-02260-6

[27] Alam, T. M., Shaukat, K., Hameed, I. A., Khan, W. A., Sarwar, M. U., Iqbal, F., & Luo, S. (2021) , "A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining", Biomedical Signal Processing and Control, 68(April), 102726. https://doi.org/10.1016/j.bspc.2021.102726

[28] Kantabutra S, Jindaluang, W., & Chouvatut, V. (2014), "Under-sampling by algorithm with performance guaranteed for class-imbalance problem", International Computer Science and Engineering Conference (ICSEC). https://doi.org/10.1109/ICSEC.2014.6978197

[29] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2011), "SMOTE: Synthetic Minority Over-Sampling Technique", Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

[30] Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023), " Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks",

Applied Sciences (Switzerland), 13(6). https://doi.org/10.3390/app13064006

[31] Dablain, D., Krawczyk, B., & Chawla, N. V. (2022), "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data", IEEE Transactions on Neural Networks and Learning Systems, 34(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

[32] Mahmudah, K. R., Indriani, F., Takemori sakai, Y., Iwata, Y., Wada, T., & Satou, K. (2021), "Classification of imbalanced data represented as binary features", Applied Sciences (Switzerland), 11(17). https://doi.org/10.3390/app11177825

[33] Chen, Y., & Shen, C. (2017), "Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition", IEEE Access, 5, 3095–3110. https://doi.org/10.1109/ACCESS.2017.2676168

[34] Pise, R., Patil, K., Laad, M., & Pise, N. (2022), "Dataset of vector mosquito images", Data in Brief, 45, 108573. https://doi.org/10.1016/j.dib.2022.108573