# Evaluating Medical Aesthetics Treatments through Evolved Age-Estimation Models

Risto Miikkulainen
Cognizant AI Labs;
The University of Texas at Austin

Elliot Meyerson
Cognizant AI Labs
San Francisco, CA, USA

Xin Qiu
Cognizant AI Labs
San Francisco, CA, USA

Ujjayant Sinha
Cognizant Technology Solutions
New Delhi, India

Raghav Kumar
Cognizant Technology Solutions
New Delhi, India

Karen Hofmann
Cognizant Technology Solutions
Bethlehem, PA, USA

Yiyang Matt Yan
AbbVie, Inc.
Irvine, CA, USA

Michael Ye
AbbVie, Inc.
Irvine, CA, USA

Jingyuan Yang
AbbVie, Inc.
Irvine, CA, USA

Damon Caiazza
AbbVie, Inc.
Irvine, CA, USA

Stephanie Manson Brown
AbbVie, Inc.
Marlow, UK

## ABSTRACT

Estimating a person's age from a facial image is a challenging problem with clinical applications. Several medical aesthetics treatments have been developed that alter the skin texture and other facial features, with the goal of potentially improving patient's appearance and perceived age. In this paper, this effect was evaluated using evolutionary neural networks with uncertainty estimation. First, a realistic dataset was obtained from clinical studies that makes it possible to estimate age more reliably than e.g. datasets of celebrity images. Second, a neuroevolution approach was developed that customizes the architecture, learning, and data augmentation hyperparameters and the loss function to this task. Using state-of-the-art computer vision architectures as a starting point, evolution improved their original accuracy significantly, eventually outperforming the best human optimizations in this task. Third, the reliability of the age predictions was estimated using RIO, a Gaussian-Process-based uncertainty model. Evaluation on a real-world Botox treatment dataset shows that the treatment has a quantifiable result: The patients' estimated age is reduced significantly compared to placebo treatments. The study thus shows how AI can be harnessed in a new role: To provide an objective quantitative measure of a subjective perception, in this case the proposed effectiveness of medical aesthetics treatments.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*GECCO '21, July 10–14, 2021, Lille, France*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8350-9/21/07…$15.00
https://doi.org/10.1145/3449639.3459378

## CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; **Neural networks**; **Computer vision tasks**; • **Mathematics of computing** → **Hypothesis testing and confidence interval computation**; • **Applied computing** → **Health informatics**.

## KEYWORDS

Neural Architecture Search, Uncertainty Estimation, Real-world Applications, Medical Aesthetics

**ACM Reference Format:**
Risto Miikkulainen, Elliot Meyerson, Xin Qiu, Ujjayant Sinha, Raghav Kumar, Karen Hofmann, Yiyang Matt Yan, Michael Ye, Jingyuan Yang, Damon Caiazza, and Stephanie Manson Brown. 2021. Evaluating Medical Aesthetics Treatments through Evolved Age-Estimation Models. In *2021 Genetic and Evolutionary Computation Conference (GECCO '21), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3449639.3459378

## 1 INTRODUCTION

One of the most impressive areas of Deep Learning applications focuses on processing faces: recognition of individuals, understanding emotion and intent based on facial expressions, whether the person is paying attention, and even generation of synthetic but realistic face images [23, 30, 34, 55, 57]. While some of this work is controversial because it can lead to biases and raise privacy issues, much of it can be beneficial to the society, for instance improving medical diagnosis, patient monitoring, and safety [16, 33, 39, 49].

One such application is presented in this paper: Evaluating effectiveness of medical aesthetics treatments. They involve injecting a toxin or a filler treatment in targeted areas of the face, altering the skin texture or other facial features [1, 2]. They can be part of a treatment for facial injuries and diseases such as facial palsy, or elective procedures to improve appearance, including lower perceived age. Success is difficult to measure, and subjective in nature.

This paper proposes a novel approach to this problem: an AI-based method for measuring the results objectively and quantitatively. The starting point is a state-of-the-art neural network trained in the age estimation task. The paper improves the state-of-the-art in three ways: (1) Instead of celebrity face datasets used in prior work [47], a dataset of actual clinical patients is used to make the training more accurate; (2) While existing networks were initially designed for other visual tasks, their design, including training setup, data augmentation, and architecture, is customized specifically for this task through evolutionary optimization [36]; and (3) the RIO method, based on a Gaussian Process model of residual errors [41], is used to measure confidence in the age estimates. In a series of experiments, each of these elements were found to improve the accuracy of the age estimation process.

The method was then applied to a clinical trial dataset where one group of patients was given a placebo injection, while other two groups were given different versions of the active injectable tratment. The neural network was used to estimate the patients age before the treatment and at several time points up to 180 days after the injection. The main finding is that the treatment has a quantifiable result: Patients with active treatment appeared consistently chronologically younger than those with placebo injections.

This study thus demonstrates how current AI can be used in a new role: as a mechanism for objective quantitative evaluation of a measure with high degree of subjectivity. Similar approaches can potentially be used in the future to automate evaluations that would otherwise be difficult to quantify, making results more consistent and reliable.

## 2 BACKGROUND

This section begins with an overview of the goals, treatments, and challenges in medical aesthetics. Deep learning approaches to age estimation are then reviewed, including the importance of obtaining a realistic dataset. Deep learning models can be improved through neural architecture search, and they can be more useful in practical application through methods that estimate confidence in their output.

**Aging and Aesthetic Medicine:** Aging is a complex biological process that is not fully understood, however, we are in constant pursuit of ways to live a longer, healthier life. In doing so, the concept of aging well is central in the field of aesthetic medicine. While aging well does not necessarily equate to living a longer, healthier life, it does help the aging population cope with aging and is a strong contributor to improved self-esteem.

Aesthetic medicine utilizes a set of invasive (e.g. surgical) and noninvasive (e.g. injectable, laser) techniques to improve appearance of the patient [1, 2, 15] in support of the concept of aging well. The global market for aesthetic medical procedures is valued at $86.2 billion in 2020, with 52.5% of the revenue share in noninvasive procedures, and it is expected to expand by 9.8% per year in the next eight years [15]. Injections of botulinum toxin type A (Botox® Cosmetic) and hyaluronic acid fillers are some of the most common noninvasive procedures, aiming at correcting facial texture and volumizing muscle tone. Facial treatments target several anatomical indications such as forehead, tear troughs, nasolabial folds, and lips.

Age-related issues, including reduction of age appearance (perceived age) to match how a patient feels inside, can be the goal for many patients who elect noninvasive aesthetic medicine procedures [25]. One of the challenges is that success is often subjective and difficult to quantify when comparing age appearance to biological age. With more objective age perception measures, it might be possible to select the appropriate procedures more accurately and reach medical and personal goals more reliably.

A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation found nine different patient-reported outcome (PRO) surveys used in the assessment of post-treatment patient satisfaction [27]. It concluded that a new patient-report tool was needed to help measure satisfaction with facial appearance following aesthetic procedures. In response to these findings the authors developed FACE-Q, a systemized and standardized PRO tool. The development and validation of FACE-Q improves the industry's knowledge around patient satisfaction, but as with all survey-based tools, it remains largely subjective.

Breakthrough advancements in computer vision and access to rare and private datasets offer new approaches in medical image assessment. A tool that objectively evaluates the age appearance of individuals can become a starting point towards further AI evaluations of patients in the healthcare industry.

**Age Estimation with Deep Learning:** Age estimation from facial images has been used as a benchmark task to evaluate various deep learning approaches. For instance, Rothe et al. [47] introduced the IMDB-WIKI dataset of celebrity face images for age estimation and showed that deep learning could be used to do it well, using the classic VGG-16 network [50] as a base model. Follow-up work expanded these results [56], by applying modern architectures such as DenseNet [20] and MobileNet [19] to the problem and showing the utility of more compact customized architectures. It has proven difficult to improve upon these results partly because of the nature of the dataset: The celebrity face images are often taken with an application of significant amounts of make-up, presumably sometimes also after medical aesthetics treatments, and the images have often been retouched to improve appearance. Such alterations make age estimation difficult and unreliable, and the results do not generalize well to actual medical datasets. Thus, in order to achieve sufficient accuracy and reliability to evaluate treatment effects, obtaining and utilizing a dataset of more realistic face images is crucial.

**Neural Architecture Search (NAS):** The age estimation architectures used in the above studies were originally developed for the standard image classification tasks such as CIFAR-10 and Imagenet. They form a good starting point for other computer vision tasks, and have been used in applications such as object detection, X-ray processing, and even malware detection [42, 45, 46], in addition to age estimation. Such architectures are large and expensive to train, and therefore their design is usually not optimized extensively to the new task. However, recent results show that large gains are possible by neural architecture search, i.e. by customizing the network design to the task [8]. Many NAS techniques use gradient descent, reinforcement learning, or Bayesian parameter optimization, and focus on hyperparameter tuning. In contrast, evolutionary NAS can optimize network design more broadly, including network structure, data augmentation, activation and loss functions, and

learning methods. In several studies, such evolutionary optimization has been shown to improve performance in image, text, X-ray, and face-feature classification tasks [3, 12, 32, 35, 43, 44]. It will be used in this paper to customize state-of-the-art computer vision architectures to age estimation.

**Uncertainty Estimation:** When AI systems are deployed in real-world applications, interesting challenges emerge that are sometimes overlooked in the laboratory. Trustworthiness is one such dimension: for instance, when a neural network makes a prediction, such as estimated age, it is not enough that the prediction is as accurate as possible; it is also important to know how reliable the prediction is, i.e. what the confidence intervals are around it. Several techniques have been developed that return confidence intervals in addition to the predicted value, combining Bayesian reasoning or Gaussian Processes with Neural Networks [9, 10, 24, 28, 29, 40]. These methods require significant modifications to the model infrastructure and training pipeline, and are complex to implement and expensive to train.

In contrast, the RIO technique [41] allows deriving confidence estimates on any pretrained point prediction neural network (or other point prediction model). The idea is to train a separate Gaussian Process (GP) to model the residual errors of the network, with a kernel that combines both the input and output of the network. The GP then provides the confidence intervals for any future samples given to the network. Furthermore, the mean of the GP distribution can be used to fine-tune the output of the network, making it more accurate. RIO was shown more accurate than other methods (SVGP, ANP, and NNGP; [18, 24, 29]) in a number of benchmark tasks. It also improved accuracy of DenseNet in the IMDB age-estimation task, reducing MAE from 7.43 to 6.35, while providing accurate confidence intervals.

Uncertainty estimation is particularly important in the application presented in this paper. In order to demonstrate the benefit of medical aesthetics treatments, the estimated improvements must be statistically reliable. RIO provides a way of estimating such reliability.

## 3 EVOLVING AGE ESTIMATION NETWORKS

This section describes the real-world age estimation problem that needed to be solved, the vision models that were used to tackle this problem, and how evolution was applied to optimize these models.

### 3.1 Problem Setup

The goal is to produce a model that, given an input image, estimates the age of the individual in the image, such that the overall mean absolute error (MAE) of the model is minimized. Formally, the problem consists of a dataset $\mathcal{D} = (\mathcal{X}, \mathbf{y}) = \{(\mathbf{X}_i, y_i)\}_{i=1}^{N}$ of $N$ RGB images $\mathbf{X}_i \in \mathbb{R}^{h \times w \times 3}$ and their corresponding integer age labels $y_i \in [a_{\min}, \ldots, a_{\max}] \subset \mathbb{N}$, where $h$ and $w$ are the height and width (in pixels) of each image, and $a_{\min}$ and $a_{\max}$ are the minimum and maximum ages (in years) of individuals for which predictions need to be made. The goal is to find a prediction model $\mathcal{F}(\mathbf{X}) = \hat{y}$ that minimizes MAE, i.e., minimizes

$$\mathcal{L}_{\text{MAE}}(\mathcal{F}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \Big| \, y_i - \mathcal{F}(\mathbf{X}_i) \, \Big|. \tag{1}$$

Age estimation is a regression problem, with the additional structure that each label is an integer. This structure can be exploited by extending methods developed for classification. The most popular state-of-the-art such models have been developed for visual multi-class classification problems. Therefore, they can be most directly adapted to age estimation by treating each integer age as a class and having the last layer of the model output a probability $p(a \mid \mathbf{X})$ for each age $a \in [a_{\min}, \ldots, a_{\max}]$ via a softmax. This model can then be trained using the standard multi-class classification loss, i.e., the cross-entropy loss

$$\mathcal{L}_{\text{CE}}(\mathcal{F}, \mathcal{D}) = -\frac{1}{N} \sum_{i=1}^{N} \log(p(y_i \mid \mathbf{X}_i)), \tag{2}$$

from which age estimates can be deduced by computing the expected value over ages

$$\mathcal{F}(\mathbf{X}) = \sum_{a=a_{\min}}^{a_{\max}} a \cdot p(a \mid \mathbf{X}). \tag{3}$$

Importantly, $\mathcal{L}_{\text{MAE}} = 0 \iff \mathcal{L}_{\text{CE}} = 0$ and $\lim_{\mathcal{L}_{\text{CE}} \to 0} \mathcal{L}_{\text{MAE}} = 0$, making cross-entropy a valid and intuitively reasonable loss for this problem. Notice that, since stochastic gradient descent can be performed on both, plugging Eq. 3 into Eq. 1 also suggests a possible training loss, i.e.,

$$\mathcal{L}_{\text{MAE}}(\mathcal{F}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \Big| \, y_i - \sum_{a=a_{\min}}^{a_{\max}} a \cdot p(a \mid \mathbf{X}) \, \Big|. \tag{4}$$

Opportunities like these different possible losses define the dimensions that evolution can explore to optimize these models.

### 3.2 Evolutionary Optimization Method

Experiments were run using the LEAF platform for evolutionary optimization [32, 36, 38]. The experiments utilized hyperparameter optimization of learning algorithm and data augmentation parameters, population-based training (PBT) [22, 31] to integrate training and evolution, a domain-specific version of loss-function optimization [12, 13], and ensembling of evolved solutions. Inside of LEAF, hyperparameter optimization is performed via the genetic algorithm component of CoDeepNEAT [36], which includes standardized mutation and crossover operators for handling continuous, integer, Boolean, and categorical (i.e., Enum) parameters within hierarchical structures [32]. Using PBT means that models in evolution can be initialized with the weights trained in previous generations [22], so training from scratch is not required; this technique yields substantial computational savings. Loss-function optimization evolves the structure of the loss surface that stochastic gradient descent traverses, thus it can be viewed as evolving the training environment to align with application goals. The ensembling process is implemented in the standard way, i.e., by averaging model predictions, taking advantage of the fact that the complete evolutionary process yields multiple high-performing models, which can have complementary behaviors.

The space of possible parameter values for the domain is shown in Table 1. The parameters span various data types and are grouped into classes based on how they affect the model: *Opt* parameters relate to the backpropagation process, *Aug* parameters relate to

| Parameter | Possible Values | Type | Class |
|---|---|---|---|
| Algorithm | [adam, rmsprop] | Enum | Opt |
| Initial Learning Rate (LR) | [1e-5, 1e-3] | Float | Opt |
| Momentum | [0.7, 0.99] | Float | Opt |
| (Weight Decay) / LR [26] | [1e-7, 1e-3] | Float | Opt |
| Patience (Epochs) | [1, 20] | Int | Opt |
| SWA Epochs [21] | [1, 20] | Int | Opt |
| Rotation Range (Degrees) | [1, 60] | Int | Aug |
| Width Shift Range | [0.01, 0.3] | Float | Aug |
| Height Shift Range | [0.01, 0.3] | Float | Aug |
| Shear Range | [0.01, 0.3] | Float | Aug |
| Zoom Range | [0.01, 0.3] | Float | Aug |
| Horizontal Flip | {True, False} | Bool | Aug |
| Vertical Flip | {True, False} | Bool | Aug |
| Cutout Probability [7] | [0.01, 0.999] | Float | Aug |
| Cutout Max Proportion [7] | [0.05, 0.5] | Float | Aug |
| Pretrained Base Model | Keras App. [5] | Enum | Arch |
| Base Model Output Blocks | {B0, B1, B2, B3} | Subset | Arch |
| Loss function $\lambda$ in Eq. 5 | [0, 1] | Float | Arch |

**Table 1: *Space of possible parameter values for evolution.* The parameters span various data types and are grouped into classes based on how they affect the model: *Opt* relate to the backpropagation process; *Aug* relate to data augmentation; and *Arch* relate to the structure of the model function. Together they form a comprehensive approach to optimizing neural network designs through evolution.**

data augmentation, and *Arch* parameters relate to the structure of the model function.

Unless otherwise cited, the data augmentation parameters can be found in Keras' `ImageDataGenerator`. The pretrained base models can be selected from Keras' applications module. The Base Model Output Blocks refers to the set of layers of the base model that are concatenated before being fed into the final classification layer. B0 is the output of the final layer of the base model, and B1-B3 are the outputs of the three preceding blocks of layers. As a single-parameter domain-specific variant of loss function optimization, the model can tradeoff between Eq. 2 and Eq. 4 via

$$\mathcal{L}_{\text{Opt}}(\mathcal{F}, \mathcal{D}) = \lambda \cdot \mathcal{L}_{\text{CE}}(\mathcal{F}, \mathcal{D}) + (1 - \lambda) \cdot \mathcal{L}_{\text{MAE}}(\mathcal{F}, \mathcal{D}), \quad (5)$$

for a tunable parameter $\lambda \in [0, 1]$. In the space of age probabilities, $\mathcal{L}_{\text{CE}}$ has a unique global optimum at $p(a = y_i \mid \mathbf{X}_i) = 1$, while $\mathcal{L}_{\text{MAE}}$ has multiple alternative global optima, e.g., $p(a = y_i - 1 \mid \mathbf{X}_i) = 0.5$ with $p(a = y_i + 1 \mid \mathbf{X}_i) = 0.5$. However, $\mathcal{L}_{\text{MAE}}$ has the intuitively satisfying property that mistakes closer to the correct age have lower loss, e.g., $p(a = y_i - 1 \mid \mathbf{X}_i) = 0.5$ with $p(a = y_i \mid \mathbf{X}_i) = 0.5$ has lower loss than $p(a = y_i - 10 \mid \mathbf{X}_i) = 0.5$ with $p(a = y_i \mid \mathbf{X}_i) = 0.5$. The parameter $\lambda$ allows the model to potentially exploit both properties: uniqueness of optima and locality. Models were trained with Keras [5]. The learning rate was decayed by a factor of 10 whenever the patience was exceeded w.r.t. validation MAE.

Notice that, since all the evolved parameters are architecture-agnostic, the base model can be changed during evolution. In particular, it is possible to start evolution with a relatively lightweight base model and scale up in stages (S*) as evolution converges, picking up from where it left off. This stage-wise approach is similar to scale-agnostic methods that have achieved state-of-the-art NAS performance for largescale models in prior work [43, 53, 58]. Such efficiency is critical in largescale real-world applications such as age estimation, where the final deliverable model is very expensive to train and evaluate.

## 3.3 Datasets

The methods were evaluated on two datasets of different sizes (Table 2). The smaller dataset (D0) consisted of 10,837 training images and 2692 test images, with ages ranging from 18 to 79. The age classification layer of the models for this dataset thus have 62 units (i.e., classes), one for each age. These images were of 3,719 unique patients before treatment, with multiple images per patient, e.g., with differing face angle or facial expression. The larger dataset (D1) consisted of 18,537 training images and 3733 test images, with ages ranging from 18 to 80 (yielding 63 age classes), and 5,998 unique patients. D1 is a scaled-up and refined version of D0, in which more studies and patients were added and potentially misleading images were systematically removed. For both datasets, the training and test sets were split so that no patient appeared in both sets. The raw images for each dataset were very high resolution (≈6000×4000 pixels) pre-treatment images from various clinical trials and were downsampled to varying degrees depending on the architecture. All images in D0 and D1 are of patients before any treatment has been applied, so the models can learn to estimate age without conflating age with treatment effects.

The value of such a realistic dataset was demonstrated in a preliminary experiment. A version of DenseNet-121 was trained on the IMDB dataset, resulting in validation error of 7.43 years. This result is similar to prior results on the IMDB and WIKI datasets with several architectures such as DenseNet, MobileNet, and SSR-Net [56]. In contrast, a related architecture, DenseNET-169, in multiple instantiations achieved a significantly lower validation error, down to 3.65. These results suggest that such more realistic datasets serve as a better foundation for building age-prediction models.

## 3.4 Optimization

For D0, evolution was run with a population size of 20 for 25 generations using ResNet-50 [17] as the base model (S0) followed by 25 generations using DenseNet-121 [20] with each candidate trained for 20 epochs (S1). As a rule-of-thumb, a population size of 20 is the minimum required for producing reliable results using LEAF while minimizing computational cost. In both S0 and S1, performance plateaued by the 25th generation. Since D0 is a smaller development dataset, the goal was to evaluate the basic evolutionary method, so no PBT or loss function optimization was used, and image resolution was kept fixed at $224 \times 224$ for a fair comparison to Human Design. The results are shown in Figure 1.

RN-50 Base is the ResNet-50 baseline without evolution, i.e., the best fitness in the initial population. The Human Design is the best model for this application and dataset developed by professional data scientists. For D0, this design is based on a ResNet-50 base model. This model is substantially better than the baseline, and evolution further improves upon it substantially.

| Dataset ID | Application | # Studies | # Images | # Patients ∩ D0 | # Patients ∩ D1 | # Patients ∩ D2 |
|---|---|---|---|---|---|---|
| D0 | Age Estimation | 9 | 13,529 | 3719 | 3415 | 0 |
| D1 | Age Estimation | 19 | 22,270 | 3415 | 5998 | 0 |
| D2 | Treatment Evaluation | 1 | 77,914 | 0 | 0 | 787 |

**Table 2: *Dataset Details.* Three datasets were used in the experiments in this paper, each consisting of high-resolution clinical photographs of patient faces. D0 and D1 were used for developing the age estimation model, while D2 was used for evaluating treatment outcomes. D0 and D1 contain high quality images from multiple clinical studies. D1 is a scaled-up and refined version of D0, in which more studies and patients were added and potentially misleading images were systematically removed. D2 contains data from a single Botox study, and has no overlap with D0 or D1. D2 contains many fewer patients than D0 and D1, but many more images, because it contains multiple images of the same patient over time after the treatment, whereas D0 and D1 consist only of pre-treatment images.**

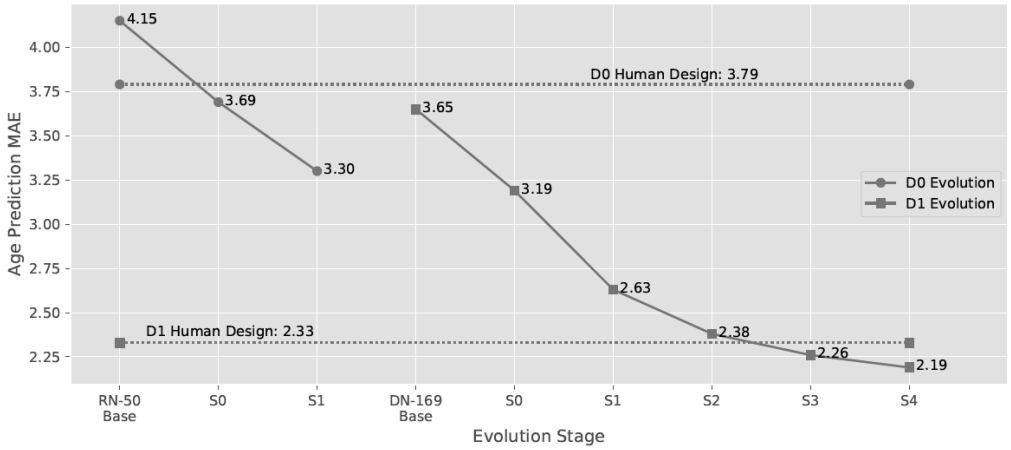


**Figure 1: *Model optimization results.* Performance of the age-estimation model is measured with respect to test set MAE. For both datasets D0 and D1, the Human Design is substantially better than the baseline. Through multiple stages, evolution is able to optimize a model that outperforms the Human Design.**

For D1, the Human Design is based on a much more sophisticated and computationally-expensive base model: EfficientNet-B6 [53], which was state-of-the-art for ImageNet classification at the time of this work [48]. This model takes ≈ 3.25hr to train for one epoch on a K80 GPU, so evolution was incrementally scaled up the full model. Evolution was run for 12 generations using DenseNet-169 as the base model and image size 224 × 224 (S0); it was then scaled to DenseNet-201 for two generations with image size 512 × 512 (S1); the number of epochs per generation was then increased to 50 (S2); next to EfficientNet-B6 for four generations with image size 528 × 528 using PBT (S3); and, finally, the three best models were ensembled (S4). Scaling up the image size in this way is motivated by the observation that larger images generally result in better performance; evolving initially with smaller image sizes makes the approach computationally feasible [53]. Increasing the number of evaluation epochs after a certain stage offers a similar computational advantage [51].

The results for D1 are similar to D0: Evolution is able to improve over the baseline quickly, and eventually achieve a lower error than the Human Design. This final model, the result of the S4 that achieves 2.19 MAE on D1, is the model used in the uncertainty analysis in Section 4.

Note that training age-estimation models is computationally expensive, making extensive comparisons to alternative hyperparameter optimization methods infeasible in this experiment. However, prior comparisons have shown LEAF to perform favorably in similar tasks [32, 36].

## 3.5 Discoveries

The results of evolutionary optimization contain some interpretable discoveries. As a simple example, in S0 on D0 evolution quickly converges on using Vertical Flip but not Horizontal Flip in data augmentation. Horizontal Flip is commonly used in object and face tasks, but the raw images in D0 and D1 happen to all be rotated by 90° degrees; evolution quickly adapts to this data peculiarity. Similarly, in S1 on D0 evolution converges to setting Width Shift Range at around 5X of Height Shift Range, whereas for standard object classification they are usually equal. Again, on this dataset such a setting makes sense: Height Shift (of 90°-rotated images) does not yield much regularization since the opposite side of the face is still visible, whereas Width Shift allows obfuscation of the

forehead or chin, which could otherwise be easy areas for the model to overfit.

The MAE loss was overall preferable to the cross-entropy; this was especially clear during PBT, when the extended training showed that cross-entropy was more prone to early overfitting. However, in the final ensemble, the cross-entropy played a nontrivial role in the three constituent models, which had loss functions defined by $\lambda = 0$, $\lambda = 0.05$, and $\lambda = 0.43$. This loss function diversity (which has been the focus of previous work [31]) highlights the value of using PBT to push the performance of large models.

The highest level blocks of the base image models were surprisingly sufficient for the age estimation problem, given that the problem is seemingly so different from object classification. Although the best models in S1 and S2 on D1 included inputs from lower-level blocks merged with some additional convolutional layers, by the end of S3, the best model used only the input from the output of the final block, following it with a single classification layer. This result suggests that although there is potential for pushing such large models further through evolved architectural innovations, significantly more compute or alternative evolutionary methodologies will be required.

That said, even without drastic architectural modifications, the results demonstrate that large SotA image models can be further optimized to new datasets, and beyond human designs. Notably, the 2.19 MAE achieved by the final evolved model is substantially lower than the recorded MAE of humans trying to guess age from images. In studies that have assessed this human ability, the MAE ranges from 3-4 for highly-controlled image settings [4, 11, 52] to 6-8 for larger datasets in more diverse settings [6, 37, 54]. The high accuracy of the evolved deep model suggests that not only does it allow us to skip expensive human assessment, but it may actually be easier to trust than human assessment. This quantifiable trustworthiness is a property that is exploited in the next section.

# 4 EVALUATING TREATMENT OUTCOMES

This section first describes the RIO uncertainty estimation method and the dataset used in the treatment outcome experiments. A pre-deployment experiment is then presented demonstrating that RIO can estimate uncertainty and improve accuracy in this domain, followed by the main experiment that demonstrates that the treatment is effective.

## 4.1 RIO

RIO [41] is a methodology that can be directly applied on top of any pretrained point-prediction model, i.e., regression model. RIO estimates the predictive uncertainty of the original model quantitatively, and calibrates the original prediction to make it more accurate.

The main idea of RIO is fitting a Gaussian Process (GP) to predict the distribution of residual errors, i.e., the signed difference between the ground-truth and prediction, of the original model. More specifically, the GP is trained to predict

$$r_i = y_i - \hat{y}_i, \quad \text{for } i = 1, 2, \ldots, n\ . \tag{6}$$

where $y_i$ is the label, $\hat{y}_i = \mathcal{F}(\mathbf{X_i})$ is the prediction made by original model, and $r_i$ is the residual error. RIO utilizes a special kernel $k_c$ consisting of both an input kernel $k_{\text{in}}$ and an output kernel $k_{\text{out}}$, and thus takes into account both the input features and the original model outputs when calculating the covariance matrix of the GP:

$$k_c\Big((\mathbf{X}_i, \hat{y}_i), (\mathbf{X}_j, \hat{y}_j)\Big) = k_{\text{in}}(\mathbf{X}_i, \mathbf{X}_j) + k_{\text{out}}(\hat{y}_i, \hat{y}_j), \quad \text{for } i, j = 1, 2, \ldots, n\ . \tag{7}$$

During the training phase, the hyperparameters of the RIO model are optimized by maximizing the log marginal likelihood of sampling the original residual errors on training data, i.e. $\log p(\mathbf{r}|\mathcal{X}, \hat{\mathbf{y}})$.

During the deployment phase, the trained RIO model predicts a distribution of residual error given a new data point $\mathbf{x}_*$ and its corresponding prediction $\hat{y}_*$ by the original regression model, i.e. $\hat{r}_* \sim \mathcal{N}(\bar{\hat{r}}_*, \text{var}(\hat{r}_*))$. By adding the predicted distribution back to the original prediction, a calibrated prediction with its associated uncertainty is obtained:

$$\hat{y}'_* \sim \mathcal{N}\Big(\hat{y}_* + \bar{\hat{r}}_*, \text{var}(\hat{r}_*)\Big). \tag{8}$$

## 4.2 Dataset

In order to evaluate the effectiveness of treatment, a clinical dataset D2 was prepared based on a Botox study, containing 3925 images that were taken before treatment and 68,799 images taken after treatment, at one and two weeks and monthly over six months. Two different treatment versions with different injection volumes were included. In addition, the dataset contained 5190 after-treatment images taken at similar time points, where, instead of an actual treatment, a placebo injection was given to the patient. In total there were 787 patients aging from 21 to 76, of which 156 patients were in the placebo group. Table 2 compares D2 to the datasets used in Section 3.

Note that some patients may look younger/older than their true age. In order to remove such inherent biases, the prediction errors of pre-treatment images were calculated for each patient using the final (S4) model from Section 3. The data included several pre-treatment images (with different poses) for each patient, and the average prediction error was used as the inherent age bias for each patient. These patient-wise biases were then deducted from all predictions made by the S4 model on pre-treatment, placebo, and post-treatment images.

## 4.3 Pre-deployment Evaluation of RIO

In order to verify that RIO can estimate uncertainty and improve accuracy in the age-prediction domain, a preliminary RIO model was trained to predict the residual errors of the S4 model on placebo images of dataset D2.

Of the placebo images, 80% were used as training data, and the remaining 20% as testing data. The input kernel of the RIO model takes the softmax probability vector of S4 model, and the output kernel takes the expected age of S4 model. As in the original RIO work [41], radial basis function (RBF) was used for I/O kernel, the number of inducing points was 50, L-BFGS-B optimizer was used with maximum optimization iteration of 1000. MAE was used to evaluate the prediction accuracy before and after applying RIO. To measure the quality of uncertainty estimation, coverage percentages of testing points by 95%/90%/68% confidence intervals (i.e., the

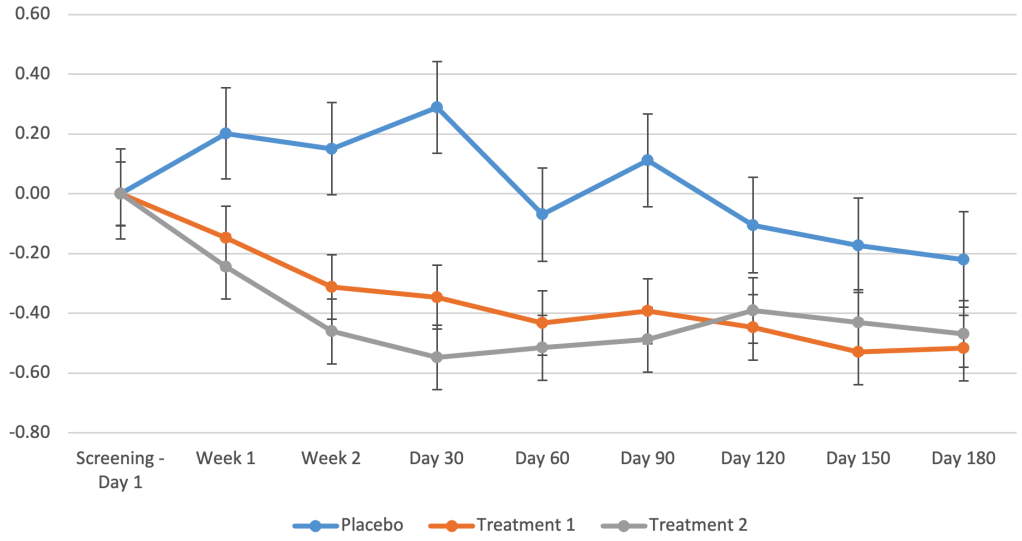


**Figure 2: *Comparing Treatment Effects with Placebo Effects.* The vertical axis shows the perceived age difference from pre-treatment images to images taken at different times after treatment. The error bars indicate standard error on RIO values, averaged across individuals. Whereas the estimated age differences with placebo treatment is centered around zero, the actual Botox treatments (of which there were two versions) reduce the apparent age substantially, demonstrating that the treatments are effective.**

| Metric | Value |
|---|---|
| Original MAE | 1.61 |
| MAE with RIO | 1.48 |
| 95% CI Coverage | 94.2% |
| 90% CI Coverage | 89.2% |
| 68% CI Coverage | 69.2% |

**Table 3: *Pre-deployment Evaluation of RIO.* CI coverage means the percentage of testing outcomes that are within the estimated CI. RIO reduces the prediction error of the S4 model and provides accurate uncertainty estimation for its prediction.**

percentage of testing outcomes that are within the corresponding confidence intervals estimated by RIO) were calculated.

The results are shown in Table 3. RIO provides reliable uncertainty estimation with accurate confidence intervals, and by adding the calibration, improves the prediction accuracy of the S4 model.

## 4.4 Measuring Age Reduction

After the pre-deployment study showed that RIO can be deployed reliably in the age estimation domain, it was retrained with the full set of 5190 placebo images, in order to obtain the most accurate uncertainty model with the available data. The placebo images are representative of those who may want aesthetics treatments, thereby forming a relevant training set. The pre-treatment images were excluded from the RIO training since they have already been used to remove the inherent age biases of patients. The S4 age prediction model was again used to generate the age predictions during training. S4 and RIO were then applied to all post-treatment images of both treatments. Figure 2 illustrates the results, comparing them to those of the placebo group in the pre-deployment study.

The age estimations of the placebo images vary somewhat but are centered around zero. In contrast, the age estimations of both series of treatment images are substantially lower. They decrease rapidly during the first 1-2 months as the treatment takes hold and then stabilize, as expected. Eventually they reach a final age reduction of about 0.5 years. Since the patients actually aged 0.5 years during this time, their appearance in the end is thus estimated to be one full year younger. (Note that this reduction is due to a single injection; a typical treatment consists of multiple injections, with a larger cumulative effect.) Treatment1 works slightly faster than Treatment2, but the difference is insignificant towards the end.

The results thus demonstrate that these medical aesthetics treatments are effective, reducing the perceived age substantially. While such a reduction was previously perceived only subjectively, the AI approach makes it objective and quantitative.

## 5 DISCUSSION AND FUTURE WORK

The experiments in this paper quantify an important outcome of a medical aesthetics treatment: The patient's age appears to be substantially reduced as a result. Up to this point, this outcome has been measured subjectively—the AI age estimation and uncertainty measurement techniques in this paper make it possible to quantify it objectively. Building on a medically relevant dataset, evolutionary optimization of network design plays a crucial role in this process, customizing the neural networks that were originally developed for other computer vision applications to this task. The resulting networks have improved accuracy over the original designs as

well as over designs optimized by human experts. In turn, the RIO uncertainty measurement mechanism makes it possible to demonstrate that the age reduction result is reliable.

The approach can be extended in the future to other types of medical aesthetics treatments, as well as sequences of treatments that consist of multiple injections, and combinations of treatments of different types. Any treatment for which image and age data is available can be evaluated in the same manner. Similar models can be developed to estimate other aspects of the outcome, such as whether the facial features and facial movements appear natural. The ground truth may be established based on human judgment, or through a discriminator (similar to that in a GAN [14]), trained to distinguish pre- and post-treatment faces.

Such a quantitative analysis opens intriguing possibilities for expanding the role of AI in medical aesthetics. Using the same data, models can be trained to predict the effects of treatments, i.e. generate face images that are likely to result from them. Age estimation can be applied to the resulting images, predicting their success quantitatively. Such models can be further conditioned with the treatment type, making it possible to compare alternatives. Further, using such a model as a surrogate, it may be possible to evolve another model to make treatment recommendations that optimize the outcomes. They can be multi-objective, resulting in a number of choices that balance the different quality metrics as well as cost and side-effects. Such a tool could be highly valuable to physicians and patients in evaluating treatment alternatives.

## 6 CONCLUSION

In domains like aesthetic medicine, it is difficult to evaluate the treatment outcomes in an objective and quantitative manner. This paper demonstrates how AI can be harnessed in this role. The design of a deep learning neural network is customized through evolution to estimate the patient age from facial images, and RIO is used to estimate the uncertainty in the estimate. For the first time, this study was able to demonstrate quantitatively that aesthetic treatments can potentially reduce the perceived chronological age of patients. The approach can serve as a foundation for a number of future extensions in evaluation, prediction and optimization of medical aesthetics treatments, thus empowering physicians and patients to reach more ambitious treatment objectives.

## REFERENCES


[1] Abelsson, A. and Willman, A. 2020. Ethics and aesthetics in injection treatments with Botox and Filler. *Journal of Women & Aging* (2020), 1–13.

[2] Arsiwala, S. Z. 2018. Trends for Facial Injectable Therapies in Medical Aesthetics. *Journal of Cutaneous and Aesthetic Surgery* 11 (2018), 45–46.

[3] Bingham, G., Macke, W., and Miikkulainen, R. 2020. Evolutionary Optimization of Deep Learning Activation Functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*.

[4] Burt, D. M. and Perrett, D. I. 1995. Perception of age in adult Caucasian male faces: computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 259, 1355 (1995), 137–143.

[5] Chollet, F. and others, . 2015. Keras. https://keras.io. (2015).

[6] Clifford, C. W., Watson, T. L., and White, D. 2018. Two sources of bias explain errors in facial age estimation. *Royal Society Open Science* 5, 10 (2018), 180841.

[7] DeVries, T. and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552* (2017).

[8] Elsken, T., Metzen, J. H., Hutter, F., and others, . 2019. Neural architecture search: A survey. *J. Mach. Learn. Res.* 20, 55 (2019), 1–21.

[9] Gal, Y. and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*. 1050—-1059.

[10] Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. M. A. 2018. Conditional Neural Processes. In *Proceedings of the 35th International Conference on Machine Learning*. 1704–1713.

[11] George, P. A. and Hole, G. J. 2000. The role of spatial and surface cues in the age-processing of unfamiliar faces. *Visual Cognition* 7, 4 (2000), 485–509.

[12] Gonzalez, S. and Miikkulainen, R. 2020. Improved Training Speed, Accuracy, and Data Utilization Through Loss Function Optimization. In *Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC)*.

[13] Gonzalez, S. and Miikkulainen, R. 2020. Optimizing Loss Functions Through Multivariate Taylor Polynomial Parameterization. *arXiv:2002.00059* (2020).

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2672–2680.

[15] Grand View Research, . 2020. Aesthetic Medicine Market Size, Share & Trends Analysis Report By Procedure Type (Invasive Procedures, Non-invasive Procedures), By Region (North America, Europe, Asia Pacific, Latin America, Middle East & Africa), And Segment Forecasts, 2021 - 2028. (2020). https://www.grandviewresearch.com/industry-analysis/medical-aesthetics-market, accessed 2/4/2021.

[16] Grifantini, K. 2020. Detecting Faces, Saving Lives: How facial recognition software is changing health care. *IEEE Pulse* (March/April 2020).

[17] He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Hensman, J., Matthews, A., and Ghahramani, Z. 2015. Scalable Variational Gaussian Process Classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. 38. 351–360.

[19] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861* (2017).

[20] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.

[21] Izmailov, P., Wilson, A., Podoprikhin, D., Vetrov, D., and Garipov, T. 2018. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. 876–885.

[22] Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., and others, . 2017. Population based training of neural networks. *arXiv:1711.09846* (2017).

[23] Karras, T., Laine, S., and Aila, T. 2020. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* (2020). 10.1109/TPAMI.2020.2970919.

[24] Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O., and Teh, Y. W. 2019. Attentive Neural Processes. In *Proceedings of the International Conference on Learning Representations*.

[25] Klassen, A. F., Cano, S. J., Scott, A., Snell, L., and Pusic, A. L. 2010. Measuring patient-reported outcomes in facial aesthetic patients: Development of the FACE-Q. *Facial Plastic Surgery* 26, 4 (2010), 303.

[26] Kornblith, S., Shlens, J., and Le, Q. V. 2019. Do better imagenet models transfer better?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2661–2671.

[27] Kosowski, T. R., McCarthy, C., Reavey, P. L., Scott, A. M., Wilkins, E. G., Cano, S. J., Klassen, A. F. D., Carr, N., Cordeiro, P. G., and Pusic, A. L. 2009. A Systematic Review of Patient-Reported Outcome Measures after Facial Cosmetic Surgery and/or Nonsurgical Facial Rejuvenation. *Plastic and Reconstructive Surgery* 123, 6 (2009), 1819–1827.

[28] Lakshminarayanan, B., Pritzel, A., and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 6405–6416.

[29] Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., and Sohldickstein, J. 2018. Deep neural networks as Gaussian processes. In *Proceedings of the International Conference on Learning Representations*.

[30] Li, S. and Deng, W. 2020. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* (2020).

[31] Liang, J., Gonzalez, S., Shahrzad, H., and Miikkulainen, R. 2021. Regularized Evolutionary Population-based Training. In *Proceedings of the Genetic and Evolutionary Computation Conference*.

[32] Liang, J., Meyerson, E., Hodjat, B., Fink, D., Mutch, K., and Miikkulainen, R. 2019. Evolutionary Neural AutoML for Deep Learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 401–409.

[33] Ma, D., You, F., Gong, Y., Tu, H., Liang, J., and Wang, H. 2020. A Fatigue Driving Detection Algorithm Based on Facial Motion Information Entropy. *Journal of Advanced Transportation* 8851485 (2020).

[34] Masi, I., Wu, Y., Hassner, T., and Natarajan, P. 2018. Deep Face Recognition: A Survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 471–478.

[35] Meyerson, E. and Miikkulainen, R. 2018. Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back. In *International Conference on Machine Learning*. 3511–3520.

[36] Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., and Hodjat, B. 2020. Evolving Deep Neural Networks. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, C. F. Morabito, C. Alippi, Y. Choe, and R. Kozma (Eds.). Elsevier, New York.

[37] Moyse, E. and Brédart, S. 2012. An own-age bias in age estimation of faces. *European Review of Applied Psychology* 62, 1 (2012), 3–7.

[38] Mutch, K. 2021. Studio Go Runner. (2021). https://github.com/leaf-ai/studio-go-runner/tree/0.13.1

[39] Naqvi, R. A., Arsalan, M., Rehman, A., Rehman, A. U., Loh, W.-K., and Paul, A. 2020. Deep Learning-Based Drivers Emotion Classification System in Time Series Data for Remote Applications. *Remote Sensing* 12, 3 (2020).

[40] Neal, R. M. 1996. *Bayesian Learning for Neural Networks*. Springer, Berlin.

[41] Qiu, X., Meyerson, E., and Miikkulainen, R. 2020. Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel. In *Proceedings of the International Conference on Learning Representations*.

[42] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., and others, . 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv:1711.05225* (2017).

[43] Real, E., Aggarwal, A., Huang, Y., and Le, Q. V. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4780–4789.

[44] Real, E., Liang, C., So, D., and Le, Q. 2020. AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. 8007–8019.

[45] Ren, S., He, K., Girshick, R., and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv:1506.01497* (2015).

[46] Rezende, E., Ruppert, G., Carvalho, T., Ramos, F., and De Geus, P. 2017. Malicious software classification using transfer learning of ResNet-50 deep neural network. In *Sixteenth IEEE International Conference on Machine Learning and Applications*. IEEE, 1011–1014.

[47] Rothe, R., Timofte, R., and Van Gool, L. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2 (2018), 144–157.

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., and others, . 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[49] Sato, A., Takaki, S., Yokose, M., and Goto, T. 2019. Automatic prediction model with facial recognition and machine learning of patient's image in ICU patients. *European Journal of Anesthesiology* 36 e-Suppement 57 (2019).

[50] Simonyan, K. and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

[51] So, D., Le, Q., and Liang, C. 2019. The evolved transformer. In *International Conference on Machine Learning*. 5877–5886.

[52] Sörqvist, P. and Eriksson, M. 2007. Effects of training on age estimation. *Applied Cognitive Psychology* 21, 1 (2007), 131–135.

[53] Tan, M. and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. 6105–6114.

[54] Voelkle, M. C., Ebner, N. C., Lindenberger, U., and Riediger, M. 2012. Let me guess how old you are: Effects of age, gender, and facial expression on perceptions of age. *Psychology and aging* 27, 2 (2012), 265.

[55] Wang, M. and Deng, W. 2018. Deep Face Recognition: A Survey. *arXiv:1804.06655* (2018).

[56] Yang, T.-Y., Huang, Y.-H., Lin, Y.-Y., Hsiu, P.-C., and Chuang, Y.-Y. 2018. SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation.. In *International Joint Conference on Artificial Intelligence*.

[57] Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T. C., and Qu, H. 2020. EmotionCues: Emotion-Oriented Visual Summarization of Classroom Videos. *IEEE Transactions on Visualization and Computer Graphics* (2020).

[58] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8697–8710.