applied sciences

MDPI

Article

FocalMatch: Mitigating Class Imbalance of Pseudo Labels in Semi-Supervised Learning

Yongkun Deng, Chenghao Zhang, Nan Yang and Huaming Chen *

School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2008, Australia
* Correspondence: huaming.chen@sydney.edu.au

Abstract: Semi-supervised learning (SSL) is a popular research area in machine learning which utilizes both labeled and unlabeled data. As an important method for the generation of artificial hard labels for unlabeled data, the pseudo-labeling method is introduced by applying a high and fixed threshold in most state-of-the-art SSL models. However, early models prefer certain classes that are easy to learn, which results in a high-skewed class imbalance in the generated hard labels. The class imbalance will lead to less effective learning of other minority classes and slower convergence for the training model. The aim of this paper is to mitigate the performance degradation caused by class imbalance and gradually reduce the class imbalance in the unsupervised part. To achieve this objective, we propose FocalMatch, a novel SSL method that combines FixMatch and focal loss. Our contribution of FocalMatch adjusts the loss weight of various data depending on how well their predictions match up with their pseudo labels, which can accelerate system learning and model convergence and achieve state-of-the-art performance on several semi-supervised learning benchmarks. Particularly, its effectiveness is demonstrated with the dataset that has extremely limited labeled data.

Keywords: semi-supervised learning; class imbalance; data privacy

check for updates

Citation: Deng, Y.; Zhang, C.; Yang, N.; Chen, H. FocalMatch: Mitigating Class Imbalance of Pseudo Labels in Semi-Supervised Learning. *Appl. Sci.* **2022**, *12*, 10623. https://doi.org/10.3390/app122010623

Academic Editor: Xiaobo Zhou

Received: 30 September 2022
Accepted: 19 October 2022
Published: 20 October 2022

1. Introduction

Machine learning (ML) is one of the most important and popular fields in artificial intelligence. The core concept of ML is about the data-driven model [1]. It has evolved rapidly in recent decades due to the explosive growth in the amount of available data and the increase in computational power. The main feature of machine learning is to automatically improve performance through experience [1,2]. Due to this feature, machine learning has rapidly become the fundamental technique of many modern applications, including computer vision, natural language processing (NLP), fraud detection, medical analysis (both physically and mentally), the agriculture industry, the energy sector, mechanical engineering, network security, etc. [3–12]. Machine learning has spawned many branches, such as supervised learning and unsupervised learning. The main difference between supervised and unsupervised learning is whether or not labeled data is used. Supervised learning usually achieves better performance than unsupervised learning on the same task by leveraging valuable information from labeled data.

Although the amount of available data has dramatically increased over the last few decades, labeled data still represents a small fraction. Labeling data is either complicated, costly (time-consuming and/or expensive), or both. For example, some samples can only be labeled by the expert, such as the medical analysis. Furthermore, as the number of available data increases, people are more and more concerned about data privacy. Even if a significant number of labels have previously been collected, it still remains unknown if the labels will be available for model learning, or if more attention will be needed to handle the data [13]. Therefore, it is critical for the model to generate artificial labels for unlabeled data instead of manually labeling the data due to privacy concerns. The lack of labeled data has given rise to a different research area called semi-supervised learning (SSL). SSL uses datasets that contain only a small amount of labeled data and a considerable amount of unlabeled data. By leveraging massive volumes of unlabeled data, SSL can significantly improve the model performance with much less labeled data.

In recently proposed semi-supervised learning frameworks, pseudo-labeling [14] has been widely used. Pseudo-labeling is based on the assumption that the learning model should generate hard labels for unlabeled data on its own (i.e., through the predicted class distributions) and then use these generated labels as targets for unlabeled data. FixMatch [15] is a state-of-the-art semi-supervised learning method that produces pseudo (one-hot) labels from weakly augmented samples and utilizes the cross-entropy loss to ensure the consistencies between pseudo labels and the predictions of the same samples (strongly augmented). The generated pseudo labels of unlabeled data help FixMatch to achieve entropy minimization [16] in the unsupervised learning part. However, FixMatch and other semi-supervised learning methods that utilize pseudo-labeling have a tendency to assign pseudo labels to certain classes, particularly in the initial stages of the training process. This introduces the problem of class imbalance. The overall loss can be dominated by classes with a large number of pseudo labels. Hence, the model can only learn useful information from the majority classes of pseudo labels and ignores other classes. As a result, most classes of pseudo labels usually become the easy samples (with high prediction accuracy), and other classes become the hard samples. This further exacerbates the class imbalance issue since the model rarely gives high prediction confidence for hard samples. Hence, the pseudo labels of hard samples are less likely to be considered valid pseudo labels. Therefore, the corresponding loss item will not be added to the overall unsupervised loss.

In this paper, we propose a new method FocalMatch that combines FixMatch with focal loss [17] to address the problem of class imbalance that occurs in the unsupervised learning part. In light of the systematic analysis of the unsupervised learning part in FixMatch, the focal loss is proposed to enhance the learning process of FocalMatch by providing the capability of loss contribution adjustment for different samples. This is particularly achieved by the automatic evaluation of how close their predictions are to their pseudo labels. As a result, the overall unsupervised loss will not be overwhelmed by a large number of easy samples. The workflow of FocalMatch is shown in Figure 1. Experiments show that FocalMatch significantly reduces the difference in the total number of pseudo labels generated for each class and provides a smoother learning curve in comparison with other state-of-the-art models.

In brief, the main contributions of this work can be summarized in the following sections:

- We propose **FocalMatch**, a novel but simple semi-supervised learning method that combines FixMatch and focal loss, which effectively mitigates the performance degradation caused by class imbalance and gradually reduces class imbalance that occurs in the unsupervised learning part when generating pseudo labels.
- FocalMatch adjusts the loss weights of different unlabeled data based on the proximity of their predictions to their pseudo labels. Hence, the loss will not be overwhelmed by easy samples. Thus, the model can effectively learn valuable information from all classes.
- FocalMatch outperforms most state-of-the-art semi-supervised learning methods on several benchmarks, especially when the quantity of labeled data is severely limited. Experiments show that FocalMatch significantly reduces the difference between the number of pseudo labels generated for each class. FocalMatch also has a smoother training curve and converges faster compared to FixMatch.

For the following sections, we first discuss the related work in Section 2, which includes semi-supervised learning and class imbalance. Next, We discuss the materials and methods of our proposed FocalMatch in Section 3. We then introduce the experiments we have performed, including the experiment setting and baseline methods used in Section 4. We compare the experiment outcomes in Section 5 and conduct an ablation study to investigate the effectiveness of our method in Section 6. Finally, we give a summary of this paper in Section 7, followed by Appendix A, which gives a detailed experiment setting.

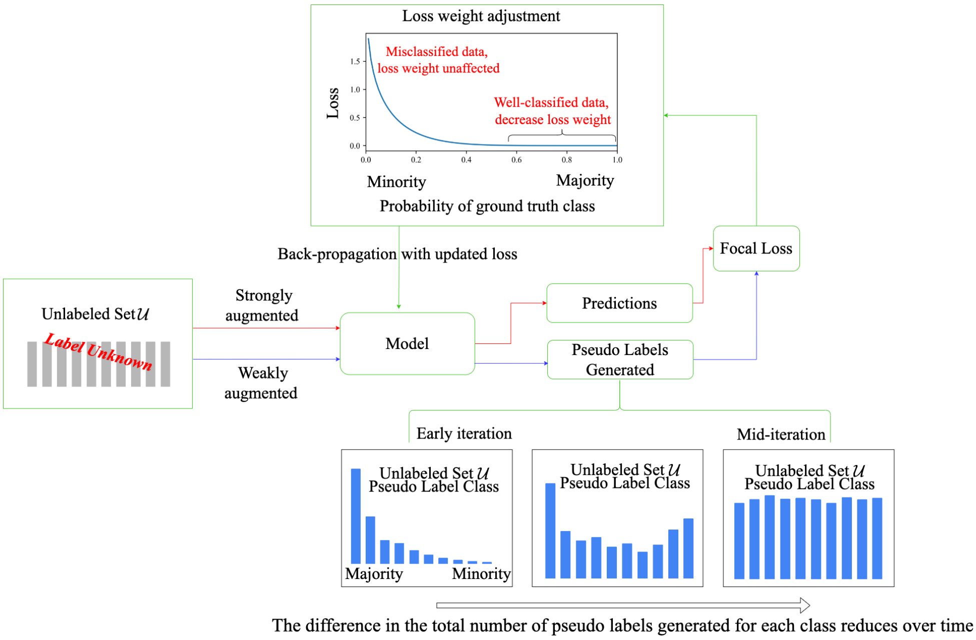


Figure 1. Framework of FocalMatch. Unlabeled data is fed into the model, and the model generates pseudo labels based on the weakly-augmented unlabeled data. At early iterations, the model prefers certain classes for pseudo labels, which causes severe class imbalance. Then, the model makes predictions on the same unlabeled data but with strong augmentation. The model calculates the loss between predictions and pseudo labels via focal loss [17] to ensure consistency. Focal loss adjusts the weight of different data based on how close their predictions are to their pseudo labels. For well-classified data (i.e., the majority of pseudo labels), their loss contribution is reduced. Therefore, the model can rapidly eliminate class imbalance at early iterations.

2. Related Work

2.1. Semi-Supervised Learning

Semi-supervised learning (SSL) is a popular machine learning technique that combines supervised and unsupervised learning, i.e., both labeled and unlabeled data are used for training. Typically, semi-supervised learning uses smaller labeled datasets because one of the main ideas behind SSL is to address the problem of insufficient labeled data [18,19]. Semi-supervised learning has become increasingly important in many areas, for example, medical image analysis [20] and natural language processing [21] since labeled data in these scenarios is usually either expensive or hard to obtain. A detailed introduction to semi-supervised learning can be found in [22].

In the study of semi-supervised learning, consistency regularization [23] is a commonly used technique. It assumes that distorted versions of the same input sample should yield similar predictions from the model. Consistency regularization has been applied in many state-of-the-art SSL methods. In order to generate distorted samples, data augmentation is usually beneficial. ReMixMatch [24] and UDA [25] both use strong data augmentations to improve the consistency regularization between different versions of images. Pseudo-labeling [14] is another widely used technique in SSL that generates hard (one-hot) artificial labels for unlabeled data from the model predictions. In several recently proposed SSL methods, pseudo-labeling is combined with consistency regularization. FixMatch [15] generates one-hot pseudo labels from predictions on weakly-augmented data with a pre-defined high threshold and ensures consistency against strongly-augmented data.

2.2. Class Imbalance

In the majority of machine learning, the training dataset is considered well-balanced (i.e., each class contains a similar number of samples) [26]. However, class distribution is usually imbalanced (i.e., some classes contain considerably more samples than other classes) in real-world scenarios, including fraud detection [27], medical diagnosis [28], software failure prediction [29], etc. The class imbalance problem has a detrimental impact on machine learning, such as the convergence and generalization ability of the model [30]. A more extensive introduction of class imbalance is provided in [31].

As [32] suggests, the mainstream solution to class imbalance can be summarized into two approaches: data-level methods and algorithmic-level methods. Data-level methods aim to eliminate class imbalance by modifying the training dataset, such as oversampling [33] (randomly selects samples from minority class and duplicates them) and undersampling [34] (randomly selects samples from majority class and discards them). However, these sampling methods may degrade the final performance, such as causing overfitting [35]. Algorithmic-level methods, instead, aim to address the class imbalance problem by modifying the learning algorithms. Threshold moving is one of the most famous algorithmic-level methods. The main idea behind threshold moving is to adjust the output (e.g., weights) of the model continuously to accommodate the imbalanced distributions of samples [36]. In the machine learning area, both algorithmic-level and data-level methods are commonly used. For example, ref. [37] proposes a novel hybrid sampling method to address class imbalance based on generative adversarial network.

Our proposed FocalMatch combines FixMatch (the state-of-the-art semi-supervised learning framework) and focal loss (an algorithmic-level method) to address the class imbalance problem that occurs when generating pseudo labels in the unsupervised learning part. It surpasses the traditional cross-entropy loss function used by the model instead of modifying the original dataset, which is more in line with the application of semi-supervised learning and ensures data privacy. FocalMatch adjusts the loss weights of different samples based on how close their predictions are to their pseudo labels. It decreases the loss weights of easy samples so that the overall unsupervised loss will not be overwhelmed by easy samples. A detailed ablation study to investigate the effectiveness of FocalMatch is discussed in Section 6. Comparing FocalMatch to other state-of-the-art models (including Π model [38], Mean Teacher [39], MixMatch [40], ReMixMatch [24], UDA [25], and FixMatch [15]), our experiments reveal that FocalMatch significantly reduces the difference in the total number of pseudo labels generated for each class and has a more gradual learning curve. FocalMatch surpasses most state-of-the-art semi-supervised learning algorithms on several benchmarks, particularly when the amount of labeled data is severely constrained.

3. Materials and Methods

3.1. Consistency Regularization and Pseudo-Labeling

Consistency regularization [23] is one of the most popular ideas in semi-supervised learning. It assumes that the distortion of a sample should not have an impact on the predictions of the model. [15] formulates the consistency loss as Equation (1):

$$\sum_{b=1}^{\mu B} \|p_m(y|\alpha(u_b)) - p_m(y|\alpha(u_b))\|_2^2, \tag{1}$$

where μ is the relative size of unlabeled data to labeled data, B is the batch size of labeled data, u_b is an unlabeled data, α is a data augmentation function, $p_m(y|\alpha(u_b))$ is the predictions (soft label) from the model on augmented u_b. Since both α and p_m are stochastic, the two items in Equation (1) are different. $\|p_m - p_m\|_2^2$ is used to measure the distance between the aforementioned two predictions.

In modern semi-supervised learning techniques, pseudo-labeling [14] is highly related to consistency regularization. It suggests that the model should generate artificial labels for the unlabeled data.The authors in [15] give the definition of the loss function of pseudo-labeling as Equation (2):

$$\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}(max(p_m(y|u_b)) \geq \tau)H(\hat{p}_m(y|u_b), p_m(y|u_b)), \tag{2}$$

where $p_m(y|(u_b))$ is the predictions (soft label) from the model on u_b, whereas $\hat{p}_m$ is the one-hot pseudo label obtained from p_m. τ is the hyperparameter that defines the threshold, and H is the cross-entropy loss.

3.2. FixMatch

FixMatch [15] is a recently proposed state-of-the-art semi-supervised learning algorithm. In FixMatch, the learning model first makes predictions on a weakly-augmented sample with probability distributions of each class (soft label). If the probability of a specific class exceeds the pre-defined threshold, that class will be adapted as a pseudo label (one-hot label) of the sample. Secondly, the learning model makes predictions on the same strongly-augmented sample and uses a cross-entropy loss to ensure the consistency regularization between the predictions of the strongly-augmented sample and the pseudo label. FixMatch combines consistency regularization and pseudo-labeling. Hence, Equation (2) can be re-formulated as:

$$\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}(max(p_m(y|\alpha(u_b))) \geq \tau)H(\hat{p}_m(y|\alpha(u_b)), p_m(y|\mathcal{A}(u_b))). \tag{3}$$

Equation (3) is simply the combination of Equations (1) and (2) provided in [15], where α refers to the weak data augmentation and $\mathcal{A}$ refers to the strong data augmentation. τ is the threshold, and H is the cross-entropy loss. In FixMatch, the one-hot pseudo label ($\hat{p}_m$) is obtained by applying arg max to the soft label (p_m) of the weakly-augmented image.

3.3. FocalMatch

The standard cross-entropy loss measures the distance between two probability distributions (i.e., the ground truth and the prediction). The lower the cross-entropy loss, the closer the two probability distributions (i.e., the prediction is closer to the ground truth). Due to this property, the cross-entropy loss is widely used in classification tasks. However, the standard cross-entropy loss treats the loss contribution of each class equally. This is generally acceptable in class balance situations. However, in class imbalance situations (i.e., the sample sizes of some classes are significantly larger than others), the loss from majority classes can dominate the overall cross-entropy loss. As a result, the model can hardly learn useful information from the minority class, which will further decrease the prediction accuracy of the minority class. Moreover, because of the difference in sample size, even if the loss of a single sample from the minority class is higher than that of a sample from the majority class (due to the lower accuracy), the total loss from the majority class may still dominate the overall cross-entropy loss. Equation (4) shows the standard cross-entropy loss. For simplicity, we use the binary classification case in the following sections:

$$CE(p_t) = -log(p_t), p_t = \begin{cases} p & if\ \ y = 1 \\ 1 - p & otherwise, \end{cases} \tag{4}$$

where y is the ground truth label, and p is the predicted probability distribution of the sample. In order to solve the class imbalance problem, ref. [17] proposed an improved version of the cross-entropy loss called focal loss. The main idea behind focal loss is to adjust the contributions of different samples. The focal loss adds a modulating factor to the standard cross-entropy loss:

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t), p_t = \begin{cases} p & if \ \ y = 1 \\ 1-p & otherwise, \end{cases} \tag{5}$$

where $(1-p_t)^{\gamma}$ is the modulating factor, and γ is a hyperparameter that is greater than or equal to 0. The existence of the modulating factor can help the model to adjust the weights of different samples. In the correct classification scenarios, p_t is closer to 1, which means that the modulating factor is closer to 0. As a result, the loss weights of these samples (easy samples) are reduced. In the misclassified scenarios, p_t is closer to 0, which means that the modulating factor is closer to 1. Therefore, the loss weights of these samples (hard samples) keep unchanged. Even if the number of easy samples is much higher than that of hard samples, the loss from hard samples will still account for a significant portion of the total loss due to the weight adjustment, and the model can learn valuable information from hard samples so that the model performance can be further improved.

Following the method described in [17], Figure 2 shows the loss curves with different γ values. Focal loss adjusts the contributions of easy samples. As γ rises, the model adjusts the loss contributions more strongly.

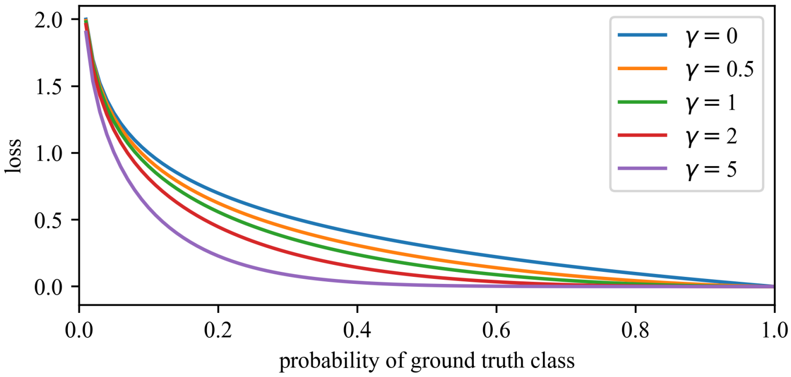


Figure 2. The loss curves of cross-entropy loss and focal loss with different γ values. Note that when γ = 0, the focal loss is identical to cross-entropy loss.

In FixMatch [15], the cross-entropy loss is used between the pseudo labels and the predictions of strongly-augmented images. However, during the training phase, we found that the learning model tends to generate class-specific pseudo labels (e.g., the number of pseudo labels for cats may be much higher than the number of pseudo labels for airplanes). As a result, a class imbalance of pseudo labels occurs in the unsupervised learning phase. Detailed information on the number of pseudo labels generated is provided in Section 6. In this scenario, the cross-entropy loss is no longer optimal. To address this problem, we propose our new method, FocalMatch, that combines FixMatch and focal loss [17]. We replace the cross-entropy loss with the focal loss for the unsupervised learning part so that the model can focus more on the minority pseudo labels. Therefore, the unsupervised loss L_u of our method can be formulated as:

$$L_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(max(p_m(y|\alpha(u_b))) \geq \tau) FL(p_t), \tag{6}$$

$$FL(p_t) = -(1-p_t)^{\gamma} log(p_t), p_t = \begin{cases} p_m(y|\mathcal{A}(u_b)) & if \ \ \hat{p}_m(y|\alpha(u_b)) = 1 \\ 1-p_m(y|\mathcal{A}(u_b)) & otherwise, \end{cases} \tag{7}$$

where u_b is unlabeled data, α and $\mathcal{A}$ refer to weak data augmentation and strong data augmentation, respectively. $p_m(y|\alpha(u_b))$ and $p_m(y|\mathcal{A}(u_b))$ are the predicted probability distributions on weakly-augmented and strongly-augmented samples, respectively. The former one is also the soft label of the unlabeled sample. In addition, $\hat{p}_m(y|\alpha(u_b)$ are the pseudo labels generated from soft labels where the confidence of a specific class is higher than τ (i.e., the hyperparameter that defines the threshold). The supervised loss L_s is the same as FixMatch [15]:

$$L_s = \frac{1}{B}\sum_{b=1}^{B} H(y_b, p_m(y|\alpha(x_b))), \tag{8}$$

where x_b is the labeled data, y_b is the corresponding label, and H is the standard cross-entropy loss. The overall loss of FocalMatch is:

$$Loss = Ls + \lambda L_u \tag{9}$$

where λ is another hyperparameter that defines the weight of unsupervised loss Lu. The detailed algorithm of our method is shown in Algorithm 1.

Algorithm 1 FocalMatch Algorithm

```
Input: $\mathcal{X} = \{(x_b, y_b) : b \in (1, \dots, B)\}, \mathcal{U} = \{u_b : b \in (1, \dots, \mu B)\}$    ▷ $\mathcal{X}$ is labeled set, $\mathcal{U}$ is unlabeled set
for i = 1 to max iteration do
    $L_s$, $L_u \leftarrow 0$
    for b = 1 to B do
        labeled_prediction $\leftarrow p_m(y|\alpha(x_b))$    ▷ Class distribution on weakly augmented $x_b$
        $L_s$ += cross_entropy_loss($y_b$, labeled_prediction)    ▷ Equation (8)
    end for
    for b = 1 to $\mu B$ do
        if max($p_m(y|\alpha(u_b))$) > $\tau$ then    ▷ Confidence of a class is higher than $\tau$
            pseudo_label $\leftarrow$ argmax($p_m(y|\alpha(u_b))$)
            unlabeled_prediction $\leftarrow p_m(y|\mathcal{A}(u_b))$    ▷ Class distribution on strongly augmented $u_b$
            $L_u$ += focal_loss(pseudo_label, unlabeled_prediction)    ▷ Equation (6)
        end if
    end for
    $L_s \leftarrow \frac{L_s}{B}$, $L_u \leftarrow \frac{L_u}{\mu B}$
    Loss $\leftarrow L_s + \lambda L_u$
    Model Update
end for
```

4. Experimental Section

4.1. Experimental Setup

4.1.1. Datasets

We conduct our experiments on three datasets: CIFAR-10 [41], CIFAR-100 [41], and SVHN [42]. CIFAR-10 and CIFAR-100 both contain 60,000 32 × 32 color (3 channels) images with 50,000 training images and 10,000 testing images. CIFAR-10 contains 10 classes with 6000 images per class, whereas CIFAR-100 contains 100 classes with 600 images per class. SVHN contains over 600,000 32 × 32 color (3 channels) images with 10 classes (digits). Considering semi-supervised learning's actual application circumstances, all experiments are undertaken with a very limited quantity of labeled data. We apply two settings to all datasets, 4 labels per class and 10 labels per class (i.e., for CIFAR-10 and SVHN: 40 labels and 100 labels in total, respectively; for CIFAR-100: 400 labels and 1000 labels in total).

4.1.2. Baselines

During the experiment, we compare FocalMatch with several sate-of-the-art semi-supervised learning models on the aforementioned three datasets: Π model [38], Mean Teacher [39], MixMatch [40], ReMixMatch [24], UDA [25], and FixMatch [15].

4.1.3. Setup

To fairly compare our approach with other SSL methods, all the experiments are implemented with PyTorch [43] using the same codebase of TorchSSL [44]. We use similar hyperparameter settings as [15,44]: all baseline methods use Wide ResNet-28-2 [45] as the backbone network, batch size 64 for labeled data, standard stochastic gradient descent with a momentum of 0.9 as the optimizer [46,47], initial learning rate of 0.03 with cosine learning rate decay [48]. There are other hyperparameters that are method-dependent: μ (unlabeled data to labeled data ratio), τ (threshold of generating pseudo labels), λ (weight of unsupervised loss), temperature (for sharpening soft labels). As suggested by [44], all method-dependent hyperparameters follow the original papers. Some hyperparameters only belong to specific methods (e.g., the weight for distribution matching loss in ReMixMatch); these parameters also follow the original papers. In addition, [15] emphasizes the importance of combining weak and strong data augmentation. We use random horizontal flip (with 50% probability) and random crop (crop size 32) for weak data augmentation on the datasets mentioned above. For strong data augmentation, we use RandAugment [49]. In Appendix A, a comprehensive set of hyperparameters is presented.

5. Results

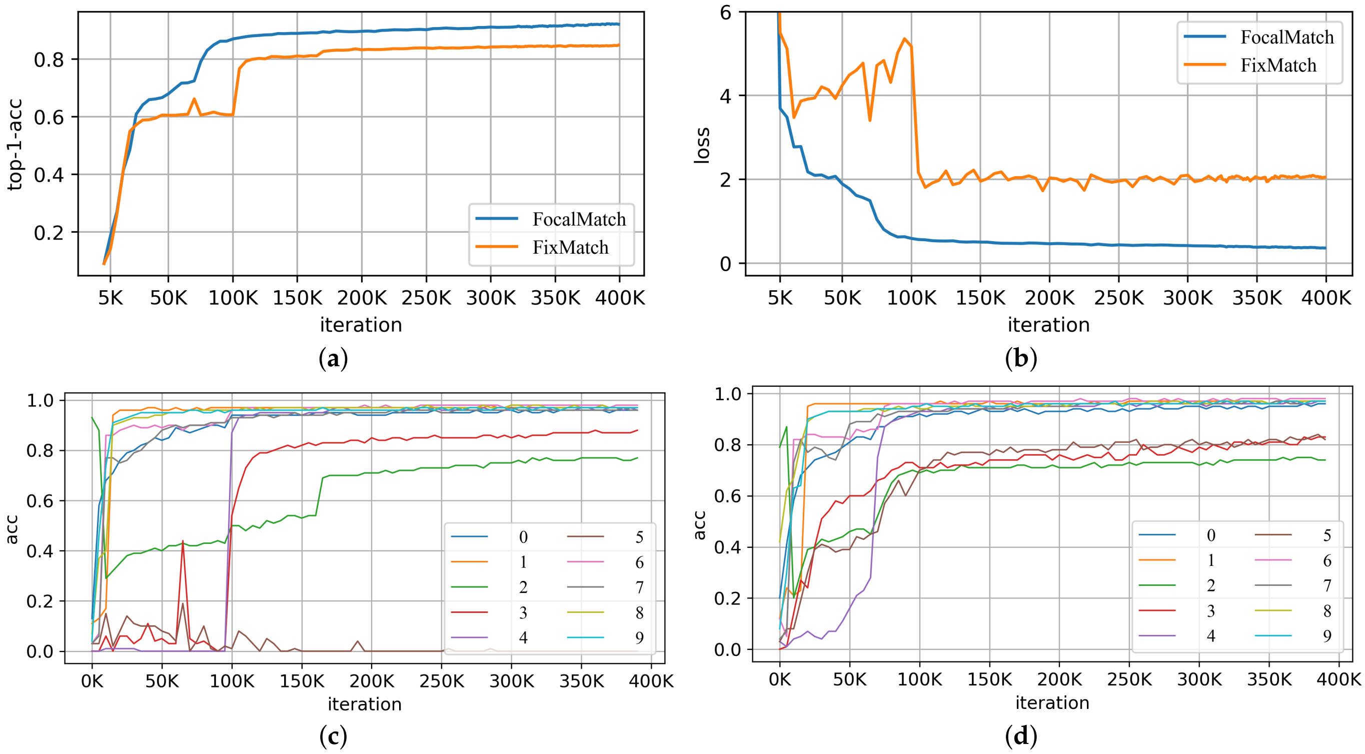
Our experiments use top-1 classification accuracy as the evaluation metric for all baseline methods and FocalMatch. The result is shown in Table 1. It shows that FocalMatch outperforms all baseline methods on most of the benchmarks. FocalMatch performs particularly well when the number of labeled data is extremely small (i.e., four labels per class). However, FocalMatch does not perform as well as expected on SVHN with 10 labels per class. FixMatch and UDA outperform their accuracy by around 0.2% and 0.5%, respectively. We believe this is due to the simplicity of the SVHN dataset. When the amount of labeled data is extremely small (i.e., four labels per class), the model is not able to produce valid pseudo labels evenly for all classes since the overall prediction confidence is not high enough. This causes a severe class imbalance problem which can be effectively alleviated by FocalMatch. When we increase the number of labeled data in the SVHN experiment (i.e., 10 labels per class), the model is confident enough to generate pseudo labels evenly, and the accuracy of each class is relatively high. Therefore, the loss contribution adjustment of FocalMatch can slow down the learning of the model. This may also explain the reason why FocalMatch achieves significant performance improvement when the number of labeled data is extremely small, whereas the performance improvement of FocalMatch reduces as the number of labeled data increases (i.e., less challenging to classify).

Table 1. Accuracy Comparison in Different Methods.

	CIFAR-10		**CIFAR-100**		**SVHN**	
	40 Labels	**100 Labels**	**400 Labels**	**1000 Labels**	**40 Labels**	**100 Labels**
Π-Model	22.92 ± 1.26	34.98 ± 1.53	12.34 ± 1.37	26.17 ± 2.31	31.33 ± 0.75	78.88± 0.32
Mean Teacher	27.13 ± 1.31	44.41 ± 2.42	14.31 ± 1.53	29.50 ± 3.67	64.04 ± 3.18	79.83 ± 4.41
MixMatch	61.64 ± 3.47	79.24 ± 2.63	22.96 ± 2.16	44.62 ± 2.47	71.4 2± 6.37	96.09 ± 0.29
ReMixMatch	90.26 ± 1.41	91.96 ± 0.75	44.03 ± 1.33	57.49 ± 0.95	76.27 ± 9.54	94.18 ± 0.48
UDA	85.31 ± 4.37	92.33 ± 0.23	43.17 ± 1.41	57.85 ± 0.71	95.36 ± 3.47	97.92 ± 0.04
FixMatch	89.94 ± 0.34	92.87 ± 0.17	43.38 ± 1.09	57.99 ± 0.69	96.79 ± 1.42	97.71 ± 0.15
FocalMatch	92.29 ± 0.27	93.09 ± 0.15	46.02 ± 0.86	58.70 ± 0.31	97.37 ± 1.25	97.53 ± 0.09

FocalMatch has substantially extended the learning ability of FixMatch by resolving the latent class imbalance issue. Our method not only outperforms FixMatch in terms of classification accuracy on CIFAR-10, CIFAR100, and SVHN (except when the number of labels per class is 10 on the SVHN dataset) but also speeds up the convergence of the model. We compare the convergence speed of our method and FixMatch in terms of the overall top-1 accuracy and loss in Figure 3a,b. It is obvious that the loss curve of FocalMatch is smoother and converges faster compared to FixMatch. Following the approach described in [44], we also compare the accuracy of FixMatch and FocalMatch for each class in Figure 3c,d. It is observed that there is a large gap between the accuracy of each class in FixMatch, which is due to the class imbalance on the pseudo labels generated in the unsupervised learning part. The total unsupervised loss of FixMatch tends to be dominated by classes with a large number of pseudo labels instead of learning from the overall unlabeled data. This could explain why the accuracy of some classes is appreciably lower than that of other classes or even not improving at all (e.g., class 5).

On the other hand, the accuracy for each class of FocalMatch evenly increases with no significant differences between the classes. It demonstrates that FocalMatch is able to extract useful features from all classes uniformly instead of a specific class. We conduct an ablation study in Section 6 to investigate the effect of focal loss on addressing the class imbalance problem of pseudo labels.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 3. The comparison of accuarcy and loss between FixMatch and FocalMatch on CIFAR-10 dataset. The numbers in legends of (**c**,**d**) represent the 10 classes in CIFAR-10 dataset. (**a**) top1 accuracy. (**b**) loss. (**c**) Accuracy for each class of FixMatch. (**d**) Accuracy for each class of FocalMatch.

6. Discussion and Ablation Study

Our method simply combines FixMatch [15] and focal loss [17]. The main idea behind focal loss is to address the problem of class imbalance. The class imbalance can make it difficult for the model to learn useful information from the minority class. Focal loss is commonly used in the object detection area since the number of images of target classes is much smaller than that of background classes. We find that focal loss is also useful for image classification when the amount of labels is imbalanced. Our experiments set an equivalent number of labels for each class. Therefore, the class imbalance problem can hardly occur in the supervised learning part. However, in FixMatch, pseudo labels are self-generated from unlabeled data. Therefore, the class imbalance can happen on the generated pseudo labels, which will affect the unsupervised learning of the model. To investigate the effectiveness of focal loss, we conduct an ablation study using different γ values in the focal loss.

Figure 4 demonstrates the number of pseudo labels generated (i.e., the predicted confidence of a specific class in the soft label is greater than τ) by the model for each iteration on the CIFAR-10 dataset. Figure 4a shows the result of not using focal loss (i.e., γ = 0), whereas Figure 4b shows the result of using focal loss with γ = 1. It is evident that when not using focal loss, there is a significant quantitative imbalance in the pseudo labels of each class. The class imbalance problem can seriously affect the ability of the model to learn from classes with a small number of pseudo labels. In the early stages, the number of pseudo labels generated from a single class (i.e., class 2) is even higher than the aggregated number of pseudo labels generated from all other classes. This indicates that the unsupervised loss is dominated by a single class instead of all classes.

In contrast, the difference between the number of pseudo labels generated for each class is notably reduced when using focal loss. Therefore, the model can extract useful information from all classes. Focal loss does not present a stricter condition to reduce the number of pseudo labels. Instead, the total number of pseudo labels generated with focal loss is much higher than without focal loss (7 billion to 10 billion). Focal loss provides a smoother learning curve for the model to learn from all unlabeled data, which also shortens the iterations required to reach a stable phase of generating pseudo labels.

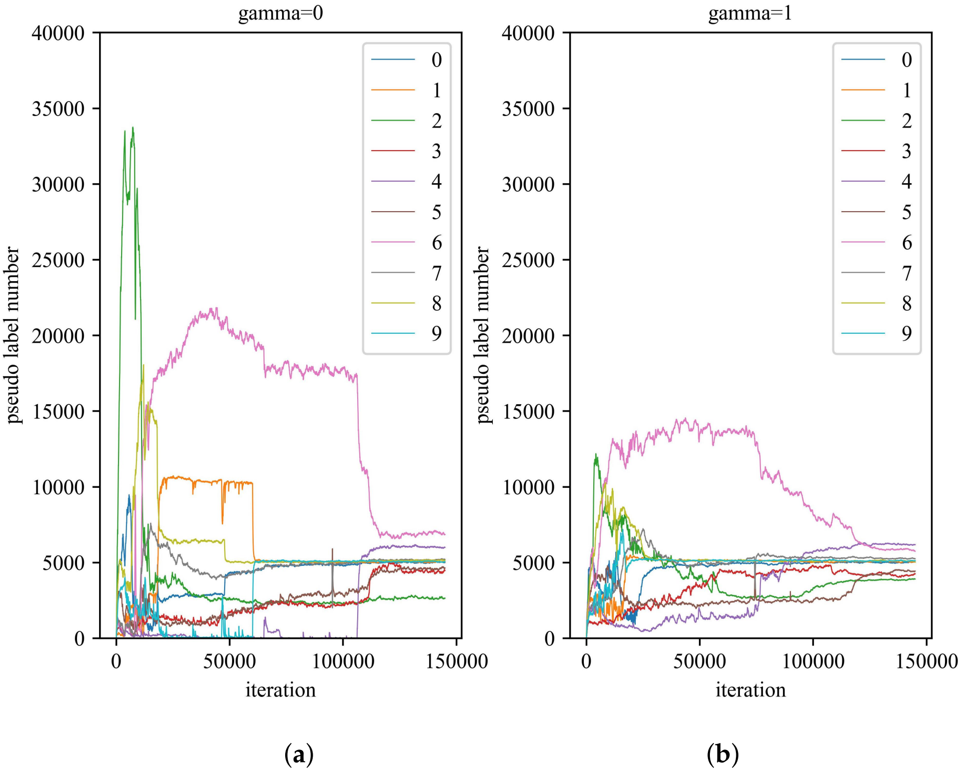


Figure 4. Ablation study on the use of focal loss. The number of pseudo labels generated of each class for the first 150,000 iterations (**a**) When gamma = 0 (i.e., not using focal loss). (**b**) When gamma = 1. The numbers in the figure legend represent the 10 classes in the CIFAR-10 dataset.

7. Conclusions and Future Work

This paper proposes FocalMatch, a new semi-supervised learning approach that combines FixMatch and focal loss. Instead of using the original cross-entropy loss for the unsupervised learning part, the focal loss is introduced in FocalMatch to effectively alleviate the problem of class imbalance that occurs on the generated pseudo labels during unsupervised learning. FocalMatch compels the model to focus more on the hard samples by adjusting the loss weights of different samples. Experiments show that FocalMatch dramatically reduces the variation in the number of pseudo labels generated for each class. In addition, FocalMatch outperforms all baseline methods and achieves state-of-the-art performance on several commonly used benchmarks, especially when the number of labeled data is extremely small. FocalMatch also provides a smoother learning curve and a higher convergence speed compared to FixMatch. The original focal loss contains an additional hyperparameter α that further adjusts loss contributions by the class frequency [17]. For semi-supervised learning methods that utilize pseudo-labeling, the number of pseudo labels generated for each class is usually unstable; therefore, it is hard to define the value of α beforehand. In future work, we plan to add α to FocalMatch and adjust the value of α and the modulating factor (i.e., γ) dynamically so that the model can converge more smoothly in different stages of training and is able to handle different tasks more efficiently.

Author Contributions: Conceptualization, N.Y., H.C. and Y.D.; methodology, Y.D.; software, C.Z.; formal analysis, Y.D.; investigation, Y.D and C.Z.; writing—original draft preparation, Y.D.; writing—review and editing, N.Y. and H.C.; supervision, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this study, there are three public datasets been used for performance evaluation, they are: CIFAR-10 [41], CIFAR-100 [41] and SVHN [42].

Acknowledgments: We thank Li Wang, Fan Huang and Yuning Zhang for their feedback on this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Experimental Details

Appendix A.1. Baseline Methods

In our experiments, we compare our method with Π model [38], Mean Teacher [39], MixMatch [40], ReMixMatch [24], UDA [25], FixMatch [15] since they are similar to our method. Among these models, ReMixMatch, UDA, and FixMatch all leverage data augmentations to ensure the consistency regularization [23] between different versions of the same image. Furthermore, pseudo-labeling [14] is also commonly used in these baseline methods. MixMatch, ReMixMatch, and UDA utilize sharpening on soft labels (i.e., the predicted class distribution of the model) to generate pseudo labels. FixMatch further applies arg max to soft labels to produce one-hot pseudo labels.

Appendix A.2. Hyperparameter Settings

We use similar hyperparameter settings as [15,44]: all baseline models and our model use Wide ResNet-28-2 [45] as the backbone network, batch size of 64 for labeled data, standard stochastic gradient descent with the momentum of 0.9 as the optimizer [46,47], initial learning rate of 0.03 with cosine learning rate decay [48], weight decay of 5×10^{-4}, EMA with the moment of 0.999. For method-dependent hyperparameters:

- μ: unlabeled data to labeled data ratio. **1** for Π model, Mean Teacher, MixMatch and ReMixMatch; **7** for UDA, FixMatch and FocalMatch.
- λ: weight of unsupervised loss. **10** for Π model, **50** for Mean Teacher, **100** for MixMatch, **1** for ReMixMatch, UDA, FixMatch and FocalMatch.
- **T**: temperature for sharpening soft labels. **0.5** for MixMatch and ReMixMatch, **0.4** for UDA.
- τ: threshold of generating pseudo label. **0.8** for UDA, **0.95** for FixMatch and FocalMatch.

As suggested by [44], all method-dependent hyperparameters follow the original papers. Table A1 provides a summary of method-dependent hyperparameters.

Table A1. Hyperparameters for Different Methods in Table 1 [1].

	Π Model	Mean Teacher	MixMatch	ReMixMatch	UDA	FixMatch	FocalMatch
μ	1	1	1	1	7	7	7
λ	10	50	100	1	1	1	1
T	-	-	0.5	0.5	0.4	-	-
τ	-	-	-	-	0.8	0.95	0.95

[1] These hyperparameter settings follow [44].

Appendix A.3. Training Details

The maximum training iteration of all experiments is set to 400,000. For the first 80 percent of the total iterations, we evaluate the overall accuracy of baseline models every 5000 iterations. For the rest 20 percent iterations, we increase the evaluation frequency to every 1000 iterations. In the meantime, we also record the accuracy of each class for FixMatch and our method on CIFAR-10 every 5000 iterations. All experiments are conducted based on codebase TorchSSL [44].

References

1. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NU, USA, 1997; Volume 1, p. 2.
2. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
3. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
4. Nadkarni, P.M.; Ohno-Machado, L.; Chapman, W.W. Natural language processing: An introduction. *J. Am. Med. Inf. Assoc.* **2011**, *18*, 544–551. [CrossRef] [PubMed]
5. Awoyemi, J.O.; Adetunmbi, A.O.; Oluwadare, S.A. Credit card fraud detection using machine learning techniques: A comparative analysis. In Proceedings of the 2017 international conference on computing networking and informatics (ICCNI), Ota, Nigeria, 29–31 October 2017; pp. 1–9.
6. Nageswaran, S.; Arunkumar, G.; Bisht, A.K.; Mewada, S.; Kumar, J.; Jawarneh, M.; Asenso, E. Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. *BioMed Res. Int.* **2022**, *2022*, 1755460. [CrossRef] [PubMed]
7. Sajja, G.S.; Mustafa, M.; Phasinam, K.; Kaliyaperumal, K.; Ventayen, R.J.M.; Kassanuk, T. Towards Application of Machine Learning in Classification and Prediction of Heart Disease. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1664–1669. [CrossRef]
8. Bhola, J.; Jeet, R.; Jawarneh, M.M.M.; Pattekari, S.A. Machine learning techniques for analysing and identifying autism spectrum disorder. In *Artificial Intelligence for Accurate Analysis and Detection of Autism Spectrum Disorder*; IGI Global: Hershey, PA, USA, 2021; pp. 69–81.
9. Pallathadka, H.; Jawarneh, M.; Sammy, F.; Garchar, V.; Sanchez, D.T.; Naved, M. A Review of Using Artificial Intelligence and Machine Learning in Food and Agriculture Industry. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April 2022; pp. 2215–2218. [CrossRef]
10. Arumugam, K.; Swathi, Y.; Sanchez, D.T.; Mustafa, M.; Phoemchalard, C.; Phasinam, K.; Okoronkwo, E. Towards applicability of machine learning techniques in agriculture and energy sector. *Mater. Today Proc.* **2022**, *51*, 2260–2263. [CrossRef]
11. Akhenia, P.; Bhavsar, K.; Panchal, J.; Vakharia, V. Fault severity classification of ball bearing using SinGAN and deep convolutional neural network. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2022**, *236*, 3864–3877. [CrossRef]
12. Sajja, G.S.; Mustafa, M.; Ponnusamy, R.; Abdufattokhov, S. Machine learning algorithms in intrusion detection and classification. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 12211–12219.
13. Arai, H.; Sakuma, J. Privacy preserving semi-supervised learning for labeled graphs. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 124–139.
14. Lee, D.H. *Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*; Workshop on challenges in representation learning, ICML. Citeseer; 2013 ; Volume 3, p. 896. Available online: https://scholar.google.com.au/scholar?q=The+Simple+and+Efficient+Semi-Supervised+Learning+Method+for+Deep+Neural+Networks&hl=en&as_sdt=0&as_vis=1&oi=scholart (accessed on 18 October 2022).
15. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
16. Grandvalet, Y.; Bengio, Y. Semi-supervised learning by entropy minimization. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 529–536.
17. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

18. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
19. Zhu, X.J. *Semi-Supervised Learning Literature Survey*; Technical Report 1530, Computer Sciences; University of Wisconsin-Madison: Madison, WI, USA, 2005.
20. Cheplygina, V.; de Bruijne, M.; Pluim, J.P. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **2019**, *54*, 280–296. [CrossRef]
21. Liang, P. Semi-Supervised Learning for Natural Language. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005.
22. Ouali, Y.; Hudelot, C.; Tami, M. An overview of deep semi-supervised learning. *arXiv* **2020**, arXiv:2006.05278.
23. Bachman, P.; Alsharif, O.; Precup, D. Learning with pseudo-ensembles. *arXiv* **2014**, arXiv:1412.4864. [CrossRef]
24. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv* **2019**, arXiv:1911.09785.
25. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6256–6268.
26. Japkowicz, N. The class imbalance problem: Significance and strategies. In Proceedings of the MICAI 2000: Advances in Artificial Intelligence: Mexican International Conference on Artificial Intelligence, Acapulco, Mexico, 11–14 April 2000; pp. 111–117.
27. Olszewski, D. A probabilistic approach to fraud detection in telecommunications. *Knowl.-Based Syst.* **2012**, *26*, 246–258. [CrossRef]
28. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [CrossRef]
29. Wang, S.; Yao, X. Using class imbalance learning for software defect prediction. *IEEE Trans. Reliab.* **2013**, *62*, 434–443. [CrossRef]
30. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef]
31. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 1–54. [CrossRef]
32. Yap, B.W.; Rani, K.A.; Rahman, H.A.A.; Fong, S.; Khairudin, Z.; Abdullah, N.N. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Kuala Lumpur, MA, USA, 16–18 December 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 13–22.
33. Ling, C.X.; Li, C. Data mining for direct marketing: Problems and solutions. *Kdd* **1998**, *98*, 73–79.
34. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
35. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
36. Collell, G.; Prelec, D.; Patil, K.R. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing* **2018**, *275*, 330–340. [CrossRef] [PubMed]
37. Zhu, B.; Pan, X.; vanden Broucke, S.; Xiao, J. A GAN-based hybrid sampling method for imbalanced customer classification. *Inf. Sci.* **2022**, *609*, 1397–1411. [CrossRef]
38. Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; Raiko, T. Semi-supervised learning with ladder networks. *arXiv* **2015**, arXiv:1507.02672. [CrossRef]
39. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv* **2017**, arXiv:1703.01780. [CrossRef]
40. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. *arXiv* **2019**, arXiv:1905.02249. [CrossRef]
41. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. 2009. Retrieved 17 August 2022. Available online: https://www.cs.toronto.edu/~kriz/cifar.html (accessed on 18 October 2022).
42. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning. In Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011. 2011. Available online: http://www.iapr-tc11.org/dataset/SVHN/nips2011_housenumbers.pdf (accessed on 18 October 2022).
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703. [CrossRef]
44. Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18408–18419.
45. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
46. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, PMLR, Atlanta, GA, USA, 17–19 June 2013; pp. 1139–1147.
47. Polyak, B.T. Some methods of speeding up the convergence of iteration methods. *Ussr Comput. Math. Math. Phys.* **1964**, *4*, 1–17. [CrossRef]
48. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
49. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.