Predicting Potential Propensity of Adolescents to Drugs via New Semi-Supervised Deep Ordinal Regression Model

Alireza Ganjdanesh[1], Kamran Ghasedi[1], Liang Zhan[1], Weidong Cai[2], Heng Huang[1,3]⋆

[1] Electrical and Computer Engineering, University of Pittsburgh, PA 15261, USA
[2] School of Computer Science, University of Sydney, Sydney NSW 2006, Australia
[3] JD Finance America Corporation
alireza.ganjdanesh@pitt.edu, kamran.ghasedi@gmail.com, liang.zhan@pitt.edu, tom.cai@sydney.edu.au, heng.huang@pitt.edu

Abstract. Addiction to drugs between young people is one of the most severe problems in the real world, and it imposes a huge financial and emotional burden on their families and societies. Therefore, predicting potential inclination to drugs at earlier ages can prevent lots of detriments. In this paper, we propose a new semi-supervised deep ordinal regression model to predict the possible propensity of adolescents to marijuana using the diffusion MRI-derived mean diffusivity (MD) from 148 Regions of Interest (ROIs). The traditional deep ordinal regression models cannot be directly applied to our biomedical problem which only has a small number of labeled data, not enough to train the deep learning models. Thus, we design a semi-supervised learning mechanism for deep ordinal regression, such that both labeled and unlabeled data can be used to enhance the model training. In our experiments, we use the ABCD dataset, which contains MRI images of the adolescents under study and their answers in the Likert scale to a questionnaire containing questions about Marijuana. Experimental results on the ABCD dataset validate the superior performance of our new method. Our study provides an inexpensive way to predict the drug tendency using brain MRI data.

Keywords: Adolescent · Marijuana · Deep Learning · Ordinal Regression · Semi-Supervised Learning · diffusion MRI · mean diffusivity.

1 Introduction

Predicting a potential tendency of adolescents to drugs in the future enables us to take effective preventive actions against the risk of their addiction to drugs. One of the approaches to do so is to study brain condition and its possible correlation with different behavioral patterns. In this regard, a study called Adolescents Brain Cognitive Development (ABCD)[1] is in progress, which is the largest

⋆ Corresponding Author. This work was partially supported by NSF IIS 1836945, IIS 1836938, IIS 1845666, IIS 1852606, IIS 1838627, IIS 1837956, and NIH AG049371.

[1] https://www.addictionresearch.nih.gov/abcd-study

long-term study of brain development in the United States. This study aims to monitor the brain condition of children from the age of 9-10 to primary stages of adulthood using diffusion and functional Magnetic Resonance Imaging (dMRI and fMRI respectively), and by doing so analyzing the factors that impact different aspects of the young people's life such as the potential inclination to drugs. ABCD dataset is the fruit of this project. It contains dMRI and fMRI images of the cases under study. Also, these cases have answered some questionnaires on the Likert Scale, and their answers can reflect their viewpoint regarding drugs. Therefore, developing a method that can predict answers of a new case to the questionnaires based on their MRI features can be a solution to the goal of drug tendency prediction.

In ABCD data, the answers to the questions (labels) are on the Likert Scale where answers 1, 2, 3, 4, and 5 correspond to Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, and Strongly Agree respectively. This is neither a traditional regression problem nor a multi-class classification one because the typical regression and classification tasks don't consider the order relations between response variables. But in this study, the answers' values have order meanings. For example, if the correct answer is 5, predicting 4 should have a lower negative impact than predicting 1. Thus, the ordinal regression models should be used for such an answer prediction.

In recent research, the deep ordinal regression models have achieved much better results in various applications than traditional ordinal regression methods. However, these deep ordinal regression models cannot be applied to our study. Because the existing deep learning methods require large amount labeled data to train satisfied models but the ABCD dataset does not contain the answers for all the cases whose MRI data are available. This is not surprising, and at the most circumstances in biomedical applications, unlabeled data are abundant and labeled data are rare because in biomedical research providing labels is expensive or difficult as it need human expert supervision.

To address the above challenging problems, in this paper, we focus on designing new semi-supervised deep learning model that addresses the ordinal regression task, and at the same time, reduces the need for massive labeled training data. Our new approach shows superior performance on predicting the possible propensity of adolescents to marijuana using the diffusion MRI-derived mean diffusivity (MD) from 148 Regions of Interest (ROIs) of ABCD data.

2 Related Work

Ordinal Regression refers to the supervised machine learning problems in which the labels are categorical, and concurrently, the categories have meaningful order between them. In the literature, there are several types of methods proposed for ordinal regression problems.

In the first category, the methods try to address the problem from the regression perspective. They learn some mapping function that maps the samples to real numbers and suggest a way to find some decision boundaries to determine the rank of a given sample based on the interval that its mapped value lies in [15]. In the second category, the related methods try to reformulate the ordinal regression problem so that it enables us to leverage the power of prominent classification methods. They split the problem into a sequence of binary classification sub-problems and determine the class of the input based on the aggregation of the answers of the binary classifiers [9,3,10].

However, all of the mentioned methods are based on handcrafted features. In recent years, deep learning has obtained the state of the art results on different tasks in classification such as object recognition [7]. The main reason for its success is its superior ability to learn how to extract useful features for classification. As a result, modern approaches to the ordinal regression problem have focused on designing their methods based on deep learning. Niu *et al.* [13] proposed the first solution to ordinal regression using deep learning in the context of age estimation. They converted the problem with R possible ranks to R - 1 binary classification problems such that the *i*-th problem determines whether the rank of a sample is bigger than *i* or not. They have used a Convolutional Neural Network (CNN) as the feature extractor for their network. Liu *et al.* [11] suggested the second deep learning based method for ordinal regression based on the idea of the first category by mapping the samples to real numbers, but they did not use these real numbers to determine the rank of the samples. Rather, they proposed a loss function on these numbers to show the network the natural order between the ranks. Similar to [13], Liu *et al.* used CNN for feature extraction, and they called their method as CNNPOR.

Semi-supervised learning aims to reduce the need for labeled data for training models, especially deep neural networks. The main importance of semi-supervised techniques is in the areas that labeled data is scarce and there is ample unlabeled data. Numerous methods have been proposed in this regard, and the general idea of almost all of them is to add a term to the loss function that is calculated using the unlabeled data set that ultimately benefits generalization of the trained network [14,8,16,1,4,17]. Sajjadi *et al.* [14], Laine and Aila [8], and Tarvainen and Valpola [16] addressed consistency regularization for semi-supervised training. Berthelot *et al.* [1] applied sharpening to enforce the predictions to have lower entropy and impose entropy regularization [4]. Verma *et al.* [17] and Berthelot *et al.* [1] used MixUp idea to make the network more robust. The problem is that these methods both are originally intended for image data or their performance is validated on image data. In addition, they have been used to improve the classification task. Our method addresses the semi-supervised training in the ordinal regression task.

3 Proposed Method

3.1 Motivations and Model Design

In this section, we propose our method for solving an ordinal regression problem with the semi-supervised mechanism. Because our problem is semi-supervised learning, we first explain how to use the labeled set, *i.e.* supervised learning, and then we will focus on incorporating the unlabeled data to enhance the performance of our model.

Our method should use the labeled data of the training set to not only provide the neural network model the corresponding rank of each input but also suggest a trick to teach the natural order between the possible ranks to it. We address the former by framing it as a classification task and do so for the latter by introducing a mapping and imposing an order between mapped values of inputs from different ranks.

To leverage the unlabeled part of the training set, we add new terms to the loss function of the supervised training so that it encourages the network to make more consistent predictions, predict more confident scores, and have convex behavior with the inputs and their corresponding labels. We show a quantitative description of our approach in the following subsections.

3.2 Problem Formulation

Let us consider an ordinal regression problem with rank R and a set of labeled samples $L = \{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ where $\mathcal{Y} = \{1, 2, ..., R\}$ along with a set of unlabeled ones $U = \{(x_i) | x_i \in \mathcal{X}\}$. We denote the subset of samples with rank k as $\mathcal{X}_k$. . We show the development of our loss function step by step, and then we will provide our proposed architecture.

3.3 Loss Functions

Cross Entropy Loss. We convert the labels in the labeled part of the training set into the one-hot format (a vector with the length R and its k-th element being one and other elements being zero if the rank of a sample is k) and define a part of the loss function as the cross-entropy between the targets and softmax of the outputs of the network to minimize the KL divergence between the target distribution and $P_{model}(y|x;\theta)$ where θ is the vector of the model parameters.

$$\mathcal{L}_{CE} = \frac{1}{|L|} \sum_{(x_i, y_i) \in L} H(y_i, P_{model}(y|x_i;\theta)) \,. \tag{1}$$

Ordinal Loss. To guide the network to learn the natural order between the ranks, we map the samples to real numbers and define a loss over the mapped values of the samples from different ranks. Considering the activations of the penultimate layer of the deep neural network for an input x as $f(x)$, we define a linear mapping $\mathcal{M}(f(x))$ from f to real numbers. Then, we enforce the network to generate larger values for the samples of the class k than the values for the ones of the class $k-1$. To do so, given a batch of samples $X_k \subseteq \mathcal{X}_k$ and $X_{k-1} \subseteq \mathcal{X}_{k-1}$ we add the following term to the loss function:

$$\mathcal{L}_{Ordinal} = \sum_{k=2}^{R} \sum_{\substack{x_k \in X_k \\ x_{k-1} \in X_{k-1}}} ReLU(1 - \mathcal{M}(f(x_k)) + \mathcal{M}(f(x_{k-1}))) \tag{2}$$

Where $ReLU(x)$ is Rectified Linear Unit [12] with output equal to x given $x > 0$ and zero otherwise. The advantage of this loss function is that it only needs to consider pairwise comparison between adjacent ranks because if $\mathcal{M}(f(x_{k-1})) < \mathcal{M}(f(x_k))$ and $\mathcal{M}(f(x_k)) < \mathcal{M}(f(x_{k+1}))$, then $\mathcal{M}(f(x_{k-1})) < \mathcal{M}(f(x_{k+1}))$ $(x_i \in \mathcal{X}_i)$, i.e, adjacent ranks comparison implies farther ranks comparison.

3.4 Semi-Supervised Learning

Consistency Regularization. Consistency regularization aims to make the prediction of the network for a sample and its augmented versions, varieties of the sample that have the same conceptual meaning in the problem context, as close as possible. For example, in image classification, the classifier should output the same distribution for an image that is a rotated version of the original one because rotation does not change the class of a sample. To apply consistency regularization, we produce a guessed label for each unlabeled sample in two steps. At first, we generate several augmented samples from the unlabeled sample by adding Gaussian noise with a small variance to it. After that, we enter the original sample as well as augmented ones to the network and determine the average of the output distributions of the network as the guessed label for the original sample.

One of the techniques for training consistent classifier in a semi-supervised training is to motivate the network to show convex behavior in its predictions, i.e make a similar prediction for a linear combination of two unlabeled samples to the same linear combination of its predictions for them. [17] To implement this idea, we generate new samples by mixing samples in the dataset. Given two samples $(x_1, p_1), (x_2, p_2)$, we produce (x_3, p_3) as following:

$$\beta \sim Beta(\alpha, \alpha) \tag{3}$$

$$\beta := max(1 - \beta, \beta) \tag{4}$$

$$x_3 = \beta * x_1 + (1 - \beta) * x_2 \tag{5}$$

$$p_3 = \beta * p_1 + (1 - \beta) * p_2 \tag{6}$$

Entropy Minimization. Entropy minimization idea is originated in the information theory context where the uncertainty of a distribution is measured with its entropy. As a result, minimizing the entropy of the network output distribution is equivalent to enforcing the network to make more confident predictions, and we use sharpening to do so. If we denote the network output distribution prediction with vector p, the sharpened vector q gets calculated as following:

$$q_i = \frac{p_i^{\frac{1}{T}}}{\sum_{j=1}^{R} p_j^{\frac{1}{T}}} \tag{7}$$

where R is the length of p (number of the possible ranks in the ordinal regression problem), and T is the distribution temperature.

Based on the above ideas and to leverage their advantages, we use the Algorithm 1 to prepare inputs for our loss function for the semi-supervised training that we will introduce in the next subsection.

Algorithm 1 Mixing Up Labeled and Unlabeled Set

```
Input: A set of labeled samples $\mathcal{L} = \{(x_{l_i}, y_i)\}$, a set of unlabeled samples $\mathcal{U} = \{(x_{u_j})\}$ for $1 \leq i, j \leq N$, $\alpha$, $T$, Gaussian noise standard deviation $\sigma$, number of augmentations $M$, Deep Network $net$
for $m = 0$ to $M$ do
    if $m$ is 0 then
        $y_{pred_{j,m}} = net(x_{u_{j,m}})$
    else
        $noise \sim Gaussian(0,\, \sigma)$
        $x_{u_{j,m}} = x_{u_j} + noise$
        $y_{pred_{j,m}} = net(x_{u_{j,m}})$
    end if
end for
for $j = 1$ to $N$ do
    $y_{pred_j} = Average(y_{pred_{j,0}}, \ldots, y_{pred_{j,M}})$
    $y_{guess_j} = Sharpen(y_{pred_j}, T)$
end for
$\mathcal{U}_1 = \{(x_{u_j}, y_{guess_j}) \mid 1 \leq j \leq N\}$
$\mathcal{C} = Shuffle(Concatenation(\mathcal{L}, \mathcal{U}_1))$
$\hat{\mathcal{L}} = \{(MixUp(\mathcal{L}, \mathcal{C}[1:N])\}$
$\hat{\mathcal{U}} = \{(MixUp(\mathcal{U}_1, \mathcal{C}[N+1:2N])\}$
Return $\hat{\mathcal{L}}, \hat{\mathcal{U}}$
```

Semi-Supervised Training. Now we introduce the loss functions that we use to perform the semi-supervised training.

$$\mathcal{L}_l = \frac{1}{|\hat{\mathcal{L}}|} \sum_{(x_i, y_i) \in \hat{\mathcal{L}}} H(y_i, P_{model}(y|x_i; \theta)) \tag{8}$$

$$\mathcal{L}_u = \frac{1}{R|\hat{\mathcal{U}}|} \sum_{(x_j, y_j) \in \hat{\mathcal{U}}} \|y_j - P_{model}(y|x_j; \theta)\|_2^2 \tag{9}$$

In these equations, $\mathcal{L}_l$ has the notion of consistency regularization, and $\mathcal{L}_u$ aims to push the network to show convex behavior.

3.5 Proposed Loss Function and Model Architecture

We train our model based on the following loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + c_1 * \mathcal{L}_{Ordinal} + c_2 * \mathcal{L}_l + c_3 * \mathcal{L}_u \tag{10}$$

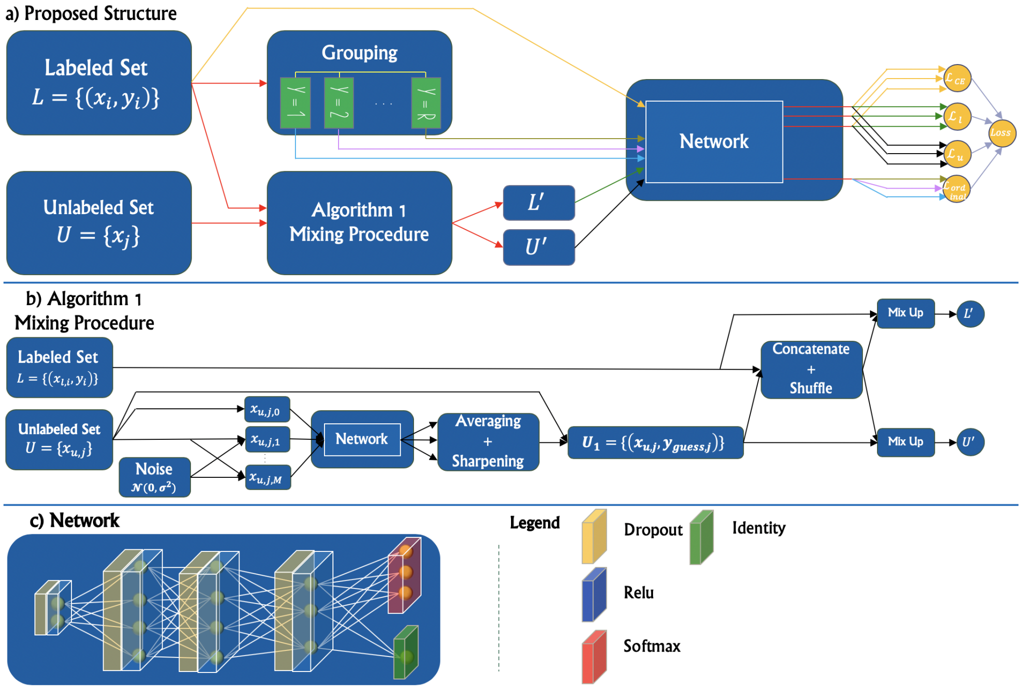

Fig. 1. The architecture proposed for ORSS problem. **a)** The path for calculation of each part of the loss function is specified with a separate color. **b)** Block diagram of the Algorithm 1. **c)** The network used for prediction.

Where c_1, c_2, and c_3 are hyperparameters. This loss function combines all of the motivations for ordinal regression and semi-supervised learning mentioned in the above sections. As we discussed, our experiments were on non-image data. Therefore, we used Multi-Layer Perceptron as the DNN feature extractor. Because our method solves the ordinal regression problem in a semi-supervised setting, we name it ORSS. Figure 1 shows the proposed procedure for training the network. We have shown the calculation path for each loss function with distinct colors. For example, the path used to compute the cross-entropy loss is the yellow one in Fig.1.a.

4 Experiments

We evaluated our method on 10731 subjects (mean age=118.96 ± 7.5 months) from the ABCD cohort. We extracted mean diffusivity (MD) from 148 Regions of Interest (ROIs) for each subject. Then, we put these measurements in a vector with length 148 and concatenated with three extra measures to the vector: the mean MD for left hemisphere, the mean MD for right hemisphere, and the mean

MD for the whole brain. In aggregate, we obtained a MD vector with the length of 151 (=148+3) for each person, from the dMRI images.

ABCD dataset contains questionnaires containing questions that ask the opinion of people regarding statements about drugs on the Likert Scale. For example, one of the questions is 'Marijuana helps a person relax and feel better.' If one of the adolescents answers completely agree to this question, it may suggest that they may have an inclination to drugs in the future.

In our experiments, our goal was to predict the answers to the question mentioned above using brain MRI features. As the answers are on the Likert scale, our task is Ordinal Regression, and because we did not have answers for some people whom we had dMRI features of them, we tried semi-supervised training to enhance the performance of our model. In summary, we had dMRI features for 10731 people (a vector of length 151 for each person). Among them, we had answers of 3663 ones to the question (labeled), and 7068 were unlabeled. We randomly split the labeled part into train, validation, and test subsets with the ratio 0.7, 0.15, and 0.15 respectively.

We compared our model with multi-class logistic regression, K-nearest neighbor, thresholded ridge regression, thresholded lasso regression, and multi-layer perceptron (MLP) with softmax logistic regression loss as classification baselines; Label propagation [2] and MixMatch as the semi-supervised training baselines; and we replaced the CNNPOR structure of the Liu *et al.* approach [11], specified as MLPPOR, as the ordinal regression baseline. We used to metrics to compare the methods. The first one is accuracy which is standard metric for classification networks, and also, we used Mean Absolute Error (MAE) which enables us to compare the performance of the methods in terms of the distance that their prediction has to the correct class label which is important in ordinal regression tasks because as we mentioned earlier, if the correct label is 5, predicting 4 is better than predicting 1 when the classes have natural order between them.

Table 1 summarizes the results from different methods. As can be seen, our method outperforms all other methods when comparing by both accuracy and MAE. Having lower MAE compared to MLPPOR which is the state of the art ordinal regression method shows that our method can effectively learn the order between classes. In addition, our method has better accuracy performance which shows that it can properly employ the unlabeled data to build a better classifier. For deep models, we used 5 different random seeds for initialization and reported the average of the results as the performance of the model.

We observed that using Dropout [5] ($p = 0.5$) in the input layer improves the performance of the network, but applying Dropout for other layers had negative impacts on the performance. In addition, hyperparameter setting ($K = 5$, $alpha = 0.5$, $T = 0.5$, $c_1 = 2$, $c_2 = 1$, and $c_3 = 1$) yielded the best result when our metric was accuracy, and ($K = 5$, $alpha = 0.5$, $T = 0.5$, $c_1 = 2$, $c_2 = 1$, and $c_3 = 1$) was the best one when the metric was MAE. We employed Adam optimizer [6] with parameters learning rate = 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay equal to 0.0001.

Table 1. Results on the ABCD dataset.

Methods	Accuracy	MAE
Multi-Class Logistic Regression	34.7%	1.2309
K-nearest neighbour	33.3%	1.2818
Thresholded Ridge Regression	31.9%	1.1655
Thresholded Lasso Regression	31.6%	1.1582
Label Propagation (semi-supervised baseline)	33.4%	1.3255
MLP with softmax Logistic Regression Loss	35.7%	1.2432
MLPPOR [11]	36.5%	1.1380
MixMatch [1]	36.4%	1.1571
ORSS (Ours)	**38.6%**	**1.0810**

4.1 Ablation Study

In order to further analyze the importance of each term in the loss function, we removed each of them one at a time and examined the performance of the method. We did not perform all hyperparameter search again and used the best setting that we obtained for the 'MAE' metric above. At each time, we change one of the 'c_i's ($i = 1, 2, 3$) to zero while keeping all other hyperparameters unchanged. Again, we reported the results by averaging the performance of 5 different initializations. The results are shown in Table 2.

Table 2. Results on the ABCD dataset.

Ablation	Accuracy	MAE
$c_1 = 0$	36.5%	1.1869
$c_2 = 0$	36.2%	1.1570
$c_3 = 0$	37.1%	1.1680

5 Conclusion

In this paper, we proposed a new framework for semi-supervised training of an ordinal regression problem. We developed the idea behind each part of the proposed network and loss function extensively and showed that our method outperforms modern methods of ordinal regression and semi-supervised learning on the ABCD dataset. In future, we will investigate more advanced brain MRI features and conduct extensively experiments on more adolescent behavior correlations as well as evaluate the gender effect on this problem.

6 Acknowledgment

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study[1], held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under aware numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U01DA041123, U01DA041147. A full list of supporters is available at here[2]. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from the NIMH Data Archive Digital Object Identifier (DOI)[3]. (NDA Release 2.0). DOI can be found at here[4]. (NDA Release 2.0)

References

1. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems. pp. 5050–5060 (2019)
2. Chapelle, O., Schlkopf, B., Zien, A.: Semi-Supervised Learning. The MIT Press, 1st edn. (2010)
3. Frank, E., Hall, M.: A simple approach to ordinal classification. In: European Conference on Machine Learning. pp. 145–156. Springer (2001)
4. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: Advances in neural information processing systems. pp. 529–536 (2005)
5. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
9. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Advances in neural information processing systems. pp. 865–872 (2007)

[1] https://abcdstudy.org

[2] https://abdstudy.org/nih-collaborators

[3] https://doi.org/10.15154/1503885

[4] https://nda.nih.gov/study.html?tab=cohortid=693

10. Lin, H.T., Li, L.: Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Computation **24**(5), 1329–1367 (2012)
11. Liu, Y., Wai Kin Kong, A., Keong Goh, C.: A constrained deep neural network for ordinal regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 831–839 (2018)
12. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
13. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4920–4928 (2016)
14. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in neural information processing systems. pp. 1163–1171 (2016)
15. Shashua, A., Levin, A.: Taxonomy of large margin principle algorithms for ordinal regression problems. Advances in neural information processing systems **15**, 937–944 (2002)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems. pp. 1195–1204 (2017)
17. Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. arXiv preprint arXiv:1903.03825 (2019)