Quantifying Student Attention using Convolutional Neural Networks

Andreea Coajă[1][a] and Cătălin V. Rusu[1,2][b]
[1]*Department of Computer-Science, Babeş-Bolyai University, Romania*
[2]*Institute for German Studies, Babeş-Bolyai University, Romania*

Keywords: Machine Learning, Prediction, Image Recognition, Classification.

Abstract: In this study we propose a method for quantifying student attention based on Gabor filters, a convolutional neural network and a support vector machine (SVM). The first stage uses a Gabor filter, which extracts intrinsic facial features. The convolutional neural network processes this initial transformation and in the last layer a SVM performs the classification. For this task we have constructed a custom dataset of images. The dataset consists of images from the Karolinska Directed Emotional Faces dataset, from actual high school online classes and from volunteers. Our model showed higher accuracy when compared to other convolutional models such as AlexNet and GoogLeNet.

1 INTRODUCTION

E-Learning offers students a quick and comfortable way of accessing online teaching sessions and learn at their own pace, using a large variety of resources (Deng and Wu, 2018). It has in recent years seen rapid development, especially with the explosion in popularity of learning platforms such as Udemy and Coursera. The COVID-19 worldwide pandemic has additionally forced schools and universities to quickly adapt and change how they teach students, driving interest in E-Learning further.

At the same time society puts an ever-growing emphasis on social media over traditional, in-person interactions. Nowadays people are spending substantial amounts of time on social media, with 16 to 24 year olds investing the largest amount, around 3 hours a day (Padmanathan et al., 2020). This shift has also sparked an increase in interest into developing machine learning based systems that are able to augment these interactions, such as systems for detecting facial emotions. Choudhury et al. (Choudhury, 2019) offer a comprehensive review of how face detection methodologies have evolved in the last 40 years and which are the state-of-the-art techniques suitable for face detection. This suggests that artificial intelligence has the potential to further improve E-Learning in the near future.

In this new normal, heavily driven by online interactions, student engagement is crucial for successful learning (Ellis and Bliuc, 2019). It is however hard for students to pay attention without interruptions throughout an entire lecture. Direct student communication, quizzes, and project work together with the use of engaging teaching methodologies can partly help students stay attentive however student attention remains a challenge in the virtual classroom. This is further complicated by the fact that in online teaching there is no easy way for the instructor to see the entire classroom or communicate with individual students. This separation between instructors and students has changed the way instructors evaluate and teach students. Recent studies have shown that oral presentations and exams are no longer used by the instructors for evaluations, with importance given to more interactive teaching methods and project-related evaluations (Motogna et al., 2020). Such changes might not be sufficient to keep students continuously engaged and attentive as studies suggest that students not only do not pay attention in the first and last few minutes of a lecture but also they do not pay attention continuously for 10-20 min and their engagement level alternates between attentive and not attentive throughout a lecture (Bunce et al., 2010).

With studies also showing that student-centric interactions at different times throughout a lecture, not only decrease student attention lapses but also to increase interaction (Bunce et al., 2010), student attention during online lectures can be a good indicator of

[a] https://orcid.org/0000-0002-0024-4705
[b] https://orcid.org/0000-0002-2056-8440

Coajă, A. and Rusu, C.
Quantifying Student Attention using Convolutional Neural Networks.
DOI: 10.5220/0010816500003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 293-299
ISBN: 978-989-758-547-0; ISSN: 2184-433X

the impact, quality and efficiency of online teaching. Methods capable of quantifying the attention level of a student based on video information from online lectures can form the basis of a system that is able to help instructors adapt their style and methodology throughout a lecture (Robal et al., 2018).

1.1 Related Work

Attention tracking is a complex task that has been extensively studied, with several successful methods being developed in recent years (Massé et al., 2017; Robal et al., 2018). Smith et al. (Smith et al., 2003) proposed a system for analyzing human driver visual attention based on global motion and color statistics. The system computes eye blinking, occlusion and rotation information to determine with reasonable success the driver's visual attention level.

Another interesting approach is introduced in (Eriksson and Anna, 2015) where the authors tried to detect if a student is attentive or not by using two distinct face detection methods implemented in OpenCV: the Viola-Jones method and the multi-block local binary pattern. Both algorithms obtained similar values for sensitivity and precision on images where the subjects were required to only look towards the front of the lecture hall. This posed a limitation on their approach as it generated high numbers of false positives when subjects were repeatedly performing poses with their faces tilted downwards. This essentially suggests the subject is not attentive when he or she is looking down, which is not always the case as students could in such situations be taking notes.

Robal et al. (Robal et al., 2018) proposed an eye tracking system to detect the position of the subject's eyes in order to quantify attention. Two different approaches were tested: a hardware eye tracker, Tobii, and also two software trackers, namely WebGazer.js (Papoutsaki, 2015) and tracking.js (TJS) (Lundgren et al., 2015). The hardware-based system achieved the highest accuracy 68.2%, followed by TJS with a recorded average performance of 58.6%.

Another approach was proposed in (Deng and Wu, 2018), where a comparison between different combinations of machine learning algorithms, such as Principal Component Analysis, Gabor feature extraction, K-nearest neighbors, Naive Bayes and Support Vector Machine (SVM) is presented. The most accurate combination was found to be Gabor feature extraction and SVM, with an accuracy of 93.1%.

Deep learning is a class of machine learning algorithms with a deep topology that are used to solve complex problems. It has in the last decade gained a lot of attention and has been used in computer vision with a high degree of success (Pak and Kim, 2017). The most distinctive characteristic of a deep learning approach is that it can automatically extract the best image features required for recognition directly via training, without the need of domain expertise or hard coded feature extraction. Convolutional neural networks (CNNs) are special types of artificial neural networks that satisfy the deep learning paradigm and have won numerous image recognition competitions in recent years (Krizhevsky et al., 2012; Szegedy et al., 2015). Schulc et al. (Schulc et al., 2019) developed a deep learning approach that detects attention and non-attention to commercials using webcam and mobile devices. The model combined a CNN with long short-term memory (Hochreiter and Schmidhuber, 1997) and achieved an accuracy of 75%.

For the past few decades, banks of Gabor filters have been widely used in computer vision for extracting features in face recognition tasks. They are based on a sinusoidal function with particular frequency and orientation that allows them to extract information from the space and frequency domains of an image. A few works have explored the integration of Gabor filters and CNN with promising results (Alekseev and Bobe, 2019).

In this paper, we propose a system that uses a CNN for the task of detecting attentive and not attentive states during online learning. The input images are put first through a Gabor filter, which extracts intrinsic facial features and servers as the input for the CNN. The last layer is a SVM that predicts the label. Our contribution is as follows: (i) we have build a dataset containing images from real online lectures; (ii) we have showed that a convolutional neural network can be effectively applied to the task of quantifying attention; (iii) we have showed that our method has significantly better performance when compared to other approaches or well-known convolutional neural network models such as AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2015).

The rest of the paper is organized as follows. In the next section we discuss CNNs, Gabor filters and the SVM. Section 3 describes our dataset, highlights our proposed solution and reports the obtained results. A discussion with some conclusions follows in the last section.

2 THEORY

2.1 Convolutional Neural Networks

Neural networks have attracted great interest in the scientific community for decades as they are theoretically able to model any linear or nonlinear relationship between input and output given sufficient data. Recently deep learning and in particular CNNs consisting of large numbers of hidden layers area able to successfully select the best features of the input data directly by propagating the training of a classifier back through the convolution layers without relying on domain expertise (Pak and Kim, 2017; Krizhevsky et al., 2012).

CNNs are thus special models of deep multi-layer neural networks. They are inspired by the visual cortex of the human brain and have been proven to be particularly effective in image processing tasks, such a image classification and object recognition (Rawat and Wang, 2017). At its core a typical CNN consists of a stack of different layers, typically convolutional and pooling. The networks performs convolutions on input images with multiple filters to form feature maps. This is typically followed by a pooling operation, where only the relevant information of the feature maps are pooled together. Such layers eventually transition to a fully connected layer at the end to produce the final result of the task, such as a label.

The convolutional layer is a key part of a CNN, with most of the computational effort taking place here. It is essentially a feature extractor and is capable of extracting fundamental features from an image such as edges, objects or textures (Rawat and Wang, 2017). It consists of several filters that are the same size but smaller than the input image. Each filter slides across the input image step by step and the dot product between the input and filter is computed, which results in an activation map. Then these activation maps are added together to form the output of the layer.

The pooling layer performs a down-sampling operation and summarizes rectangular patches into single values. It reduces the number of parameters, the size and noise of data and retains only the relevant and important features of the input by applying different activation methods, such as average activation, which takes the average value of each patch in the feature map or maximum activation, which takes the maximum value. This essentially helps reduce overfitting and makes the output more invariant to position.

In image classification, a typical CNN transforms the original image layer by layer from pixel values into assigned scores for each class. Each unit in the final output layer stores thus a probability for one particular class.

2.2 Support Vector Machine

The support vector machine (SVM) was developed for binary classification and has shown significantly better classification performance on reasonably sized datasets (Vapnik, 1995). Given training data and its corresponding labels $(x_n, y_n), n = 1, \ldots, N, x_n \in \mathbb{R}, t_n \in \{-1, +1\}$ the objective of a SVM is to find the optimal hyperplane that separates two classes by solving the following unconstrained optimization problem:

$$\min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max\left(0, 1 - w^T x_n t_n\right) \quad (1)$$

where w is a weight vector, C is a parameter that determines the trade-off between the maximization of the margin and the minimization of the classification error. Since Equation 1 with the standard hinge loss is not differentiable a common alternative is based on the squared hinge loss (Rawat and Wang, 2017):

$$\min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max\left(0, 1 - w^T x_n t_n\right)^2 \quad (2)$$

Equation 2 is not only differentiable but at the same gives more weights to errors and is known as the L2-SVM.

2.3 Gabor Filters

Gabor filters were proposed in 1949 by Denis Gabor and are typically used for the analysis of two-dimensional signals such as images in order to find local regions that have certain frequencies. The use of Gabor filters is mainly motivated by the fact that they offer maximum resolution in both space and frequency and they can extract spatial frequency information with minimal uncertainty. They haven proven to be particularly useful in object detection (Jain et al., 1997), face recognition (Chung et al., 1999) and movement analysis (Chen and Kubo, 2007).

The Gabor function is essentially a sinewave modulated by a Gaussian:

$$g(x, y, \theta, \lambda, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \quad (3)$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$, λ and θ represent the wavelength and orientation of the sinusoidal factor, ψ is the phase offset, σ

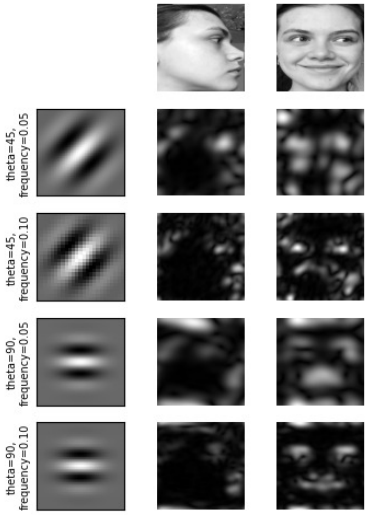

Figure 1: Examples of Gabor filters (left). Filtered images (right). The top row show the input images.

is the standard deviation of the Gaussian and γ is the spatial aspect ratio.

The response of a Gabor filter is computed by convolving the filter function with the image. Some examples of Gabor filters with different parameters and the corresponding filtered images are shown in Figure 1.

3 PROPOSED METHOD

3.1 Data

For other classification tasks such as facial expression recognition there are a few readily available datasets. For attention recognition there is however no publicly available dataset that contains data labeled as attentive and not attentive. We have manually built a dataset having three major sources of data: (i) from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist et al., 1998) we have taken pictures featuring front and side-views of subjects; (ii) online video recordings of lectures from the 'Eudoxiu Hurmuzachi' National College in Radauti, Romania; (iii) volunteers from the Babes-Bolyai University. The volunteers received specific instructions in advance. These instructions entailed the specific expressions they are required to pose during the photo session. Figure 2 shows a few examples of attentive and not attentive pictures for one particular subject.

After data collection was completed labels from one of the two categories were manually assigned to each picture in the dataset. When the level of attention could not be reliably determined, the respective image was discarded from the dataset. Table 1 summarises the eligible positions for each state.

Table 1: Eligible positions for attentive and not attentive states.

Attention
neutral emotion (front)
smile (front)
teeth smile (front)
frown (front)
neutral emotion, eyes to the left (front)
neutral emotion, eyes to the right (front)
smile, eyes to the left (front)
smile, eyes to the right (front)
teeth smile, eyes to the left (front)
teeth smile, eyes to the right (front)
hand on the forehead (front)
No Attention
looking down (front)
looking down (right-side)
looking down (left-side)
neutral emotion, looking forward (front)
smile, looking forward (right-side)
smile, looking forward (left-side)
hand over the mouth, looking forward (front)
hand over the mouth, looking down (front)
hand over the mouth, looking up (front)

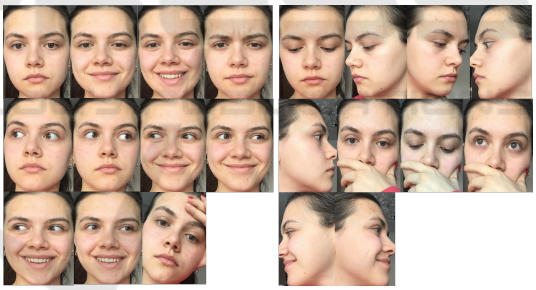

Figure 2: Attentive (left) and not attentive (right) positions.

3.1.1 Data Organization

The dataset thus combines images from actual online lectures with images from volunteers to obtain a total of 16347 photos. To avoid overfitting the dataset was artificially augmented using label-preserving transformations. We employed two distinct forms of data augmentation:

- rotate by -10 degrees
- rotate by 10 degrees

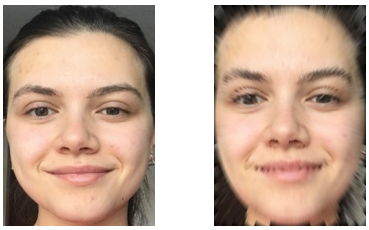

Figure 3: Original photo (left) Log-polar photo (right).

3.1.2 Data Preprocessing

The input images were scaled down to 50x50 and the faces of subjects were extracted using the DNN Face Detector from OpenCV (Liu et al., 2016). Because in some real life online learning scenarios images do not always have excellent quality we are also interested in seeing how a loss in image quality affects the classification accuracy of our model. To this end we have additionally created a low-resolution dataset where images were scaled down in size and quality using log-polar transformations. This was achieved by remapping the picture from a 2D Cartesian coordinate system (x, y) to a 2D log-polar coordinate system (ρ, θ) using the following equations:

$$\rho = \log\left(\sqrt{x^2 + y^2}\right) \tag{4}$$

$$\theta = \arctan\left(\frac{y}{x}\right) \tag{5}$$

Figure 3 shows such a compressed image together with the original.

3.2 Model

In this section we present our model based on a convolution neural network for attention detection. Our solution uses features based on Gabor filters instead of raw pixel values as the input for the CNN and in the last layer a L2-SVM to predict the label. The choice of implementing Gabor filters is biologically motivated since they are modelled after the response of cells in the early visual cortex. They additionally remove variations in light and are robust to shifts and deformations. Other studies have also explored Gabor filters for CNNs and have reported better performance across several datasets (Alekseev and Bobe, 2019). The architecture of the model, including the number of layers and the size of the kernels implemented in each layer are summarized in Table 2.

3.3 Results

Table 3 shows the average results for different values of hyper parameters in our model. The best combination of parameters achieved an accuracy of about 95%. The two-layer CNN performed slightly worse than the four-layer CNNs. With respect to kernel size, 3×3 was the best. All parameter variations were tested using the same dataset.

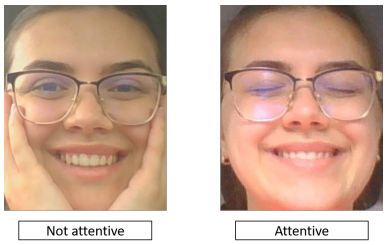

Figure 4: Samples of detection errors. Labels indicate the predicted (mistaken) class probability.

We divided the dataset into six sets: five sets were used for training and one for testing. We conducted 6-fold cross validation. The hyper parameters marked with "*" in Table 4 were used for the CNN model. All models were tested on both datasets: the original and the low resolution one obtained using the log-polar transformation. Table 4 summarizes our results. The CNN achieved a 95.15% accuracy, which is significantly higher than that of the AlexNet and higher than the one obtained by GoogLeNet. As expected all models performed worse on the LogPolar dataset however AlexNet showed a greater decrease in performance with about 15% compared to the 8% decrease for our model and 3% for GoogLeNet, which performed best on this dataset. Our results matched other studies that have also reported improved classification accuracy by replacing the softmax operator with a SVM (Rawat and Wang, 2017). Figure 4 shows samples of incorrectly labeled images by our model. These images would be a challenge to classify, even for human observers.

4 CONCLUSIONS

In this paper, we have outlined the effectiveness of convolutional neural networks in detecting attentive and not attentive states for online teaching. First, our results show that CNNs are capable of achieving very good classification results on datasets consisting of recordings in real-life teaching scenarios. Second, we created a proper dataset with images from actual online classes and volunteers. Third, we built a CNN model for classifying attentive and not attentive states and evaluated its performance. Lastly, we have found that our model outperforms other convolutional neural models such as AlexNet and GoogLeNet.

We are additionally working on implementing an application that is able to label attentive states in live

Table 2: Model architecture.

Number of layer	Type of layer
1	Input 50*50*3
2	Convolutional nr=96, size=3*3, kernel_initializer=custom_gabor
3	Convolutional nr=32, size=3*3
4	Max Pooling size=2*2
5	Convolutional nr=32, size=3*3
6	Convolutional nr=32, size=3*3
7	Max Pooling size=2*2
8	Batch Normalization
9	Flatten
10	Dense units=1024
11	Dropout rate=0.2
12	Dense units=1, linear, regularizer=l2(0.01)

Table 3: Experimental results for different hyper parameters.

Layers	Nr. of kernels	Size of kernels	Avg. accuracy
4*	96-32-32-32	3*3, 3*3, 3*3, 3*3	95.76%
4	96-32-32-32	5*5, 5*5, 5*5, 5*5	95.18%
3	32-32-32	3*3, 5*5, 5*5	93.37%
2	32-32	3*3, 3*3	95.63%

Table 4: Comparison between CNN, AlexNet and GoogLeNet on both datasets.

Method	Dataset	Accuracy
CNN	normal dataset	95.15%
GoogLeNet	normal dataset	92.54 %
AlexNet	normal dataset	85.75%
CNN	LogPolar dataset	87.97%
GoogLeNet	LogPolar dataset	89.05 %
AlexNet	LogPolar dataset	72.16%

streams of videos. It was successfully piloted this summer semester in the computer science department at the Babes-Bolyai University. As future steps, we plan to integrate such an implementation of our attention detector in online lectures and measure the impact of attention on the learning performance of students.

REFERENCES

Alekseev, A. and Bobe, A. (2019). Gabornet: Gabor filters with learnable parameters in deep convolutional neural network. In *2019 International Conference on Engineering and Telecommunication (EnT)*, pages 1–4. IEEE.

Bunce, D. M., Flens, E. A., and Neiles, K. Y. (2010). How long can students pay attention in class? a study of student attention decline using clickers. *Journal of Chemical Education*, 87(12):1438–1443.

Chen, Y.-W. and Kubo, K. (2007). A robust eye detection and tracking technique using gabor filters. In *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, volume 1, pages 109–112.

Choudhury, A. D. e. a. (2019). Evolution of face recognition technologies.

Chung, K.-C., Kee, S. C., and Kim, S. R. (1999). Face recognition using principal component analysis of gabor filter responses. In *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No.PR00378)*, pages 53–57.

Deng, Q. and Wu, Z. (2018). Students' attention assessment in elearning based on machine learning. In *IOP Conference Series: Earth and Environmental Science*, volume 199, page 032042. IOP Publishing.

Ellis, R. A. and Bliuc, A.-M. (2019). Exploring new elements of the student approaches to learning framework: The role of online learning technologies in student learning. *Active Learning in Higher Education*, 20(1):11–24.

Eriksson, J. and Anna, L. (2015). Measuring student attention with face detection:: Viola-jones versus multiblock local binary pattern using opencv.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jain, A. K., Ratha, N. K., and Lakshmanan, S. (1997). Object detection using gabor filters. *Pattern Recognition*, 30(2):295–309.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Lundgren, E., Rocha, T., Rocha, Z., Carvalho, P., and Bello, M. (2015). tracking. js: A modern approach for computer vision on the web. *Online]. Dosegljivo: https://trackingjs. com/[Dostopano 30. 5. 2016]*.

Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91(630):2–2.

Massé, B., Ba, S., and Horaud, R. (2017). Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2711–2724.

Motogna, S., Marcus, A., and Molnar, A.-J. (2020). Adapting to online teaching in software engineering courses. In *Proceedings of the 2nd ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence*, EASEAI 2020, page 1–6, New York, NY, USA. Association for Computing Machinery.

Padmanathan, P., Bould, H., Winstone, L., Moran, P., and Gunnell, D. (2020). Social media use, economic recession and income inequality in relation to trends in youth suicide in high-income countries: a time trends analysis. *Journal of affective disorders*, 275:58–65.

Pak, M. and Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, pages 1–3.

Papoutsaki, A. (2015). Scalable webcam eye tracking by learning from user interactions. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 219–222.

Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9):2352–2449.

Robal, T., Zhao, Y., WIS, T., Lofi, C., and Hauff, C. (2018). Towards real-time webcam-based attention tracking in online learning. In *ACM Annual Meeting of Interactive User Interfaces (IUI)*.

Schulc, A., Cohn, J. F., Shen, J., and Pantic, M. (2019). Automatic measurement of visual attention to video content using deep learning. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–6. IEEE.

Smith, P., Shah, M., and da Vitoria Lobo, N. (2003). Determining driver visual attention with one camera. *IEEE transactions on intelligent transportation systems*, 4(4):205–218.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Vapnik, V. (1995). *The nature of statistical learning*. Springer, Berlin, Germany.