applied sciences

MDPI

Article

Adaptive Age Estimation towards Imbalanced Datasets

Zhiang Dong [1] and Xiaoqiang Li [2,*]

[1] School of Software Technology, Zhejiang University, Ningbo 315048, China
[2] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
* Correspondence: xqli@shu.edu.cn

Abstract: Current age estimation datasets often have a skewed long-tail distribution with significant data imbalance, rather than an ideal uniform distribution for each category. The existing age estimation algorithms that rely on label distribution do not leverage data density information to address the issue of data imbalance. To solve the aforementioned problem, this paper proposes a novel method based on cost-sensitive learning, namely Data-Imbalance Adaptive Age Regression (DIAAR), for age estimation. DIAAR consists of two main modules: the adaptive soft label (ASL) module and the Data Density Smoothing (DDS) module. The ASL module embeds soft labels in the form of probability in the age regression. It assigns different degrees of soft labels adaptively to head and tail data based on their density, which helps balance the dataset. The DDS module further addresses data imbalance by revising data density through kernel smoothing and reweighting the loss function accordingly. Experiments on two benchmark datasets show that DIAAR can effectively deal with data imbalance and improve the accuracy of age estimation, achieving an average improvement of 8% over the baseline models. Moreover, this approach can be applied to various methods based on convolutional neural network models.

Keywords: age estimation; data imbalance; data density

check for updates

Citation: Dong, Z.; Li, X. Adaptive Age Estimation towards Imbalanced Datasets. *Appl. Sci.* **2023**, *13*, 10182. https://doi.org/10.3390/app131810182

Academic Editor: Chilukuri K. Mohan

Received: 31 July 2023
Revised: 6 September 2023
Accepted: 7 September 2023
Published: 11 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Age estimation is a task that involves predicting the actual age of an individual based on their facial image and has garnered significant interest due to its wide range of application scenarios, including visual surveillance [1], human–computer interactions [2], social media [3], and face retrieval [4]. Although, with the rise of big data, it is easier to collect a large scale of images for age estimation, datasets of age estimation often appear to have a skewed long-tail distribution with serious data imbalance.

A dataset is considered imbalanced when there are significant differences in the number of samples across different categories. The classes with a large number of samples are referred to as head-classes, while those with few samples are called tail-classes. Buda et al. [5] studied the impact of data imbalance on convolutional neural networks (CNNs) and concluded that it results in reduced performances of these networks. Models trained on imbalanced datasets often suffer from overfitting for head-classes and underfitting for tail-classes. Therefore, it is imperative to develop methods that can effectively combat data imbalance.

Numerous studies have been conducted on imbalanced datasets, which can be broadly categorized into two approaches: re-sampling techniques [5–8] and cost-sensitive learning methods [9–13]. Re-sampling is a commonly used technique at the data level to address the issue of data imbalance. It involves adjusting the training dataset in a manner that maintains a balanced sample quantity between the tail-classes and head-classes. Prior studies have investigated two main types of re-sampling: under-sampling, which involves reducing the number of samples in the head-classes, and over-sampling, which involves increasing the number of samples in the tail-classes.

Even though re-sampling can help balance the datasets, it is not a foolproof solution. On the one hand, over-sampling can result in the repeated learning of samples from the tail-classes, which can limit the diversity of the training dataset and potentially reduce the robustness of the resulting model. On the other hand, under-sampling can lead to a significant reduction in the number of samples in the head-classes, thereby limiting the model's ability to fully learn the variations within this category.

In addition to re-sampling, another solution proposed at the algorithmic level is cost-sensitive learning. In these methods, weights are assigned to samples to match a given data distribution [14]. Weighting by inverse class frequency [11,15] or a smoothed version of the inverse square root of class frequency [16,17] are often adopted. Although cost-sensitive learning is simple and effective, applying it to age regression brings new challenges distinct from its classification counterpart. Firstly, the distance between continuous age labels holds meaningful information. That is, the adjacent images in the label space will be interrelated and influence each other. For example, consider a scenario where two ages have the same limited number of samples, but one belongs to the head-classes and the other belongs to the tail-classes. This situation can result in varying degrees of data imbalance between the two age groups. Moreover, unlike classification, certain target values may have no data at all, which motivates the need for target extrapolation and interpolation [18].

This paper proposes a novel method based on cost-sensitive learning for robust estimation in the presence of data imbalance, called Data-Imbalance Adaptive Age Regression (DIAAR). The proposed method consists of two key operations: adaptive soft label (ASL) and Data Density Smoothing (DDS). ASL utilizes soft labels in the form of probability during regression, and adaptively assigns different degrees of soft labels to the head and tail data in the long-tail distribution according to the data density, applying cost-sensitive learning to age estimation. DDS modifies data density through kernel smoothing and further alleviates data imbalance by reweighting the loss function with the revised data density, which can prevent over-sampling and under-sampling.

The contributions of this work can be summarized as follows:

- The adaptive soft label (ASL) module makes use of soft labels in the age regression, reducing the imbalance of the data.
- The Data Density Smoothing (DDS) module smooths the distribution of data before reweighting the loss function.
- DIARR provides comparable or even better results than other leading methods on two benchmarks, IMDB-WIKI [19] and Morph II [20].

The rest of this paper is organized as follows. In Section 2, we summarize the recent development of age estimation and data imbalance. In Section 3, the proposed method is elaborated. Experiments are presented in Section 4. And in Section 5, we sum-up our work and put forward possible future work.

2. Related Work

The use of convolutional neural networks (CNNs) to extract age features from images has become increasingly popular, resulting in significant improvements in performance. For example, Levi developed a shallow CNN architecture [21] and Huo et al. [22] proposed a method using VGG-16. Recent works [23–26] all use CNN to extract deep features for age estimation tasks.

Previous papers [24,27] categorized existing age estimation algorithms into different groups. Here, a review is given about age estimation techniques on our own. We propose that the existing methods for age estimation can be broadly classified into five main categories: multi-class classification, regression, label distribution learning, ranking, and hybrid (i.e., combining two or three different types of algorithms simultaneously).

Classification-based methods consider ages as independent categories. DEX [19] is a typical classification method. It treats age estimation as a 101-class classification problem. Therefore, the penalty imposed on different error degrees is the same during the optimization process, which is unreasonable. For instance, the penalty imposed on wrongly

predicting the age of 10 to 12 or to 20 should not be equivalent. Therefore, the relation between age labels is not exploited in multi-class classification works. Another restriction of multi-classification of age estimation is that the class imbalance in age facial datasets can result in an overfitting problem [23].

Utilizing a regressor to predict age is a common approach and can lead to improved accuracy. However, the aging process is inherently random and it may not be appropriate to apply a single pattern to fit real aging modes. Therefore, some studies have proposed the use of multiple local regressors, using a divide-and-conquer strategy to handle different age groups. However, this approach may ignore the continuous relationship between subsets and result in suboptimal performance. To address this issue, Li et al. proposed Bridgenet [24], which uses gating networks to enforce similarity between neighboring subset nodes and improve the overall performance of the age estimation model.

Label-distribution learning approaches to age estimation [22,28] were proposed to alleviate the disadvantages brought by insufficient training data with exact ages and group ambiguity. They represented each age as a distribution and KL divergence was applied to measure the similarity between a predicted distribution and ground-truth distribution. The training instances related to each age will be increased without an increase in the number of training samples. For example, Pan et al. put forward mean-variance loss [29] for robust age estimation via distribution learning.

To exploit the relative ordering of ages, ranking-based methods utilize several simple binary classifiers to judge the rank of a given facial image's age [25,30,31]. The final prediction is the combination of the results of sub-binary classifiers. For example, Ranking-CNN pre-trained several basic CNNs on a large dataset and fine-tuned them with ordinal age labels [32]. However, these methods ignore the relation between binary sub-problems.

Hybrid methods are the combination of two or more algorithms. They have become popular in age estimation recently [33–35], because fusing two or three methods can bring complementary advantages.

The research on imbalanced datasets can be roughly divided into the data-level and algorithm-level. Re-sampling is considered one of the primary methods for alleviating data imbalance in imbalanced datasets, addressing the issue from a data-centric perspective. Re-sampling is to adjust the training dataset in a proper way to make sure the sample quantity of the minority class and majority class in the training dataset is basically consistent. It can be roughly divided into over-sampling and under-sampling.

Over-sampling is to increase the sample of the minority class dataset in different ways so that the sample number of the minority class and the sample number of the majority class are basically the same. Random over-sampling is a common method, which randomly copies a few samples to achieve data balance. This processing method has great randomness, and it is easy to lead to sample repetition and overfitting situations. Chawle and his team [36] proposed an over-sampling technique called SMOTE, in which a new sample is randomly generated between the two selected and cycled over and over again to increase the sample number of the minority class. In addition, Hans and his researchers [37] improved the original SMOTE over-sampling technique and proposed the borderline-SMOTE method. This method generates new synthetic examples along the line between the minority example and its selected nearest neighbors. In contrast to over-sampling, under-sampling is to reject some samples in the majority of samples through appropriate methods, so as to balance the imbalanced dataset [38]. This approach works better when dealing with a large sample set that contains very few minority classes.

At the algorithm level, the learning classifier is improved to make the model have better classification and prediction ability on imbalanced data. It can be divided into integrated learning, cost-sensitive learning, and some other methods. Cost-sensitive learning is mainly introduced here. Long before deep learning, a lot of works have been conducted to reweight the loss according to the data distribution. In most studies, the focus is often on information from minority classes, and the mis-classification cost of majority classes is much lower than that of minority classes. Adacost [39] is a typical cost-sensitive method

based on such ideas. Its idea is to assign different mis-classification costs to each category. Specifically, more weight will be given to mis-classification-positive categories, so as to improve the recognition rate of positive category samples.

3. Method

In this section, the proposed method to alleviate the data imbalance in age estimation is presented. Figure 1 shows the overall framework of DIAAR. DIAAR contains two key modules, the adaptive soft label (ASL) module and the Data Density Smoothing (DDS) module, which reinforce the model result and alleviate the data imbalance. The ASL and DDS modules are elaborated in Sections 3.1 and 3.2, respectively, and the training objective is illustrated in Section 3.3.

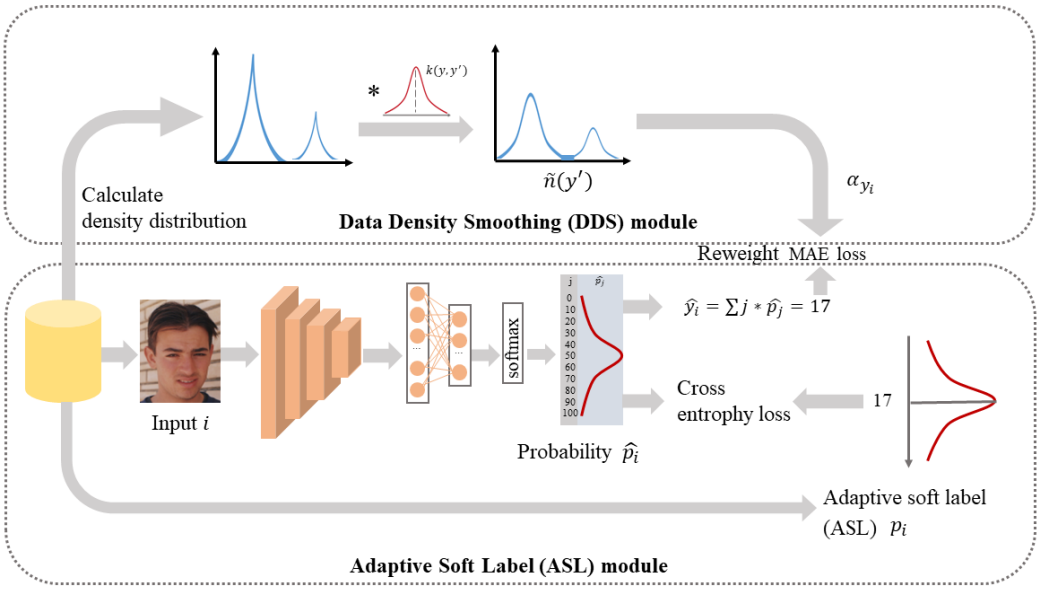


Figure 1. The overall framework of the proposed method, DIAAR.

3.1. Adaptive Soft Label Module

The illustration of the ASL module is shown in Figure 2. Specifically, we assign different discrete label distributions to sample X of different ages according to the sample distribution of training data. Here, $X \sim N(\mu, \sigma^2)$ and σ is changed dynamically. The samples in head-classes correspond to the sharper discrete Gaussian distribution (smaller σ); the samples with smaller sample sizes correspond to the flatter discrete Gaussian distribution (larger σ).

We hereby designed an age prediction model with adaptive data imbalance. Different from the previous work [28,40], we used the data-imbalance adaptive soft label. Inspired by the idea of cascade in C3AE [35], our model estimates the age based on the face images by combining regression and distribution. Through incorporating both regression and distribution components, ASL module aims to capture the underlying patterns and variations in age estimation. By minimizing loss and updating network parameters, this module makes distribution prediction and regression predictions close to the ground-truth, and finally reduces prediction error. In our method, we define two losses for age estimation. The first loss measures discrepancy between data-imbalance adaptive soft label and predicted age distribution. We adopt cross entropy loss as the measurement,

$$\mathcal{L}_{cross}(p_i, \hat{p}_i) = -\frac{1}{N}\sum_{i=1}^{N} p_i \ln \hat{p}_i \tag{1}$$

where p_i is the proposed data-imbalance adaptive soft label, and $\hat{p}_i$ is the predicted age distribution. N represents the batch size.

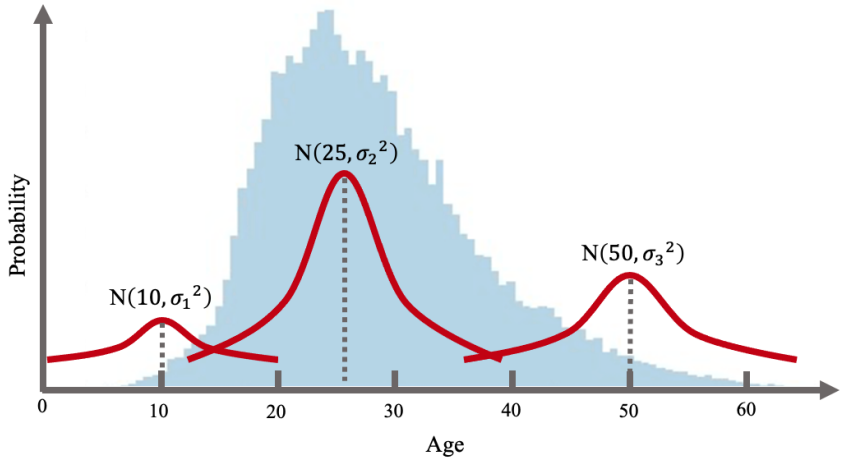


Figure 2. The data-imbalance adaptive soft label. Take IMDB-WIKI-DIR dataset as an example. The data density distribution is represented by a shaded image, and the generated soft labels for different ages are represented by red Gaussian curves.

3.2. Data Density Smoothing Module

The DDS module is shown in Figure 1. Before reweighting the loss function in this module, we applied a kernel function to smooth the data density.

The steps of DDS are shown in Figure 3. DDS convolves the data density distribution with the Gaussian kernel $k(y, y')$ to extract the kernel smooth version of the density distribution. The smoothed density distribution makes reasonable use of the correlation between the data samples of adjacent labels, and overlaps the training information corresponding to adjacent labels. Gaussian kernel $k(y, y')$ describes the degree of similarity between the target y' and any label y. A symmetric kernel function is any one that satisfies $k(y, y') = k(y', y)$ and $\nabla_y k(y, y') + \nabla_{y'} k(y', y) = 0, \forall y, y' \in \mathcal{Y}$. Here, the $\mathcal{Y}$ is on behalf of the label space. For example, Gaussian kernel and Laplacian kernel are symmetric kernel functions, and Gaussian kernel is used in all experiments. The effective label distribution formula calculated by DDS is:

$$\tilde{n}(y') = \int_y k(y, y') n(y) dy \tag{2}$$

where $n(y)$ represents the number of occurrences of label y within the training data. $\tilde{n}(y')$ is the effective density of label y'.

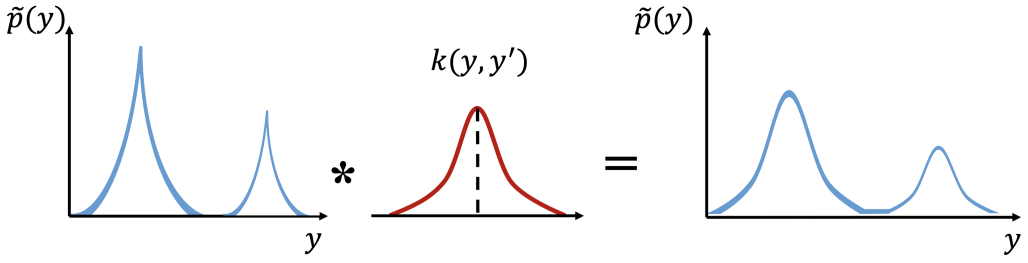


Figure 3. Data density smoothing strategy (DDS).

Modified data density through DDS is used for subsequent loss reweighting to form the loss reweighting module. This module is inserted into the adaptive soft label module in Section 3.1 to form the DIAAR, so as to deal with data imbalance more effectively.

Unlike labels in classification tasks, which are discrete, the age labels in age regression tasks are continuous values. DIAAR divides the label space $\mathcal{Y}$ into B groups in order to reasonably apply the loss function to the age regression task. It assumes that the training

dataset is $\{(x_i, y_i)\}_{i=1}^{N}$, x_i represents input image, and y_i represents age labels. The target space is divided into $[y_0, y_1)$, $[y_1, y_2)$, ... ,$[y_{B-1}, y_B)$. In this paper, we set $\delta_y \triangleq y_{b+1} - y_b = 1$; namely, the resolution of the age group division is 1. After dividing the target space into age groups, DIAAR can simply insert some classic reweighting methods. This paper adopts square-root inverse variable weighting (SQINV) [16,17] to reweight loss function, which is a simple and effective method to solve imbalanced classification problems. Therefore, the weights α_{y_i} of age y_i are calculated as Equation (3):

$$\alpha_{y_i} = \frac{1}{\sqrt{\tilde{n}(y_i)}} \tag{3}$$

where $\tilde{n}(y_i)$ represents the effective data density of label y_i.

Because of the missing of data, we need to interpolate and extrapolate age predictions. To achieve this, as shown in Figure 1, a $softmax$ expected value E is calculated as the final age regression prediction value $\hat{y}_i$. The computational process of the Distribution Regression (DR) operation is given as:

$$\hat{y}_i = E(O) = \sum_{j=0}^{100} j * \hat{p}_j, \tag{4}$$

where $O = \{0, 1, \ldots, 100\}$ and O is the output layer with 101 dimensions. $\hat{p}_j$ is the prediction distribution through the softmax layer. y_i is the discrete age corresponding to i class and symbol $*$ means multiplication.

The DR operation can not only improve the prediction accuracy of the data-imbalance adaptive age regression module, but also enable DDS to be combined with the loss reweighting to deal with classification imbalance problems.The average absolute error loss $\mathcal{L}_{MAE}$ is calculated based on the regression value $\hat{y}_i$ and age discrete label y_i. The training goal of this step is to minimize MAE loss:

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{y}_i - y_i\| \tag{5}$$

Specifically, the MAE loss of sample X_i is multiplied by the inverse of the effective label density obtained with the DDS strategy, and the process is calculated as Equation (6):

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_{i=1}^{N} \alpha_{y_i} \|\hat{y}_i - y_i\| = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sqrt{\tilde{n}(y_i)}} \|\hat{y}_i - y_i\| \tag{6}$$

Here, y_i represents the age label of sample x_i, $\hat{y}_i$ represents the regression prediction value of sample x_i, $\tilde{n}(y_i)$ represents the effective data density of label y_i, and α_{y_i} represents the weight of age label y_i.

3.3. Training Objective

The train objective of DIAAR is to minize the loss functions of ASL module and DDS module. Finally, DIAAR combines the data-imbalance adaptive soft label module with the Data Density Smoothing module, and the total loss function is shown as Equation (7):

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{cross} + \mathcal{L}_{MAE} \tag{7}$$

Here, λ is a hyper-parameter used to adjust the contribution degree of two loss terms to the overall loss function. We set $\lambda = 0.01$ according to the experiments in Section 4.5.

4. Experiments

To validate the effectiveness of our method, we report the main results in this section on datasets of age estimation.

4.1. Datasets and Setting

This section introduces the dataset, the details of evaluation in the experiments, how to implement the proposed method, and the settings to train the model.

4.1.1. Datasets

In order to evaluate the performance of the proposed method on imbalanced datasets, experiments were conducted on two benchmark datasets for age estimation: IMDB-WIKI-DIR and Morph II-DIR. The data density distribution of the IMDB-WIKI-DIR and Morph II-DIR datasets are shown in Figure 4.

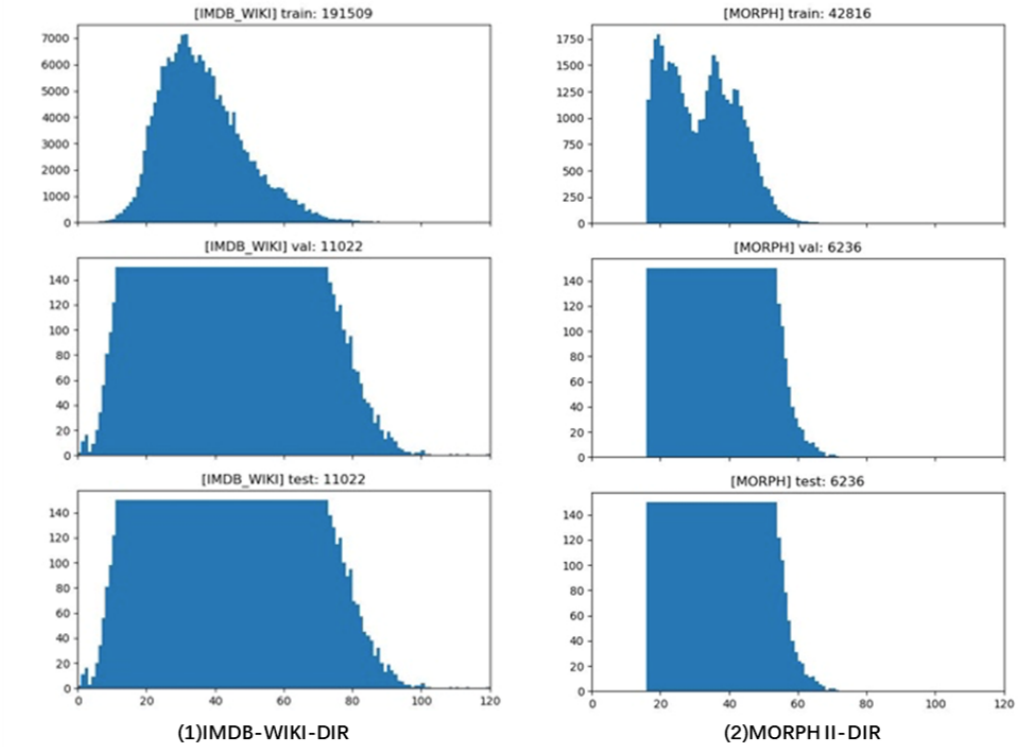


Figure 4. The data density distribution.

IMDB-WIKI-DIR: It is an imbalanced dataset which was constructed based on the IMDB-WIKI [19] dataset. We filtered low-quality images, manually divided the balanced validation set and test set, and used the remaining samples as the training set. We ensured that the number of samples corresponding to each age label of validation set and test set did not exceed 150, and the remaining samples were used as the training set. Ultimately, the IMDB-WIKI-DIR dataset included 191,509 images for training and 11,022 images for verification and testing. The minimum age was 0 and the maximum age was 100, and there was a great data imbalance in this dataset. When using the IMDB-WIKI-DIR dataset as a pre-training dataset, the target space of the dataset was limited to 0∼100, so the age estimator could be fine-tuned on the Morph II-DIR dataset.

Morph II-DIR: It is an imbalanced dataset which was constructed based on the Morph II [20] dataset. We used manual partition to ensure that the number of samples corresponding to each age label of the validation set and test set did not exceed 150, and the remaining samples were used as the training set. Ultimately, the Morph II-DIR dataset included 42,816 images for training and 6236 images for verification and testing. The minimum age range was 16 and the maximum was 77.

We carefully selected the IMDB-WIKI-DIR and Morph II-DIR datasets as our primary datasets for this study. These datasets were chosen due to their ability to effectively represent and reflect the data imbalance commonly encountered in age estimation tasks.

4.1.2. Evaluation Process and Evaluation Metrics

The experiments continued to use the evaluation process in previous works [41–43]. The model was trained on unbalanced datasets and was evaluated on the corresponding balanced test sets. To evaluate the performance of the proposed method in regions with different sample sizes, this paper divides the target space into three disjointed subsets following the evaluation settings in work [43]: Many shot (the sample size >100), Medium shot (the sample size 20∼100), and Few shot (the sample size <20). The evaluation metrics were calculated on the complete test set and the above three subsets, which were denoted as All, Many, Med., and Few. Mean Absolute Error (MAE) and Error Geometric Mean (GM) were used as evaluation metrics in the experiments.

MAE: MAE is a common evaluation metric for regression. MAE measures the error between the predicted age ($\hat{y}_i$) and ground-truth (y_i), which is computed as: $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$, where N is the number of testing images.

GM: To ensure fairness, GM was also adopted as an evaluation metric in experiments, and was defined as $GM = (\prod_{i=1}^{N} e_i)^{\frac{1}{N}} = (\prod_{i=1}^{N}|y_i - \hat{y}_i|)^{\frac{1}{N}}$, where y_i is the ground-truth, $\hat{y}_i$ is the predicted age, and N is the number of testing images.

4.1.3. Implementation Details

We modified ResNet50 as the network of DIAAR. The output of the fourth stage in the original ResNet50 goes through the global average pooling layer, finally getting the 101-dimension output. The prediction distribution $\hat{p}_j$ is the output of the softmax layer. The prediction distribution $\hat{p}_j$ and adaptive soft label participate in the similarity calculation of cross-entropy loss. $Softmax$ expectation [19] is used as the predicted age value $\hat{y}_i$ of the regression. Inputs will be resized to the size of 224 × 224. We used zero-padding to expand 20% on each side and then random crop back to the original image size. The baseline model (baseline) of DIAAR refers to the modified ResNet50.

4.1.4. Training Details

A single GPU (GeForce GTX 1080 Ti) was used in experiments. Following the settings and parameters in previous works [18,44], the batchsize was 64 and the models were trained for 150 epoches. The Adam optimizer and dynamic learning rate were applied in all experiments. A dynamic learning rate makes the model more stable in the later training period. During the training process on the IMDB-WIKI-DIR dataset, the initial learning rate was 0.001, and learning rate was decreased to 0.0001 and 0.00001 after 180 K iterations and 240 K iterations, respectively. Similarly, during the training process on the Morph II-DIR dataset, the initial learning rate was 0.001 and learning rate was decreased to 0.0001 and 0.00001 after 40 K iterations and 53 K iterations, respectively. The hyper-parameter λ in Equation (7) was set as 0.01 to balance the two loss items according to the experiments in hyper-parameter setting experiments in Section 4.5.

4.2. Comparison Experiments

This section introduces the results of experiments on two datasets, IMDB-WIKI-DIR and Morph II-DIR, and compares the results with other baseline models on these datasets.

4.2.1. Comparison on IMDB-WIKI-DIR

In order to verify the effectiveness of our method, which consists of baseline (modified ResNet50), ASL, and DDS, we compare DIAAR with the existing methods dealing with data imbalance problems such as SMOTER [45], RRT [18], and GAI [44] on the IMDB-WIKI-DIR dataset. The results are shown in Table 1. As we can see, the MAE and GM on the three disjointed subsets of DIAAR are both reduced to a certain extent compared with the baseline, especially in the Medium shot and Few shot regions. This phenomenon shows that the DIAAR method is very effective for handling tail samples. Comparing the existing methods for dealing with the imbalance of age estimates, DIAAR comes in first

place in both MAE and GM evaluation indicators. Therefore, DIAAR has great advantages in dealing with unbalanced datasets.

Table 1. Comparison experiment on IMDB-WIKI-DIR.

Metrics	MAE ↓				GM ↓			
Shot	**All**	**Many**	**Med.**	**Few**	**All**	**Many**	**Med.**	**Few**
Baseline	8.08	7.24	15.49	24.96	4.56	4.14	11.21	20.60
SMOTER [45]	8.14	7.42	14.15	25.28	4.64	4.30	9.05	19.46
SMOGN [46]	8.03	7.30	14.02	25.93	4.63	4.30	8.74	20.12
RRT [18]	7.81	7.07	14.06	25.13	4.35	4.03	8.91	16.96
BMC [44]	8.08	7.52	12.47	23.29	-	-	-	-
GAI [44]	8.12	7.58	12.27	23.05	-	-	-	-
DIAAR	7.79	7.20	12.78	23.38	4.30	4.14	7.35	13.35

4.2.2. Comparision on Morph II-DIR

To demonstrate the effectiveness and versatility of our method, we apply the method to three convolutional neural networks, including (1) ResNet50 [47], the denoted baseline model whose backbone network was modified according to ResNet50; (2) C3AE [35], an extremely compact yet efficient cascade context-based age estimation model for age estimation; and (3) SqueezeNet [48], a compact convolutional neural network including several popular modules, convolutional layers, down-sampling layers, and fully connected layers. The results are shown in Table 2.

We can see from the table that the MAE decreases from 2.88 to 2.66 on the complete test set by applying our method to ResNet50. That is to say, our method obtained an improvement of 8% over the baseline model (ResNet50). In addition, the prediction accuracy of the proposed method on the three disjointed subsets is improved to some extent. This shows that our method is very effective for processing tail classes. The evaluation indexes of three different CNNs combined with our method are also listed in the table. Experimental results show that our method combined with existing CNNs have a consistent performance improvement in the age estimation task.

Table 2. Comparison experiment on Morph II-DIR.

Metrics	MAE ↓				GM ↓			
Shot	**All**	**Many**	**Med.**	**Few**	**All**	**Many**	**Med.**	**Few**
ResNet50 [47]	2.88	2.76	4.38	8.43	1.78	1.72	2.96	5.69
ResNet50 + OURS	2.66	2.51	4.34	7.40	1.62	1.54	3.04	5.12
C3AE [35]	3.36	3.21	5.25	9.76	2.16	2.09	3.34	7.71
C3AE + OURS	3.10	2.86	6.24	11.18	1.92	1.80	5.15	10.26
SqueezeNet [48]	3.32	3.14	5.87	10.36	2.08	1.99	4.43	9.28
SqueezeNet + OURS	3.09	2.86	5.90	11.06	1.92	1.82	4.39	10.07

4.3. Ablation Experiments

To validate the effectiveness of three parts of the proposed method, ASL (adaptive soft label) and DDS (Data Density Smoothing), we compare different combinations of two parts with the baseline. We can see from Table 3 that SQINV, a loss reweighting method of classification imbalance, is ineffective and may even drag down model age-estimation accuracy. The results show that using DDS before the reweighting loss fuction (SQINV + DDS) is better than only using SQINV, which reduces the error by 24% for the Few shot region and continuously improve the performance of all regions. We can conclude that DDS provides comprehensive and fair treatment for all target values, and the imbalanced age estimation problem is significantly improved.

It can also be observed from the experimental results that adding ASL to the baseline can reduce MAE from 2.88 to 2.75, which indicates that the adaptive soft label (ASL) is very

effective. Even if used alone, ASL is competitive among many methods of arrangement and combination, as shown in Table 3.

Finally, it is shown in Table 3 that the combination of the two key operations proposed in this paper (SQINV + DDS + ASL) can achieve the best test error of 2.72 MAE. Overall, DIAAR has advantages in imbalanced age estimation.

Table 3. Ablation experiments on Morph II-DIR.

Metrics	**MAE ↓**				**GM ↓**			
Shots	**All**	**Many**	**Med.**	**Few**	**All**	**Many**	**Med.**	**Few**
Baseline	2.88	2.76	4.38	8.43	1.78	1.72	2.96	5.69
SQINV	2.92	2.87	3.49	5.95	1.85	1.82	2.27	4.12
SQINV + DDS	2.73	2.62	3.90	6.42	1.71	1.64	2.62	4.21
ASL	2.75	2.61	4.26	7.32	1.67	1.61	2.96	5.45
SQINV + DDS + ASL	2.72	2.57	4.44	7.56	1.65	1.57	2.83	5.88

4.4. Loss Function Reweighting Strategies

The loss function reweighting in our method can be devised as two strategies: (1) $\mathcal{L}_{cross}(re)$, where the loss reweighting is applied to cross-entropy loss directly; (2) $\mathcal{L}_{cross} + \mathcal{L}_{MAE}(re)$, where the regression age used to calculate the MAE loss is obtained by applying the DR operation and then the loss reweighting is applied to MAE loss. To explore which kind of strategy is more effective, comparison experiments of the two strategies are designed. Morph II-DIR is used as the unbalanced dataset. The MAE and GM metrics of the two strategies are shown in Table 4. We can see that joint using cross-entropy loss and mae loss (reweighted) together outperform using cross-entropy loss (reweighted). This is reasonable because $\mathcal{L}_{cross} + \mathcal{L}_{MAE}(re)$ strengthens the relationship between age labels and characteristics by adding the mean absolute error loss.

Table 4. Loss function reweighting strategies on Morph II-DIR.

Metrics	**MAE ↓**				**GM ↓**			
Shots	**All**	**Many**	**Med.**	**Few**	**All**	**Many**	**Med.**	**Few**
$\mathcal{L}_{cross}$	2.82	2.65	4.87	8.45	1.71	1.63	3.41	6.36
$\mathcal{L}_{cross}(re)$	2.85	2.72	4.13	7.39	1.76	1.69	2.72	5.12
$\mathcal{L}_{cross} + \mathcal{L}_{MAE}(re)$	2.77	2.65	3.98	6.54	1.71	1.62	2.79	4.56

4.5. Hyper-Parameter Setting

The hyper-parameter λ in Equation (7) is used to balance the influence of $\mathcal{L}_{cross}$ and $\mathcal{L}_{MAE}$. Therefore, we conducted hyper-parameter experiments on Morph II to study the influence of λ. Experimental results on the whole test set are shown in Table 5. By setting the value range from 0.001 to 1, we found that $\lambda = 0.01$ is the best hyper-parameter which obtains the lowest MAE. Therefore, we take the same setting ($\lambda = 0.01$) in other experiments.

Table 5. Hyper-parameter setting experiments on Morph II-DIR.

Metrics	**MAE ↓**				**GM ↓**			
λ	0.001	0.01	0.1	1	0.001	0.01	0.1	1
SQINV + DDS + ASL	2.79	2.72	2.79	3.00	1.73	1.65	1.74	1.86

5. Conclusions

In this paper, we propose Data-Imbalance Adaptive Age Regression (DIAAR), a novel approach for robust age estimation, which utilizes data density information to address the problem of imbalanced age estimation. Our method employs two key techniques, adaptive

soft label (ASL) and Data Density Smoothing (DDS), which were shown to be effective in reducing the negative impact of data imbalance on age estimation, as demonstrated by our experiments on the IMDB-WIKI-DIR and Morph II-DIR datasets. The effectiveness of adaptive soft label (ASL) and Data Density Smoothing (DDS) was verified with related ablation experiments. Our proposed approach yielded a significant average improvement of 8% over the baseline models. Additionally, we have investigated the impact of pre-training settings under the background of unbalanced data, and the results indicate that the effectiveness of pre-training is affected by the distribution of pre-trained datasets and finetuned datasets. Our study has focused on proposing a data-imbalanced adaptive model to eliminate the interference of different data distribution or pretraining strategies and exploring experimental settings with optimized parameters for forthcoming research in the future.

Author Contributions: Methodology, Z.D.; Software, Z.D.; Validation, Z.D.; Resources, X.L.; Writing—original draft, Z.D.; Writing—review & editing, X.L.; Supervision, X.L. 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: Not applicable.

Acknowledgments: We appreciate the High Performance Computing Center of Shanghai University and Shanghai Engineering Research Center of Intelligent Computing Systems for providing the computing resources and technical support.

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

References

1. Song, Z.; Ni, B.; Guo, D.; Sim, T.; Yan, S. Learning universal multi-view age estimator using video context. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 241–248. [CrossRef]
2. Geng, X.; Zhou, Z.-H.; Zhang, Y.; Li, G.; Dai, H. Learning from facial aging patterns for automatic age estimation. In Proceedings of the 14th ACM international Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 307–316. [CrossRef]
3. Rothe, R.; Timofte, R.; Van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.* **2018**, *126*, 144–157. [CrossRef]
4. Lanitis, A.; Draganova, C.; Christodoulou, C. Comparing different classifiers for automatic age estimation. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2004**, *34*, 621–628. [CrossRef] [PubMed]
5. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2017**, *106*, 249–259. [CrossRef] [PubMed]
6. Ren, M.; Zeng, W.; Yang, B.; Urtasun, R. Learning to reweight examples for robust deep learning. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 4334–4343.
7. Geifman, Y.; El-Yaniv, R. Deep active learning over the long tail. *arXiv* **2017**, arXiv:1711.00941.
8. Zou, Y.; Yu, Z.; Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
9. Ting, K.M. A comparative study of cost-sensitive boosting algorithms. In Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000.
10. Zhou, Z.H.; Liu, X.Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Knowl. Data Eng.* **2005**, *18*, 63–77. [CrossRef]
11. Huang, C.; Li, Y.; Loy, C.C.; Tang, X. Learning deep representation for imbalanced classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5375–5384.
12. Khan, S.H.; Hayat, M.; Bennamoun, M.; Sohel, F.A.; Togneri, R. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 3573–3587. [PubMed]
13. Sarafianos, N.; Xu, X.; Kakadiaris, I.A. Deep imbalanced attribute classification using visual attention aggregation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 680–697.
14. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277.
15. Wang, Y.X.; Ramanan, D.; Hebert, M. Learning to model the tail. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.

16. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2013**, *26*, 1–9.
17. Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; Van Der Maaten, L. Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 181–196.
18. Yang, Y.; Zha, K.; Chen, Y.; Wang, H.; Katabi, D. Delving into deep imbalanced regression. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 11842–11851.
19. Rothe, R.; Timofte, R.; Van Gool, L. Dex: Deep expectation of apparent age from a single image. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 10–15. . [CrossRef]
20. Ricanek, K.; Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 341–345. [CrossRef]
21. Levi, G.; Hassner, T. Age and gender classification using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 34–42. [CrossRef]
22. Huo, Z.; Yang, X.; Xing, C.; Zhou, Y.; Hou, P.; Lv, J.; Geng, X. Deep age distribution learning for apparent age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 17–24. [CrossRef]
23. Gao, B.-B.; Zhou, H.-Y.; Wu, J.; Geng, X. Age Estimation Using Expectation of Label Distribution Learning. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 712–718. [CrossRef]
24. Li, W.; Lu, J.; Feng, J.; Xu, C.; Zhou, J.; Tian, Q. Bridgenet: A continuity-aware probabilistic network for age estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1145–1154. [CrossRef]
25. Liu, H.; Lu, J.; Feng, J.; Zhou, J. Ordinal deep learning for facial age estimation. *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)* **2017**, *29*, 486–501. [CrossRef]
26. Liu, X.; Zou, Y.; Kuang, H.; Ma, X. Face image age estimation based on data augmentation and lightweight convolutional neural network. *Symmetry* **2020**, *12*, 146. [CrossRef]
27. Yang, T.-S.; Huang, Y.-I.; Lin, Y.-E.; Hsiu, P.-I.; Chuang, Y.-U. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; Volume 5, pp. 1078–1084. [CrossRef]
28. Yang, X.; Gao, B.-B.; Xing, C.; Huo, Z.-W.; Wei, X.-S.; Zhou, Y.; Wu, J.; Geng, X. Deep label distribution learning for apparent age estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 102–108. [CrossRef]
29. Pan, H.; Han, H.; Shan, S.; Chen, X. Mean-variance loss for deep age estimation from a face. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5285–5294. [CrossRef]
30. Chang, K.Y.; Chen, C.S.; Hung, Y.P. A ranking approach for human ages estimation based on face images. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3396–3399. [CrossRef]
31. Li, C.; Liu, Q.; Liu, J.; Lu, H. Learning ordinal discriminative features for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2570–2577. [CrossRef]
32. Chen, S.; Zhang, C.; Dong, M.; Le, J.; Rao, M. Using ranking-cnn for age estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5183–5192. [CrossRef]
33. Choi, S.E.; Lee, Y.J.; Lee, S.J.; Park, K.R.; Kim, J. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognit. (PR)* **2011**, *44*, 1262–1281. [CrossRef]
34. El Dib, M.Y.; El-Saban, M. Human age estimation using enhanced bio-inspired features (EBIF). In Proceedings of the IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 1589–1592. [CrossRef]
35. Zhang, C.; Liu, S.; Xu, X.; Zhu, C. C3AE: Exploring the limits of compact model for age estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12587–12596. [CrossRef]
36. 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]
37. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 878–887.
38. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* **2020**, arXiv:2010.16061.
39. Fan, W.; Stolfo, S.J.; Zhang, J.; Chan, P.K. AdaCost: Misclassification cost-sensitive boosting. In Proceedings of the Sixteenth International Conference on Machine Learning, San Francisco, CA, USA, 27–30 June 1999; Volume 99, pp. 97–105.
40. Yang, X.; Geng, X.; Zhou, D. Sparsity Conditional Energy Label Distribution Learning for Age Estimation. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 9–15 July 2016; pp. 2259–2265.

41. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
42. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
43. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-scale long-tailed recognition in an open world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2537–2546.
44. Ren, J.; Zhang, M.; Yu, C.; Liu, Z. Balanced MSE for Imbalanced Visual Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 7926–7935.
45. Torgo, L.; Ribeiro, R.P.; Pfahringer, B.; Branco, P. Smote for regression. In Proceedings of the Portuguese Conference on Artificial Intelligence, Coimbra, Portugal, 8–11 September 2013; pp. 378–389.
46. Branco, P.; Torgo, L.; Ribeiro, R.P. SMOGN: A pre-processing approach for imbalanced regression. In Proceedings of the 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, Skopje, Macedonia, 22 September 2017; pp. 36–50.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
48. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.