engineering proceedings

MDPI

Proceeding Paper

Evaluation of Modified FGSM-Based Data Augmentation Method for Convolutional Neural Network-Based Image Classification †

Paulo Monteiro Monson, Vinicius Augusto Dare de Almeida, Gabriel Augusto David, Pedro Oliveira Conceição Junior and Fabio Romano Lofrano Dotto *

Department of Electrical and Computer Engineering, São Paulo University, São Carlos, SP, Brazil; gadavid@usp.br; pedro.oliveiracjr@usp.br; roger@sc.usp.br

* Correspondence: paulo.monson@usp.br

† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

Abstract: Computer vision applications demand for a significant amount of data for effective training and inference in many computer vision tasks. However, data insufficiency situations usually happen due to multiple reasons, resulting in computational models whose performance is inadequate. Traditional data augmentation techniques are presented to solve this overfitting problem, however, their application is not always possible or desirable. In this context, this paper addresses a different data augmentation technique for classification methods based on adversarial images to reduce the impact of sample imbalance utilizing the Fast Gradient Sign Method (FGSM) with added noise to enhance classifier performance. To validate the method, a set of images was used for the classification of diseases in coffee plants due to the soil's lack of nutrients. The results showed an improvement in the model performance for this classification in coffee plants proving the validity of the proposed method, which can be used as an alternative to traditional data augmentation methods.

Keywords: data augmentation; FGSM; computer vision; artificial intelligence; overfitting

Citation: Monson, P.M.; de Almeida, V.A.D.; David, G.A.; Junior, P.O.C.; Dotto, F.R.L. Evaluation of Modified FGSM-Based Data Augmentation Method for Convolutional Neural Network-Based Image Classification. *Eng. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 26 November 2024

1. Introduction

Computer vision applications require a significant amount of data for effective training and inference, such as applications involving classification, image segmentation, and regression. However, situations of insufficient data usually happen for various reasons, such as financial or technical limitations, resulting in computational models whose performance is inadequate, generating a problem known as overfitting in deep learning, which leads to model failure [1–3]. To solve this problem, data augmentation is an effective implicit regularization technique capable of artificially increasing the amount and diversity of available data while preserving the labels for training, allowing more accurate classification models to be generated and combating overfitting [1,2,4–7]. Usually, data augmentation is used to expose a model to invariances in the data domain [4]. Generally, they have more advantages than other regularization techniques [1].

In image classification models, the most common data augmentation methods include simple changes, such as transforming the image size by a few pixels, adjusting colors, contrast, brightness, rotation, cropping, conversion, blurring, sharpening, inversion, displacement, noise generation, region occlusion, etc. [1,4,5]. It can be divided into three categories: geometric transformations, color transformations, and pixel transformations [8]. For example, the work of [1] presents the Albumentations algorithm, based on the concept of fast and flexible image augmentation with various image transformation

operations including basic and image-specific transformation operations such as bounding box and keypoints augmentations. Ref. [4] proposes the AutoAugment method that automates the process of searching for an effective data augmentation policy for a target data set, maximizing model accuracy. Ref. [2] proposes the generation of merged samples through the intelligent combination of resources between two or more samples using Convolutional Neural Networks (CNN), a method known as Smart Augmentation. Ref. [6] presents an adaptive data augmentation method that considers the characteristics of each distinct class without requiring prior domain knowledge. Ref. [7] proposes combining traditional data augmentation strategies to overcome overfitting problems in tree species classification. Ref. [9] proposes a new data augmentation technique called random image cropping and patching (RICAP), which randomly crops four images and corrects them to create a new image. These classical data augmentation approaches are simple, effective, and reliable. However, the changes are limited because the transformation produces a slightly different sample from the original. In another direction from classical data augmentation approaches, the work of [5] presents a method for generating synthetic chest X-ray images of COVID patients through the development of an Auxiliary Classifier Generative Adversarial Network (AGNAN)-based model called CovidGAN. Similarly, ref. [3] proposes a data augmentation high-quality image augmentation (HQIA) method to generate high-quality images of rice leaf diseases based on a dual generative adversarial network (GAN).

Many advances and contributions have been made until now in the data augmentation area. However, as reported by [4], the main priority of the machine learning and computer vision community is to design better network architectures, which means that less emphasis is given to finding better data augmentation methods. Thus, there are still significant research gaps in the area, especially in the application of data augmentation techniques that are not based on classic data augmentation methods. In this context, this paper presents a data augmentation method for image classification based on adversarial images. The proposed method is based on a modification of the Fast Gradient Sign Method (FGSM), a well-known method for generating adversarial images initially proposed to compromise the classifier prediction, to use the noise generated as a data augmentation technique, facilitating the convergence of the classification model.

The combination of FGSM and computer vision technologies holds great promise for advancing agricultural practices, particularly in the monitoring and management of coffee plants. To validate the proposed method, a set of images obtained from the CoLeaf-DB dataset was used to classify diseases through leaf analysis in coffee plants due to nutrient deficiencies during the plant development process. In addition, not only robustness in machine learning can be a contributing factor to innovation. According to [9], some of the contributions rely on automated systems for monitoring coffee plantations, disease detection in coffee plants and real-time decision support. By ensuring that models are robust against adversarial conditions, with added noise, these technologies can significantly improve the efficiency and effectiveness of agricultural operations.

This paper is structured as follows. Section 2 provides an overview of the dataset used in this work, Section 3 describes the modified FGSM method adopted, as well as the neural network characteristics of the classifier employed. The results obtained are analyzed in Section 4 and, finally, Section 5 presents the conclusion of the work.

2. Data Description

To evaluate the proposed approach, the neural network model was trained using the CoLeaf-DB dataset presented by [10]. The dataset presents a set of images of Peruvian coffee leaves known as Catimor, Caturra, and Borbon from coffee plantations located in the province of Jaén, Cajamarca, Peru, with a focus on detecting nutritional deficiencies in this type of plant, which are grouped according to their nutritional deficiencies, including deficiencies in Boron, Iron, Potassium, Calcium, Magnesium, Manganese, Nitrogen, and

others. The images in the dataset were captured in a controlled environment with high luminosity and 0% shadow capture [10].

The Coleaf-DB dataset consists of images divided into ten different conditions, including nutritional deficiency conditions, healthy leaves, and leaves with more than one deficiency present. The samples are divided between the conditions of Healthy leaf (6 samples), Nitrogen (N) (64), Phosphorus (P) (246), Potassium (K) (96), Magnesium (Mg) (79), Boron (B) (101), Manganese (Mn) (83), Calcium (Ca) (162), Iron (Fe) (65) and More than one deficiency (104). The dimensions used are 3000 × 4000 px in jpeg format with a horizontal and vertical resolution of 180 dpi and a depth of 24 bits as present in the dataset.

The leaves were collected by the researchers during the second productive phase, after the ripening phenological phase, after harvesting the first coffee production, and during the pre-flowering period, approximately 4 years after planting [10]. Nutritional deficiencies are classified according to leaf characteristics (internal coloration, chlorosis, and deformation) using the observation method [10]. For example, as described by [10], Figure 1a shows an image of a nutritional deficiency condition for Phosphorus, characterized by lobular internal chlorosis, with irregularly shaped yellow-brown spots and reddish areas on older leaves. Figure 1b shows an image of a nutritional deficiency condition for Boron, which manifests itself in young leaves, which are small, elongated, twisted, wrinkled, with irregular edges, deformed, and with a leathery texture.

(**a**) (**b**)

Figure 1. Leaves with nutritional deficiency condition: (**a**) Phosphorus deficiency; (**b**) Boron deficiency [10].

3. Proposed Method

As shown in the previous section, it can be seen that sample classes such as Iron (65), Manganese (83), and Potassium (96) deficiency are unbalanced concerning classes such as Phosphorus (246) and Calcium (162), for example. This imbalance could lead to lower accuracy for the classes that are sampled less during model training. To correct this imbalance between the classes and, consequently, increase the accuracy of the trained model, a data augmentation method based on FSGM is proposed.

As a form of validation, we first trained the Convolutional neural network (CNN) without data augmentation as presented in the original Coleaf-DB dataset. Subsequently, we performed data augmentation for each class using a modified FSGM method based on the Phosphorus class. Finally, with the classes regularized, the model is retrained and compared with the initial model without data augmentation. This entire sequence is illustrated in Figure 2.

Training is carried out using the deep CNN GoogLeNet, according to [11], which classifies nutritional deficiencies from the Coleaf-DB dataset containing images of coffee leaves as described in Section 2. To carry out the traininvg, the datasets were divided into two subsets, one with 80% of the images for training and 20% of the images for evaluating the neural network. The images were resized to 224 × 224 pixels to fit the GoogLeNet input, normalizing all the pixels in a similar way to that proposed by [10]. The model was trained for 25 epochs with a batch size of 16, 1125 iterations, and a learning rate of 0.0001.

Two metrics were used to evaluate the proposed method: accuracy and receiver operating characteristic (ROC) curves.

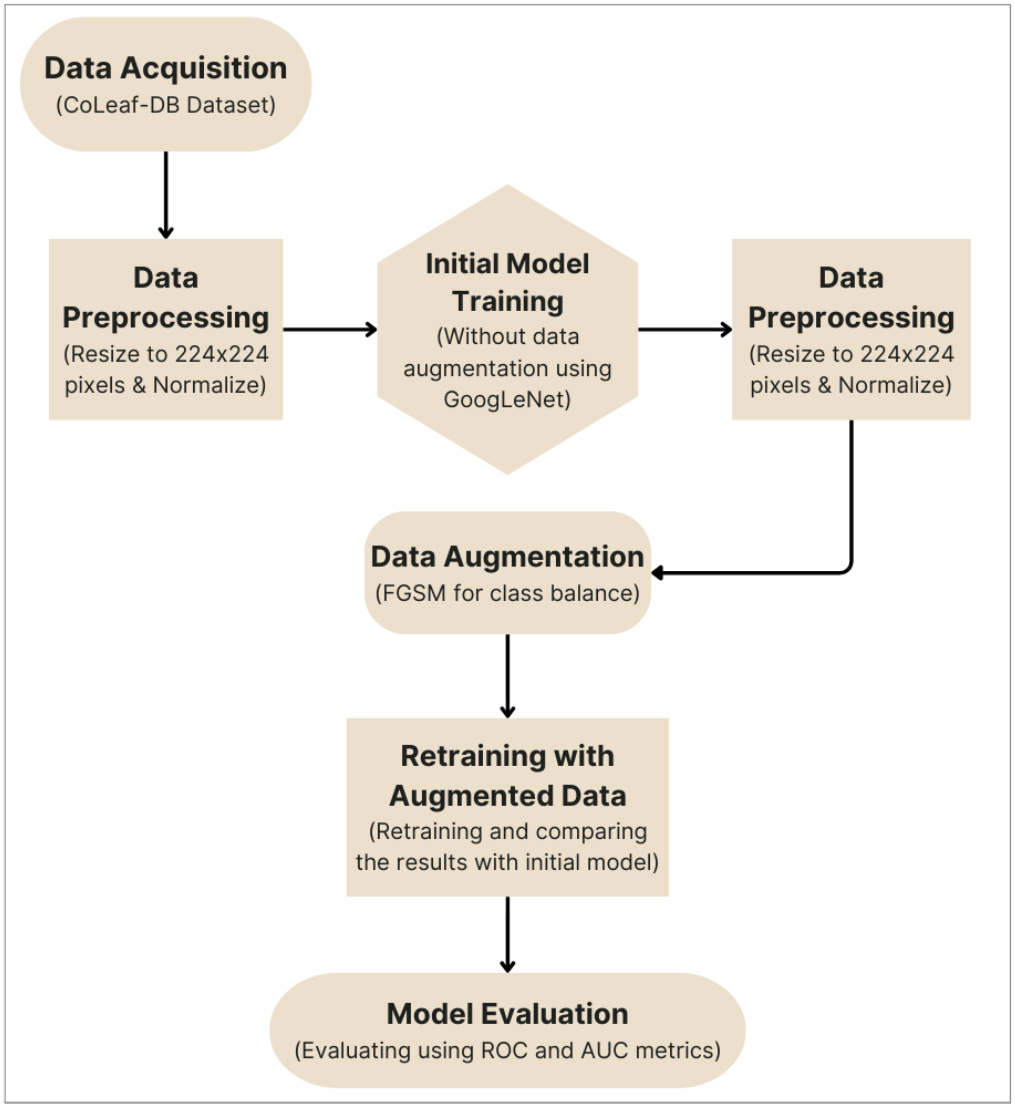

Figure 2. Flowchart illustrating the proposed methodology.

The ROC curves are capable of evaluating the accuracy of the diagnosis generated by a Deep Learning model determining different classification score thresholds by graphically demonstrating the relationship between sensitivity, which is obtained through the true positive rate (TPR), and the specificity of the model, which is obtained through the false positive rate (FRP). The area under a ROC curve (AUC) corresponds to the integral of the curve in TPR values concerning FPR values from zero to one, where the higher the AUC value, the greater the discriminatory capacity of the model, in other words, TPR is high and FRP is low. Thus, using the AUC metric, it is possible to obtain an aggregate performance measure at all possible limits, and the higher the AUC value, the better the performance of the model evaluated in terms of sensitivity and specificity [12–14]. In this work, ROC curves are obtained based on [13].

Data Augmentation Using Fast Gradient Sign Method (FGSM)

Different from traditional data augmentation methods that focus on perceptible changes in image structure through changes in dimensions, cropping, or coloring to

produce new samples, the proposed image classification method is based on adversarial images to reduce the impact of sample imbalance from unbalanced classes. To this end, the feasibility of using a modified version of FGSM capable of using noise to facilitate model convergence is evaluated. FGSM is a well-known simple and efficient adversarial sampling attack method commonly used to add imperceptible disturbances to the input images of neural network-based classifiers, which can lead to compromised prediction. Carefully crafted adversarial samples can lead a well-trained classifier to make the wrong decision and are also a threat to facial recognition methods [15–17]. To generate adversarial images, the method uses the gradients of the model's loss function in relation to the input to determine the direction and magnitude of the disturbances [17].

The method used for the current work modifies FGSM based on a white-box attack which, instead of causing a disturbance in the trained neural network, inserts a modification into the unbalanced samples through noise. In other words, an original image X undergoes a small addition of noise that produces a new image Y not identified with the original image. Noise is generated using the gradients of the loss function concerning the input image, creating an image that maximizes the loss given the input image according to Equation 1,

$$Y = X + \in * sign\ (\nabla_x J(\theta, X, Y)) \quad (1)$$

where θ is the model parameter, J is the loss gradient and $\in$ the noise magnitude applied to the input images. In this way, a new sample containing the input image plus an adversarial attack is produced as described in Figure 3. The generated sample is recognizable as a valid data sample representative of the unbalanced class, so it can improve the model's classification accuracy by increasing convergence by increasing the data provided in the input.

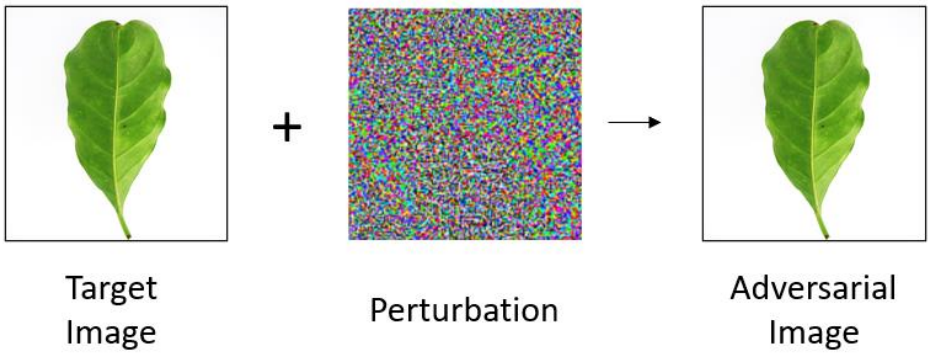

Figure 3. Data Augmentation with modified FGSM.

4. Result and Discussion

Training was carried out using the GoogLeNet deep convolutional neural network architecture, with 80% of the images used for training and 20% reserved for evaluating the model. The model was trained for 25 epochs, using a batch size of 16, and a learning rate of 0.0001. Two different scenarios were considered to assess the impact of adversarial image generation on the accuracy of the classifier.

In the first scenario, without the application of data augmentation, the model achieved an accuracy of 78.33%, with an ROC curve illustrated in Figure 4a. The model's performance is limited by the imbalance of the classes in the original dataset, which is a common problem in many image classification applications. As a result, the model has difficulty generalizing to classes with less representation in the data.

In the second scenario, a data augmentation technique was applied to balance the dataset, using a modified Fast Gradient Sign Method (FGSM) to generate adversarial

images. This approach aimed to increase the diversity of the training set and mitigate the problem of class imbalance. As a result, the model achieved significantly higher accuracy, reaching 96.15%, with the ROC curve shown in Figure 4b. This significant increase in accuracy can be attributed to the greater robustness of the classifier, which was exposed to a wider range of examples during training, including adversarial examples that simulate more difficult variations of the minority classes.

The modified FGSM method proved effective in creating examples that not only increase the size of the data set, but also promote the model's ability to deal with non-trivial variations in the images, substantially improving its generalization power.

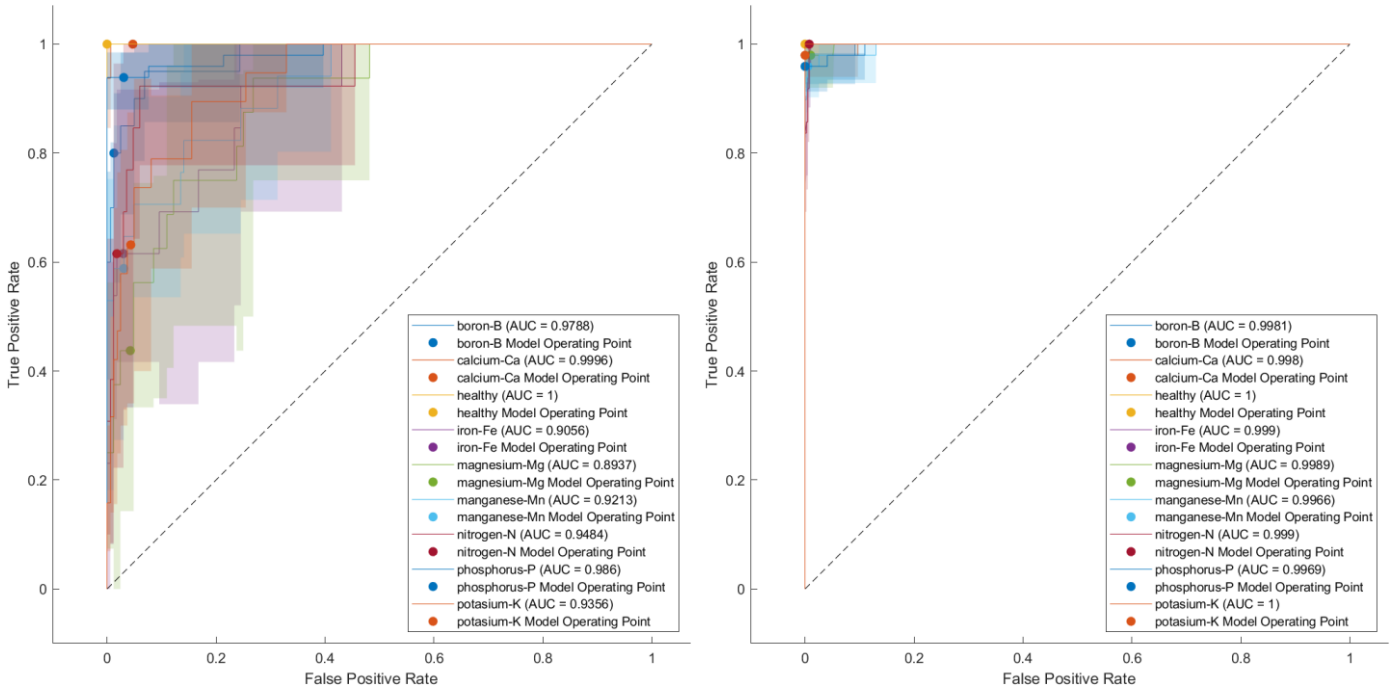

Figure 4. Data Augmentation with modified FGSM. (**a**) ROC curve without FGSM data augmentation. (**b**) ROC curve with FGSM data augmentation.

5. Conclusions

The use of adversarial images generated using the modified FGSM method proved to be an effective strategy for increasing the accuracy of classifiers in scenarios with unbalanced data sets. The data augmentation technique applied to minority classes allowed the GoogLeNet model to achieve significantly higher accuracy, from 78.33% in the scenario without data augmentation to 96.15% in the scenario with the modified FGSM. These results highlight the importance of balancing the data set and introducing challenging examples during training to increase the robustness of the model.

The approach adopted not only improves the model's ability to correctly classify images, but also makes it more resistant to adversarial perturbations, resulting in a more efficient and generalizable classifier. This paves the way for the use of similar techniques in other deep learning applications, especially those where class imbalance and the generation of adversarial examples are critical obstacles.

Author Contributions: Conceptualization, P.M.M.; methodology and software, P.M.M.; validation and formal analysis, P.M.M., G.A.D. and V.A.D.d.A.; investigation, P.M.M., G.A.D. and V.A.D.d.A.; resources, P.M.M. and P.O.C.J.; data curation, P.M.M.; writing—original draft preparation, P.M.M. and G.A.D.; writing—review and editing, P.O.C.J., V.A.D.d.A. and F.R.L.D.; visualization, P.O.C.J. and F.R.L.D.; supervision, P.O.C.J. and F.R.L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by São Paulo Research Foundation (FAPESP), grant #2023/02413-2, the Pro-Rectory of Research and Innovation (USP), grant #22.1.09345.01.2, and the National Council for Scientific and Technological Development (CNPq), grant #140775/2024-2.

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to as they are part of a research project in progress.

Conflicts of Interest:

References

1. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125.
2. Lemley, J.; Bazrafkan, S.; Corcoran, P. Smart Augmentation Learning an Optimal Data Augmentation Strategy. *IEEE Access* **2017**, *5*, 5858–5869.
3. Zhang, Z.; Gao, Q.; Liu, L.; He, Y. A High-Quality Rice Leaf Disease Image Data Augmentation Method Based on a Dual GAN. *IEEE Access* **2023**, *11*, 21176–21191.
4. Cubuk, E.D.; Zoph, B.; Mané, D.; Vasudevan, V.; Le, Q.V. AutoAugment: Learning Augmentation Strategies From Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 113–123.
5. Waheed, A.; Goyal, M.; Gupta, D.; Khanna, A.; Al-Turjman, F.; Pinheiro.CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved COVID-19 Detection. *IEEE Access* **2020**, *8*, 91916–91923.
6. Yoo, J.; Kang, S. Class-Adaptive Data Augmentation for Image Classification. *IEEE Access* **2023**, *11*, 26393–26402.
7. Chen, L.; Wei, Y.; Yao, Z.; Chen, E.; Zhang, X. Data Augmentation in Prototypical Networks for Forest Tree Species Classification Using Airborne Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16.
8. Liu, B.; Li, L.; Xiao, Q.; Ni, W.; Yang, Z. Remote Sensing Fine-Grained Ship Data Augmentation Pipeline With Local-Aware Progressive Image-to-Image Translation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16.
9. Takahashi, R.; Matsubara, T.; Uehara, K. Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 2917–2931.
10. Tuesta-Monteza, V.A.; Mejia-Cabrera, H.I.; Arcila-Diaz, J. CoLeaf-DB: Peruvian coffee leaf images dataset for coffee leaf nutritional deficiencies detection and classification. *Data Brief* **2023**, *48*, 109226.
11. Googlenet Help Center. Available online: https://www.mathworks.com/help/deeplearning/ref/googlenet.html (accessed on 11 September 2024).
12. Fan, J.; Upadhye, S.; Worster, A. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* **2006**, *8*, 19–20.
13. Compare Deep Learning Models Using ROC Curves Help Center. Available online: https://www.mathworks.com/help/deeplearning/ug/compare-deep-learning-models-using-ROC-curves.html#CompareDeepLearningModelsUsingROCCurvesExample-9 (accessed on).
14. Carter, J.V.; Pan, J.; Rai, S.N.; Galandiuk, S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery* **2016**, *159*, 1638–1645.
15. Liu, Y.; Mao, S.; Mei, X.; Yang, T.; Zhao, X. Sensitivity of Adversarial Perturbation in Fast Gradient Sign Method. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 433–436.
16. Arbena Musa, K.V.; Rexha, B. Attack Analysis of Face Recognition Authentication Systems Using Fast Gradient Sign Method. *Appl. Artif. Intell.* **2021**, *35*, 1346–1360.
17. Naqvi, S.M.A.; Shabaz, M.; Khan, M.A.; Hassan, S.I. Adversarial Attacks on Visual Objects Using the Fast Gradient Sign Method. *J. Grid Comput.* **2023**, *21*, 52.
18. Tensorflow Adversarial FGSM. Available online: https://www.tensorflow.org/tutorials/generative/adversarial_fgsm (accessed on 21 August 2024).