Addressing Class Imbalance in Image-Based Plant Disease Detection: Deep Generative vs. Sampling-Based Approaches

Nasik Muhammad Nafi[1], William H. Hsu[1]
[1]Department of Computer Science, Kansas State University, Manhattan, KS 66506, USA
nnafi@ksu.edu

***Abstract*—Deep learning-based classifiers for object recognition and classification have been used in the domain of plant disease detection, particularly lesions from leaf images. In such domains, as expected, deep neural networks perform better using balanced data sets than imbalanced ones, as they exhibit some inductive bias favoring balanced data from each class. However, data sets for plant disease detection are often imbalanced due to the rarity of disease lesions in real-world settings. While deep generative approaches such as generative adversarial networks (GANs) have been established as an effective means of augmenting high-dimensional image data, the literature lacks a detailed study of the effectiveness of GAN-based models on a plant disease detection task, compared to sampling-based approaches traditionally used to reduce the skewness of the data. In this paper, we comparatively evaluate an image classifier based on a dense convolutional neural network (CNN), trained using a GAN, versus the same CNN model used in tandem with undersampling, oversampling, and an adaptation of the Synthetic Minority Over-sampling Technique (SMOTE). The GAN-based approach is shown to attain significantly higher recall and hence F-measure and ROC AUC against each of these.**

***Keywords*—plant disease detection, class imbalance, data augmentation, sampling vs generative, generative adversarial network**

I. Introduction

Plant disease identification is crucial in agriculture because it can drastically change crop yield and quality of production. Failure to detect some viral diseases can have devastating effects on food sustainability and national economies. Therefore, research communities from different disciplines such as microbiology, agronomy, plant science are working to develop novel and accurate methods for plant disease diagnosis. However, methods that leverage domain knowledge require the involvement of domain experts and specific equipment. With the advancement in computational processing of high-dimensional data such as image, disease detection using only image data has become feasible.

Disease identification from image data can be considered as a visual anomaly detection task. Anomaly detection is the task of identifying or classifying unusual observations from data. As these anomalous data points can be linked to some sort of problem or abnormal events such as electricity pilferage, fraudulent transactions, rare diseases, product defects, etc., identification of those events are of particular interest. Due to the infrequent occurrence of anomalous events, data sets available for anomaly detection are inherently imbalanced. Plant disease data sets are no exception and they often suffer from an imbalance of different magnitude.

978-1-7281-7539-3/20/$31.00 ©2020 IEEE

The simplest approach to classify anomalous data is to identify the data points that vary significantly from common statistical properties of a distribution. Popular machine learning-based techniques for anomaly detection are decision trees, k-nearest neighbors (k-NN), k-means clustering, and support vector machine (SVM) based clustering. In the presence of imbalanced data, however, these algorithms tend to treat minority samples as noise and hence produce a strong bias towards the majority class. Skewed distributions also lead to failure in learning the true features of the minority class owing to the lack of enough representative.

Two types of approaches have been used most by researchers in order to improve the classification performance of imbalanced data sets. One approach is to hit the problem from algorithmic perspective, and another is to look at the problem from data-level. In the first approach, the classifier itself is altered at the algorithm level to bias towards the minority class, while keeping the original data unchanged. For example: cost-sensitive learning [2] and recognition-based learning [3]. Cost-sensitive learning emphasizes the cost of different kinds of misclassification. The aim of this type of learning is to limit the total cost at minimum level [4]. At the data-level, sampling or synthesizing techniques are applied to create or delete samples to accomplish a balanced data distribution [5]. In this paper, we are particularly interested in the data-level techniques.

In the last few decades, several oversampling-based techniques have been proposed to mitigate the skewness of the data [6] [7]. Recent deep generative models such as VAE and GAN have already gained success in generating a variety of complex data, such as handwritten digits, faces, road signs, bedroom scenes, and CIFAR images [8] [9] . Therefore, nowadays, it is usual to artificially generate additional anomalous data to reduce the imbalance. The quality of the new data is dependent on these data synthesis techniques and significantly affects the performance of the classifier. However, most of the previous works generally compared their works with a baseline or other works of the same category. For instance, oversampling-based techniques have been compared with other oversampling-based techniques while generative models are evaluated based on whether they can outperform the vanilla classifier trained on data set without augmentation.

We attempt to investigate the performance of data balancing techniques from different categories, and specifically how learning of the classifier is influenced by the addition of new synthetic data. We select four techniques - random undersampling, random oversampling, SMOTE, and GAN - from four different categories. As synthetic oversampling-based approaches like SMOTE are not directly applicable to high-dimensional data we adapt SMOTE to use it for image data. We carry out extensive experiments and evaluate the techniques based on precision, recall, F1 score, and AUC. We show that GAN-based augmentation outperforms simple undersampling, simple oversampling, and synthetic oversampling based approaches. We also analyze the progression of loss function during training for the training and validation sets. This helps to understand the learning process particularly to validate underfitting and overfitting as well as to get an idea of bias-variance trade-off.

II. Related Work

The issue of class imbalance can be addressed by using a higher weight on error term in the loss function if the classifier misclassifies samples from minority classes [10], or by informing the algorithm about prior class probabilities ahead of time [11]. As discussed in Section I, we are not particularly interested in algorithm-level solutions, rather our aim is to work with the data-level methods that operate on the training set and change its class distribution.

At the data-level, the most widely used approaches are different sampling methods. A straightforward approach to achieve balance in the data set is to use undersampling or oversampling. Undersampling removes majority class samples from the data set while oversampling randomly adds duplicate copies of selected samples from minority classes. Oversampling has been shown to be an effective and robust approach [12], however, balance is achieved at the cost of an increased risk of overfitting [6]. In the case of image data set there is also some variation of random oversampling which replicates random images adding slight variations such as rotation, translation, blur, center cropping, contrast, sharpening, etc. This strategy has been used for a while in a wide range of applications like plant leave classification [13], concealed cargo inspection [14], human disease detection [15].

SMOTE is one of the most popular oversampling methods for dealing with imbalance data set. SMOTE [6] augments artificial examples by interpolating neighboring data points. Some extensions of this technique are also available. For example, Han et al. proposed borderline-SMOTE method to generate synthetic samples on the borderline between two classes [7]. As the samples lying on the border are critical to learning the class boundary, borderline-SMOTE outperforms general SMOTE. DataBoost-IM generates new synthetic data by selecting difficult examples with boosting preprocessing and using information from those selected examples [16]. ADASYN tries to sample more new data around difficult samples than simple minority samples [17].

In 2014, Goodfellow et al. proposed generative adversarial networks (GAN) to generate images [8]. After its successful appearance, in the last couple of years, a good number of different architectures were proposed for different types of image generation. They have already been used in generation of medical images [18] [19], acoustic scenes [9], plant leaves [20] etc. In recent years, some GANs have been proposed specifically to reduce the class imbalance. Radford et al. (2016) proposed Deep Convolutional GANs (DCGAN) to ensure stable training in most settings [21]. This model is used as base architecture for many of the later approaches. WGAN uses the architecture of DCGAN, however, it incorporates a better loss function to approximate the data distribution [22]. Zhu et al. (2017) proposed CycleGAN which aims to create images with some particular emotions because in emotion classification some classes of emotions like disgusted are comparatively less available than other classes like happy or sad [23]. Cenggoro et al. proposed the class expert generative adversarial network (CE-GAN) for imbalance data classification which integrates class-specific data generation at an early stage of the classifier [24]. Data Augmentation Generative Adversarial Network (DAGAN) ignores the dependency on the class labels and is capable of generating novel unseen classes of data [25]. As a result, DAGAN outperforms general GANs in the few-shot learning scenario.

Sampling-based approaches are proven to be successful in many data domains. Recently, several published systems have been developed in attempts to use deep learning and GANs to detect plant disease [26] [27]. However, there is a lack of literature comparing the traditional sampling-based techniques with the most recent GAN-based techniques in the task of image-based plant disease identification. Also, due to differences in the experimental settings and the data sets on which they were trained, these existing systems are not directly comparable to one another.

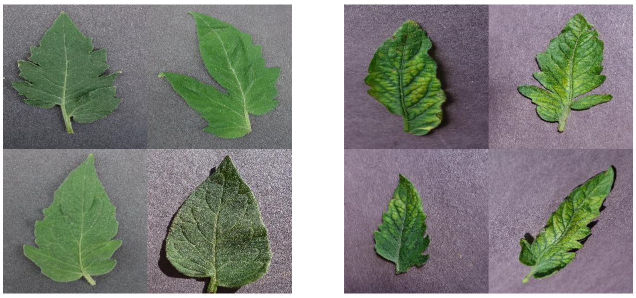
III. Methodology and Experimental Design

A. Data Set

We use a subset of the *PlantVillage* data set [28] which contains images of healthy and infected plant leaves. We select tomato plant leaves that exhibit severe imbalance across one class of disease compared to the healthy class. We consider tomato leaves infected with the mosaic virus as the minority class and healthy tomato leaves as the majority class. The data distribution is shown in Table. I. All the leaf images are captured against a similar grayish background. Fig. 1 presents some samples from both classes.

TABLE I
Number of images in the data set

Data Set	Healthy	Infected	Total
Tomato Leaf Image	1590	370	1960



(a) Healthy Leaves (b) Infected Leaves

Fig. 1. Examples of healthy and infected leaves.

B. Addressing Class Imbalance

In our study, we have identified four overall categories of data-level approaches to cope with class imbalance - undersampling, oversampling, synthetic oversampling, and generative models. Our comparative experiment design selects one method as representative from each of these categories and applies random undersampling, random oversampling, SMOTE, or a GAN, respectively. A brief description of the selected approaches are as follows:

1) Random Undersampling: The most commonly used undersampling method is random majority undersampling because of its simplicity and effectiveness. This approach randomly removes the samples from the majority class. Earlier works show in some cases undersampling outperforms oversampling [29]. However, the removal of data can lead to potential information loss.

2) Random Oversampling: The random oversampling method operates by replicating the randomly selected set of examples from the minority class so that the majority class does not have an overbearing presence during the training process. In this approach, a random image is chosen every time from the original data set until the required number of images are added to achieve the desired balance. The main drawback of this approach is the repetition of the same data, which can induce a bias towards the training instances (and anomalies represented among them).

3) Adaptation of SMOTE: Rather than adding replicated data points from the minority class, SMOTE oversamples the minority class by creating artificial data based on the original data [6]. SMOTE expands the minority class in a way that benefits the learning process by introducing at least some new information. SMOTE generates synthetic samples that lie, preferably in the feature space, between existing minority instances. At first, it finds the K-nearest neighbors of a specific sample x_i, then one of these K-nearest neighbors of x_i is randomly chosen and the euclidean distance between x_i and the selected random neighbor $\hat{x_i}$ is calculated. The distance term is multiplied by δ, which is a random number between 0 and 1, and finally, the result is added to the original sample x_i. Mathematically, the newly synthesized data point x_{new} can be represented as follows:

$$x_{new} = x_i + (\hat{x}_i - x_i) * \delta. \tag{1}$$

In the deep learning setting where we learn the feature vector of an image data inside the classifier, SMOTE is not particularly suitable for direct use. Even if we can extract the feature vector by any means from the image data and find the feature vector for a synthetic image using SMOTE, constructing the image from that feature vector without the help of any decoder or generative model is an important issue. One way to circumvent this issue is to use the whole image as its feature vector. Thus, if the width of the image is w pixel, the height of the image is h pixel, and there are 3 channels, then the size of the feature vector will be $c \times w \times h$. From here in the text, when we mention SMOTE that will denote this particular adaptation of SMOTE.

4) Wasserstein GAN: As there is a large and growing number of available GAN models, choosing one is a complicated task. In 2017, Arjovsky et al. introduced Wasserstein GAN (WGAN), an alternative to traditional GAN training [22]. Their approach achieves higher learning stability, addresses the mode collapse problem, and provides an easy method for hyperparameter tuning. WGANs measure the closeness between the model distribution and the real distribution by defining some distance function. Different distance functions have different impacts on convergence. WGAN minimizes an approximation of the Earth Mover (EM) distance which they named Wasserstein-1. They thus eliminate the need for sophisticated network architecture design and balanced training of discriminator and the generator.

WGAN incorporates DCGAN, which is one of the best Deep Convolutional Generative Adversarial Network [21], with the Wasserstein-1 loss function. The generator transforms a random input vector drawn from a uniform distribution to an image of the desired shape using a series of deconvolutional layers. The discriminator is like a classifier that uses convolutional and fully connected layers. The use of batch normalization provides higher stability to the network. While all layers of the discriminator use Leaky ReLU, in the generator the output layer uses tanh activation and the rest of the layers use ReLU activation.

C. Classifier Network Selection

The choice of the classification model is a crucial design decision for the framework developed in this paper. Since 2012, for image classification and object recognition tasks, Convolutional Neural Networks (CNNs or ConvNets) and other deep learning-based approaches have gained popularity over other approaches due to their high accuracy and robustness. CNNs are a special type of multi-layer neural network that can extract a hierarchy of features (in this domain, image-derived features) directly from image pixels without any preprocessing.

Every year in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), recently proposed approaches for visual object detection and classification compete with each other. According to the result of the ILSVRC 2015, the ResNet model outperforms most of the existing models [30]. Also, surveys of recent deep CNN models indicate the robustness of the ResNet model in image classification [31] [32]. Therefore, we selected ResNet as the representative deep learning classifier.

ResNet incorporates a residual module which allows training of very deep networks and mitigates the vanishing gradient and degradation problems [30]. ResNet makes heavy use of batch normalization. The residual module enables the training of neural networks with as many as 152 layers without much complexity. However, the commonly used variant of ResNet has 50 layers. ResNet-152 and ResNet-50 have single-model top-5 validation error of 4.49% and 5.25% respectively [30]. While ResNet-152 may lead to a slightly higher accuracy than ResNet-50, in our experiment this is not a paramount issue as long as we are comparing all data augmentation approaches using the same classifier. In consideration of accuracy and training time trade-offs, we have chosen ResNet-50 for our experiment.

D. Experimental Setup

In our study, we applied 5-fold cross-validation and 20% of the training data for validation. We made use of the *imbalanced-learn* package from *scikit-learn* library [33] for random undersampling, random oversampling, and SMOTE. For WGAN we used the official PyTorch implementation. The major parameters of the experiment are outlined as follows:

- Size of the images used: 64×64
- Number of channel in the image: 3
- Dropout rate in classifier network: 0.2
- Learning rate for the optimizer: 0.001
- Number of epochs for GAN training: 200,000
- Number of epochs for classifier training: 3,000

Parameter values are chosen either based on our empirical study or values documented in relevant published literature. We conducted the experiment on 64×64 RGB images as the original implementation of WGAN is readily available to use with the mentioned resolution. We trained the generator network for 200,000 epochs based on the convergence of the loss function as illustrated in the original WGAN paper [22]. We tested different learning rates and 0.001 appears to be a local optimal value. Similar empirical analysis motivated us to choose 0.2 as the dropout rate. For classifier training, we trained some of the models for nearly 10,000 epochs and found that the classifiers converge after 2,000 epochs. Therefore, we reported the performance of the classifier trained for 3,000 epochs.

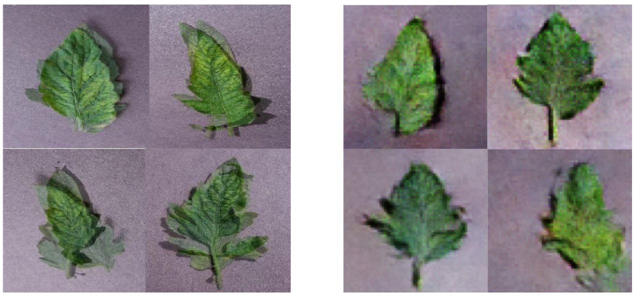
IV. Results and discussions

A. Quality of Augmented Data

The quality of the augmented image plays a significant role in the performance of the classifier. The following issues need to be considered while determining the quality of the generated images:

- Overall quality of the generated images.
- Generated images must represent the desired class.
- Generated images must not be repetitive.

The images generated by SMOTE appear to be little blurry along edges. As this algorithm calculates an instance interpolated between two similar data points, the generated images look like overlapping parts of two images. On the other hand, images generated using GANs are more clear and sharp. The shapes of the leaves are perfect in most cases. In both cases, the generated images have some yellowish texture which is a symptom of the mosaic virus. Based on our qualitative analysis, GAN-based augmentation seems to have a lot of variation than SMOTE-based one in this image domain. Fig. 2 shows some examples of the generated images.



(a) Generated by SMOTE (b) Generated by GAN

Fig. 2. Examples of generated leaf images

B. Evaluation Metrics

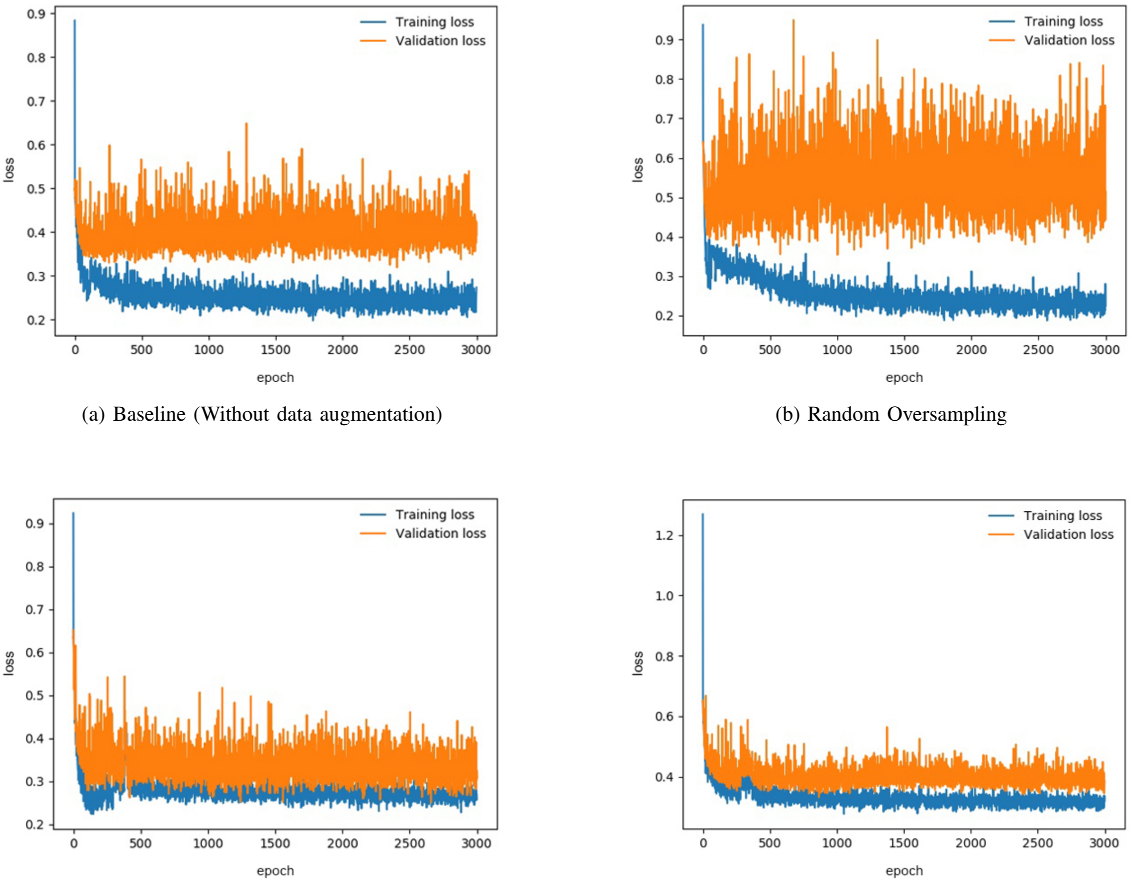
Measuring the performance of a classifier applied to imbalanced data using traditional metrics such as accuracy is difficult because they do not take into account the lower number of instances in each minority class. Rather, precision and recall have been used frequently for assessing the performance of a classifier in such cases. Recall helps to understand the number of misclassification for the positive class, which is infected leaves in our experiment. Another measure is the F1 score or F-measure, which combines precision and recall to give a better indication of the performance. Ranking order metrics such as Area Under the Curve (AUC) measure assess the performance of a classifier over all imbalance ratios and hence provide a summary of the entire range.

C. Performance Analysis

We considered the classifier trained on the imbalanced data set as the baseline. This means that, for the baseline, the training data set does not include any replicated or generated data. As a basis of discussion, we consider the infected class as positive and the healthy class as negative. Table II shows the value of the evaluation metrics for different approaches. These results are based on 5-fold cross-validation, and the highlighted results indicate the best values of each metric. According to the experimental results, SMOTE achieves the best accuracy. However, as discussed earlier, accuracy is not the best metric for an imbalanced data set. In terms of precision, the baseline appears to be the best method among the four candidates. Therefore, the rate of misclassification for negative examples is less for the baseline. As the baseline has more negative samples, this is a reasonable result. In terms of recall, F1 score, and AUC - which are particularly significant for the task in hand - the GAN-based approach outperforms other approaches. Random undersampling failed to perform any good in terms of all the metrics. This can be justified by the loss of information due to the undersampling.

TABLE II
EVALUATION METRICS FOR DIFFERENT APPROACHES (AVERAGE FOR 5-FOLD CROSS-VALIDATION)

Approach Name	Accuracy	Precision	Recall	F1	AUC
Baseline	0.8592	**0.6955**	0.4838	0.5621	0.8830
Random Undersampling	0.8372	0.5990	0.5838	0.5823	0.8557
Random Oversampling	0.8495	0.6067	0.6486	0.6202	0.8666
SMOTE	**0.8663**	0.6647	0.5919	0.6259	0.8877
GAN-based	0.8556	0.6136	**0.6622**	**0.6329**	**0.8925**



(a) Baseline (Without data augmentation)

(b) Random Oversampling

(c) SMOTE

(d) GAN-based

Fig. 3. Visualization of loss function for different data augmentation approaches

The recall for our GAN-based approach is significantly higher than the baseline and slightly better than that for random oversampling. Therefore, we can say the GAN-based approach is good at detecting the positive class which is our prime objective. Also, the F1 score and AUC are greater for the GAN-based approach which means it maintains a good balance in detecting both classes. It is evident from the results that the GAN-based augmentation is more effective to alleviate the influence of the skewed data distribution than any other sampling-based approaches.

We performed a one-sided paired t-test to analyze the significance of the GAN-based approach. The p-values at the 95% level of confidence for the F1 score of the GAN-based approach with respect to the baseline, random undersampling, random oversampling, and SMOTE-based approaches are 0.01, 0.16, 0.31, and 0.39 respectively. The null hypothesis is rejected to significantly differentiate the GAN performance (F1 score) from that of the baseline CNN only, which means that although GAN performance is uniformly slightly better than that of all random and synthetic sampling approaches, this improvement is not statistically significant for the data set used in this experiment. This indicates a need for further large scale experiments with other data sets to conclusively establish the superior precision and recall of the GAN-based approach.

In addition, we noticed that SMOTE and GAN-based approaches help to avoid overfitting. Because overfitting occurs when the hypothesis being evaluated has higher training accuracy than test set accuracy, it may reflect the degree of bias of the model towards the training data set. This can be determined by looking at the values of the loss function for both the training and validation set at the training time. If the training loss is much lower than the validation loss, we can conclude that the model is getting a good handle on classifying the training set, but failing to apply that knowledge on the validation set. Fig. 3 presents the training and validation loss function for different approaches. GAN and SMOTE-based approaches avoid overfitting as they add different samples than what exists in the original data set. The random oversampling-based approach is the one that is more prone to overfitting.

V. Conclusion and Future Work

In this work, we investigated the performance of different data balancing techniques in the plant disease detection domain. Our results demonstrate that the GAN-based approach outperforms random undersampling, random oversampling, and synthetic oversampling approaches. The variation in the samples introduced by a GAN-based approach makes it easier for the classifier to find classification boundaries. Also, the capability of avoiding overfitting indicates the robustness of the GAN-based generative approach over the sampling-based approaches in the plant disease detection task.

Continuing work at present seeks to formulate the task of infected image generation as a style-transfer problem. Further work will focus on developing effective and computationally-efficient algorithms for image data synthesis, constrained by measures of realism with respect to real-world data.

References

[1] S. Verma, A. Chug, A. Singh, S. Sharma, and P. Rajvanshi, “Deep learning based mobile application for plant disease diagnosis,” Applications of Image Processing and Soft Computing Systems in Agriculture, pp. 242–271, 2019.

[2] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, “Cost-sensitive boosting for classification of imbalanced data,” Pattern Recognition, vol. 40, no. 12, pp. 3358-3378, 2007.

[3] N. Japkowicz, “Supervised versus unsupervised binary-learning by feed-forward neural networks,” Machine Learning, vol. 42, pp. 97–122, 2001.

[4] C. X. Ling and V. S. Sheng, “Cost-sensitive learning and the class imbalance problem,” Encyclopedia of machine learning, vol. 2011, pp. 231–235, 2008.

[5] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” Journal of Big Data, vol. 6, no. 1, pp. 1-54, 2019.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[7] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” International Conference on Intelligent Computing. Springer, pp. 878–887, 2005.

[8] I. Goodfellow et al., “Generative adversarial nets,” Advances in Neural Information Processing Systems, pp. 2672–2680, 2014.

[9] S. Mun, S. Park, D. K. Han, and H. Ko, “Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane,” Proc. DCASE, pp. 93-97, 2017.

[10] Z.-H. Zhou and X.-Y. Liu, “Training cost-sensitive neural networks with methods addressing the class imbalance problem,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 1, pp. 63–77, 2006.

[11] S. Lawrence, I. Burns, A. Back, A. C. Tsoi, and C. L. Giles, “Neural network classification and prior class probabilities,” Neural Networks: Tricks of the Trade, vol.1524, pp. 299–313, 1998.

[12] C. X. Ling and C. Li, “Data mining for direct marketing: problems and solutions,” Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, p. 73–79, 1998.

[13] C. Zhang, P. Zhou, C. Li, and L. Liu, “A convolutional neural network for leaves recognition using data augmentation,” IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, pp. 2143–2150, 2015.

[14] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, “Detection of concealed cars in complex cargo x-ray imagery using deep learning,” Journal of X-ray Science and Technology, vol. 25, no. 3, pp. 323–339, 2017.

[15] A. Janowczyk and A. Madabhushi, “Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases,” Journal of Pathology Informatics, vol. 7, 2016.

[16] H. Guo and H. L. Viktor, “Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 30–39, 2004.

[17] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: adaptive synthetic sampling approach for imbalanced learning,” IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328, 2008.

[18] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using gan for improved liver lesion classification,” IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293, 2018.

[19] H.-C. Shin et al., “Medical image synthesis for data augmentation and anonymization using generative adversarialnetworks,” International Workshop on Simulation and Synthesis in Medical Imaging, pp. 1–11, 2018.

[20] Y. Zhu, M. Aoun, M. Krijn, J. Vanschoren, and H. T. Campus, “Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants,” BMVC, p. 324, 2018.

[21] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[22] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan,” arXiv preprint arXiv:1701.07875, 2017.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” IEEE International Conference on Computer Vision (ICCV),vol. 2017, pp. 2242–2251, 2017.

[24] Fanny and T. W. Cenggoro, “Deep learning for imbalance data classification using class expert generative adversarial network,” Procedia Computer Science, vol. 135, pp. 60–67, 2018.

[25] A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” arXiv preprint arXiv:1711.04340, 2017.

[26] J. A. Pandian, G. Geetharamani, and B. Annette,“Data augmentation on plant leaf disease image dataset using image manipulation and deep learning techniques,” IEEE 9th International Conference on Advanced Computing (IACC), pp. 199-204, 2019.

[27] H. Nazki, S. Yoon, A. Fuentes, and D. S. Park, “Unsupervised image translation using adversarial networks for improved plant disease recognition,” Computers and Electronics in Agriculture, vol. 168, p. 105–117, 2020.

[28] D. Hughes and M. Salathe, “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” arXiv preprint arXiv:1511.08060, 2015.

[29] C. Drummond and R. C. Holte, “C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling,” Workshop on Learning from Imbalanced Datasets II, vol. 11, pp. 1-8, 2003.

[30] K. He, X. Zhang, S. Ren, and J. Sun,“Deep residual learning for image recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[31] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi “A Survey of the Recent Architectures of Deep Convolutional Neural Networks,” arXiv preprint arXiv:1511.08060, 2019.

[32] M. Z. Alom et al., “A State-of-the-Art Survey on Deep Learning Theory and Architectures,” Electronics, vol. 8, no. 3, p. 292, 2019.

[33] F. Pedregosa et al., “Scikit-learn: Machine learning in python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.