



## Research Paper

## Automatic Detection of Acromegaly From Facial Photographs Using Machine Learning Methods

Xiangyi Kong<sup>a,b,1</sup>, Shun Gong<sup>c,1</sup>, Lijuan Su<sup>d,e,1</sup>, Newton Howard<sup>f,g</sup>, Yanguo Kong<sup>a,\*</sup><sup>a</sup> Department of Neurosurgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, No. 1 Shuafuyuan Hutong, Dongcheng District, Beijing 100730, China<sup>b</sup> Department of Breast Surgical Oncology, China National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Chaoyangqu, Panjiayuan-Nanli 17, Beijing 100021, PR China<sup>c</sup> Department of Neurosurgery, Shanghai Institute of Neurosurgery, PLA Institute of Neurosurgery, Changzheng Hospital, Second Military Medical University, 415 Fengyang Road, Shanghai 200003, China<sup>d</sup> College of Computer Science and Technology, Zhejiang University, No. 38 Zheda Road, Hangzhou, Zhejiang 310027, China<sup>e</sup> Healthcare big data lab, Tencent Technology (Shenzhen) Company Limited, Kejizhongyi Avenue, Hi-tech Park, Nanshan District, Shenzhen, 518057, China<sup>f</sup> Synthetic Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, United States<sup>g</sup> Computational Neuroscience Laboratory, Oxford University, Oxford OX1 3QD, UK

## ARTICLE INFO

## Article history:

Received 12 May 2017

Received in revised form 6 December 2017

Accepted 14 December 2017

Available online 15 December 2017

## Keywords:

Automatic acromegaly diagnosis

Artificial intelligence

Machine learning

Face recognition

Convolutional neural network

## ABSTRACT

**Background:** Automatic early detection of acromegaly is theoretically possible from facial photographs, which can lessen the prevalence and increase the cure probability.**Methods:** In this study, several popular machine learning algorithms were used to train a retrospective development dataset consisting of 527 acromegaly patients and 596 normal subjects. We firstly used OpenCV to detect the face bounding rectangle box, and then cropped and resized it to the same pixel dimensions. From the detected faces, locations of facial landmarks which were the potential clinical indicators were extracted. Frontalization was then adopted to synthesize frontal facing views to improve the performance. Several popular machine learning methods including LM, KNN, SVM, RT, CNN, and EM were used to automatically identify acromegaly from the detected facial photographs, extracted facial landmarks, and synthesized frontal faces. The trained models were evaluated using a separate dataset, of which half were diagnosed as acromegaly by growth hormone suppression test.**Results:** The best result of our proposed methods showed a PPV of 96%, a NPV of 95%, a sensitivity of 96% and a specificity of 96%.**Conclusions:** Artificial intelligence can automatically early detect acromegaly with a high sensitivity and specificity.© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Acromegaly is caused by persistent excessive secretion of growth hormone (GH), usually resulting from somatotroph adenomas. It once has been regarded as one of the rare diseases, whose estimated prevalence is around 0.07% in Europe. More recently a higher prevalence of about 130 per million has been suggested by a study in Belgium with more active surveillance for pituitary adenomas (Ribeiro-Oliveira Jr and Barkan, 2012; Melmed and Acromegaly, 1990; Chanson et al., 1989; Reddy et al., 2010). Nevertheless, current prevalence is significantly higher (1‰ in population) (Schneider et al., 2008). As acromegaly course is insidious, often resulting in a delay of 7–10 years in diagnosis after the suspected symptoms and signs onset, few patients seek care due to their appearances' changing (Utiger, 2000; Melmed,

2006). When diagnosed, 75%–80% patients have macroadenomas with some even presenting invasive growth pattern, such as a higher Ki-67 index, expansion and invasiveness (Katznelson et al., 2014). Uncontrolled GH/IGF-1 excess results in increased prevalence and mortality, whereas these could be prevented with the timely and successful disease control, thus early detection of acromegaly is crucial. The variety of disease manifestations might lead patients to refer to the doctors of different specialties (dentist, hand surgeon, ophthalmologist, gynecologist, etc.) that can result in even longer delay of the diagnosis of acromegaly.

Recently, artificial intelligence (AI) increasingly shines brilliantly through the medical area (Hamet and Tremblay, 2017; Obermeyer and Emanuel, 2016; Gencturk et al., 2013). Several peer-review/high-impact-factor journals have published studies using AI helping with assistant diagnosis (Gulshan et al., 2016; Esteva et al., 2017; Kanakasabapathy et al., 2017). Since facial-features of acromegaly are very typical: widening teeth spacing, prognathism, frontal-bone enlargement, nose enlargement, zygomatic-arch prominence, brow ridge and forehead protrusion/prominence, soft tissue swelling (lips, nose,

\* Corresponding author.

E-mail address: [kongyg@pumch.cn](mailto:kongyg@pumch.cn) (Y. Kong).<sup>1</sup> These 3 authors contribute equally to this article.

ears enlargement) and skin thickening, we herein developed a computational, automatic and handy face-recognition system which may allow doctors or patients to proactively track facial-features changes and detect acromegaly. Using 1123 face photographs, we trained and integrated several machine learning methods including Generalized Linear Models (LM), K-nearest neighbors (KNN), Support Vector Machines (SVM), Forests of randomized trees (RT) and Convolutional Neural Network (CNN) to create an Ensemble Method (EM) for facial detection of acromegaly. Other than the typical facial changes, symptoms and signs of acromegaly also include stimulation of growth of many tissues, such as skin, connective tissue, cartilage, bone, viscera, and many epithelial tissues. The metabolic effects include nitrogen retention, insulin antagonism, and lipolysis. The variety of disease manifestations of acromegaly might lead patients to refer to the doctors of different specialties (dentist, hand surgeon, ophthalmologist, gynecologist, etc.) that can result in even longer delay of the diagnosis of acromegaly. On the other hand, these doctors probably could benefit from algorithms of automatic detection of acromegaly developed by us.

## 2. Results

### 2.1. Ensemble Method of LR, KNN, SVM, RT and CNN

The goal of ensemble method was to combine the outcomes of some weak estimators in order to achieve higher classification accuracy as well as better generalizability and robustness. The three most popular methods for combining the predictions from different models were Bagging, Boosting and Voting. Herein, Bagging was used to decrease the variance and thus improve generalization. As shown in Fig. 1, to

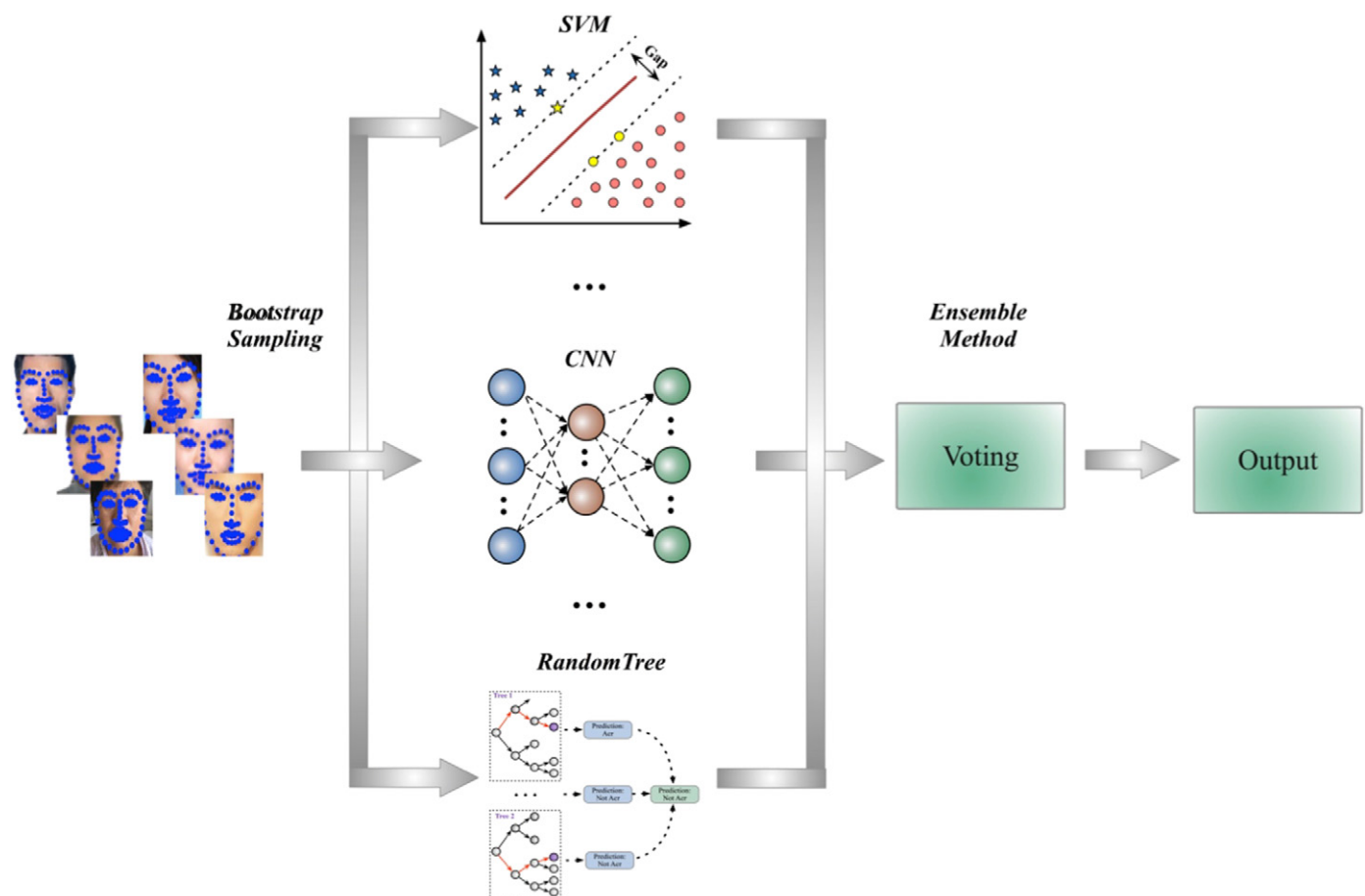
aggregate the outputs from LR, KNN, SVM, RT and CNN, the strategy of weighted arithmetic mean was used where the corresponding weights were computed by linear least squares.

### 2.2. Evaluation Metrics

With a separate dataset, consisting of acromegaly and normal facial photographs, we used several classical metrics to quantify the quality of acromegaly detection: sensitivity and specificity, positive predictive value (PPV) and negative predictive value (NPV), and f1-score, where sensitivity was defined as true positive/condition positive, specificity was defined as true negative/condition negative, PPV was defined as true positive/prediction positive and NPV was defined as true negative/prediction negative. The F1 score can be interpreted as a weighted average of the precision and recall, where the formula for the F1 score is:  $F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ .

### 2.3. Facial Landmarks

General framework of our method to automatically detect acromegaly from facial photographs using machine learning methods are shown in Fig. 3: the most left part showed the training processing, which included face detection and normalization from original input images, facial feature extraction and landmark localization, and face frontalization to improve the accuracy of acromegaly diagnosis; the middle left part showed the model training processing, which included Generalized Linear Models (LM), K-nearest neighbors (KNN), Support Vector Machines (SVM), Forests of randomized trees (RT), Convolutional Neural Network (CNN), and Ensemble Method (EM); the middle right part showed the



**Fig. 1.** Ensemble method used to combine the predictions of the basic estimators (LM, KNN, SVM, RT, CNN) to achieve higher classification accuracy as well as better generalizability and robustness for acromegaly diagnosis.

testing processing with the trained models; the most right part showed the evaluation between the ground truth, computers and doctors.

#### 2.4. Performance Evaluation

To evaluate the diagnosis performance, we compared the diagnosis results between 9 board-certified endocrinologists or neurosurgeons specialized in pituitary disease/familial with acromegaly (only through photographs) and machine learning methods in terms of the confusion matrix, sensitivity and specificity, PPV and NPV, and f1-score.

As shown in Tables 2 and 3: with the facial landmarks locations, SVM worked best with a PPV of 95%, a NPV of 88%, a sensitivity of 86% (with true positive of 98) and a specificity of 96% (with true negative of 123); with the detected facial photographs, CNN was the best, and after the face frontalization the performance was further improved with a PPV of 96%, a NPV of 92%, a sensitivity of 91% (with true positive of 109) and a specificity of 96% (with true negative of 123); the average PPV of 90%, NPV of 77%, sensitivity of 73% (with true positive of 83) and specificity of 92% (with true negative of 118) for doctors were lower than machine learning methods. The results for primary care doctors were even lower (PPV 83%, NPV 70%, sensitivity 68% (with true positive of 78), specificity 85% (with true negative of 109); the F1-score for control was 0.77, the F1-score for acromegaly patients was 0.75). The accuracy of CNN was much higher than SVM trained with hand-crafted features indicating that deep neural networks could yield conceptual abstractions in a hierarchical way, thus showing promising ability to extract underlying features automatically.

### 3. Discussion

Acromegaly is a group of disorders where excessive GH, mostly from GH-secreting pituitary adenomas causing excessive tissue-growth lasting for lots of years. The clinical manifestations of acromegaly range from subtle signs of acral overgrowth, soft-tissue swelling, arthralgias, fasting hyperglycemia, and hyperhidrosis to florid osteoarthritis, frontal bone bossing, diabetes mellitus, hypertension, and respiratory and cardiac failure. Typically, the facial-tissues are excessive with deep nasolabial-folds, the nose or the brow is enlarged, the jaw is prominent, and the zygomatic arch is prominent. Because the facial change takes place very occultly, it is usually ignored, especially by the patient himself or herself and persons who frequently see the patient. Early diagnosis of acromegaly would facilitate cures and decrease the treatment cost, hence bringing great cost-effectiveness. Based on previous work, our study is aimed at using facial pictures to screen acromegaly in adults. Preferably, acromegaly face-recognition could be applied to screen in clinical settings. One can voluntarily choose to be screened automatically in a form of self-service under a variety of conditions. When positive results were suggested after the algorithm system analyses, he or she would be recommended to see a doctor for further specialized tests and professional diagnoses, like enhanced magnetic resonance imaging (MRI) sellar-region scan and routine examinations of pituitary-hormone series.

The technologies of face-recognition have been developed for many years, generally focusing on alterations of facial features due to aging, tilt, face poses, and lighting, etc. Recently, Kruszka et al. published their study using face-recognition technology to assist doctors to diagnose 22q11.2 deletion syndrome with a specificity and a sensitivity higher than 96% (Kruszka et al., 2017). Miller et al. developed a computer program which would separate 24 facial photographs of patients with acromegaly, from those of 25 normal subjects (Miller et al., 2011; Learned-Miller et al., 2006), which obtained an accuracy of 86% (Miller et al., 2011). Using computerized similarity, Schneider et al. analyzed face photographs of 57 acromegaly patients and 60 controls (Schneider et al., 2011), with an accuracy of 91.5% for controls and 71.9% for patients (Schneider et al., 2011). However, there are some limitations for these studies: 1) Facial photographs were taken with certain

model cameras against certain backgrounds; 2) The sample size was too small (Miller et al., 2011; Schneider et al., 2011); 3) the study was performed in a specific population and cannot be extrapolated to other populations. Especially in black people, the diagnostic is often more difficult. In our next work, we will incorporate more data of patients and controls with both Caucasian and Asian background; 4) Only SVM was used, instead of integrating multiple algorithms like CNN.

Aiming at training computers for detecting certain patterns, machine learning has become ubiquitous and indispensable for solving complex problems and is opening up vast new possibilities in various medical areas. Unlike traditional machine learning primarily focusing on “feature-engineering”, deep learning can learn conceptual abstractions in a hierarchical way and discover underlying characteristics ignored by humans or unknown previously. CNN, as a typical kind of feed-forward artificial neural network has been widely applied in video, audio, and image recognition. Recently, using CNN trained on general skin lesion images, Esteva et al. matched the performances of 21 dermatologists tested across diagnostic tasks of melanoma and keratinocyte carcinoma classification (Esteva et al., 2017).

To get the optimal algorithm detecting acromegaly automatically, as shown in Fig. 1, advantages and strengths of several machine learning methods (LM, KNN, SVM, RT and CNN) were integrated. The predictions of these basic estimators were combined to achieve higher classification accuracy as well as better generalizability and robustness. We used Bagging to decrease the variance and thus improved the generalization. Our results demonstrated that the developed ensemble method could be adopted without having to specify facial soft tissue-based or facial bone-based features, with a high specificity and sensitivity. In our study, a PPV of 96%, a NPV of 95%, a sensitivity of 96% and a specificity of 96% were achieved. This is the first time that various machine learning methods were united to generate a better integrated one.

But there were still some limitations to the proposed methods: 1) We did not include photographs of side views; 2) A potential for sample selection bias might exist. 3) The data size was still not big enough when compared with Esteva's work (129,450 images) (Esteva et al., 2017) and Gulshan's work (128,175 retinal images) (Gulshan et al., 2016); 4) Stratification analyses and subgroup analyses (disease severity, disease course, gender, age, and ethnicity, etc.) were not performed but will be included in the future; 5) The study was performed on a specific population, and the controls were from the same ethnicity. So, it is unknown if the results could be extrapolated to other population.

Of note, we didn't evaluate whether patients in their study were with mild or advanced acromegaly. Usually, in advanced acromegaly cases the diagnosis could be a lot less challenging than in cases with a mild expression of the disease. In the next step of our work, we have collected hundreds of groups of facial pictures of the same patient in different disease severity stages [scored from 0 (normal stage) to 5 (most severe)] and will train the computer to automatically provide the severity evaluation score based on a certain facial picture. In this process, the machines performance in mild acromegaly will be deeply studied. Moreover, there are several other signs of acromegaly and it could be evoked to integrate them to build a more complete algorithm (i.e.: acral changes, swelling, arthralgia, diabetes, hypertension, sleep apneas) in our next work. In the present study, despite the stratification analysis not being conducted, the probability of acromegaly for the same person's facial photographs in different stages, shown in the Fig. 2, indicated that machine learning methods could not only identify acromegaly with a high accuracy, but also could distinguish faces in different periods with a significant probability estimation thus showing a potential to predict the disease severity. Studies with more patients' facial photographs in different disease-course stages (some patients' diseases were serious and facial catachrestic were more typical; while some others were wild) will be further explored.

In conclusion, diagnosis of acromegaly faces indeed significant delays, thus new methods of assuring an early diagnosis are necessary and important. Our data demonstrated the efficiency of machine



**Fig. 2.** Prediction for disease-course stages using machine learning methods: the probability estimation showed that machine learning methods provided a potential way to predict the disease severity of acromegaly. These pictures were cited from <https://www.umhs-community.org/medterm-corner-acromegaly/>; <https://www.bbsmax.com/A/kmzL0xvWzG/>; <http://www.healthgoesup.com/articles/1065/1/acromegaly-symptoms-1.html>; and <http://www.merckmanuals.com/professional/endocrine-and-metabolic-disorders/pituitary-disorders/gigantism-and-acromegaly>.

learning methods in acromegaly face detection, which outperformed neuroendocrinologists specialized in pituitary disease/familial with acromegaly. To integrate these technologies in a clinical area, this fast, scalable method is deployable on mobile devices and holds the potential for substantial clinical impact, including broadening the scope of primary care practice and other specialists, and augmenting clinical decision-making for neuroendocrinologists in real-world and clinical settings. With the wide-spread use of smart-phones this could allow the development of an application that clinicians who do not frequently diagnosis acromegaly can use to guide clinical decision making for further testing and specialist referrals. Also, if all people can download the future mobile apps with the core technology focusing on automatic acromegaly detection from facial photographs, possible and potential acromegaly patients could be reminded about the disease possibility with a probability, which is a very convenient process (only need a mobile phone autodyne or uploading a facial picture); then the possible “patients” can choose to go to the hospital to see a doctor. If the doctor highly suspects the disease, a further series of tests including biochemical assessment of GH and IGF-1 are required. In these processes, the unlocked photo information and privacy protection might be a pair of natural contradiction; thus, no matter at which phase of artificial intelligence applications and algorithm developments, privacy protection with ethical design thinking shall be always kept in mind. A series of actions and rules designed to protect privacy are deserved to be made.

## 4. Materials and Methods

### 4.1. Data Sources

The study conforms to the Declaration of Helsinki. The acromegaly face photographs were mostly from neurosurgery inpatient departments of several large general hospitals in China, and partially from previously published studies. The number used in the training and testing dataset was shown in Table 1. Totally, 527 acromegaly patients (254 women,  $54.3 \pm 12.7$  years; 273 men,  $52.7 \pm 10.1$  years) were included. The acromegaly had all been proven by growth hormone suppression

test used in clinical guidelines (Laurence et al., 2014). Additionally, elevated IGF-1 level was checked, and in those patients with biochemical acromegaly pituitary imaging was performed. Controls were recruited partly from volunteers at these hospitals and partly downloaded from the SCUT-FBP dataset (Xie et al., 2015). Controls, matched by sex and age, did not have GH excess or deficiency, hyper- or hypothyroidism, Cushing's syndrome, glucocorticoid use, androgen excess and menopause in women, obesity, pseudoacromegaly or consuming diseases (e.g. tuberculosis, malignancy, and cachexy, etc.). Androgen excess and menopause in women, also obesity and pseudoacromegaly in both genders, could change the physical appearance. These kinds of patients were also excluded. The study was performed on a specific population, and also that the controls were from the same ethnicity. We included 596 controls (287 women,  $39.88 \pm 10.6$  years; 309 men,  $40.33 \pm 11.4$  years). One separate dataset consisting of acromegaly and normal facial photographs, which had no overlap with the training dataset, was used for clinical validation. The test dataset contains age- and sex-matched 114 acromegaly patients and 128 controls. All participants, including the participants supporting the facial photographs, gave separate written informed consent for publication of this article, the copies of which are all available at any time for review.

### 4.2. Photographs

Shooting angle was allowed within plus or minus 45 degrees centered on the standard coronal plane. The backgrounds and cameras were not limited.

**Table 1**  
Number in training and testing dataset.

	Training	Testing
Normal	596	128
Acromegaly	527	114



**Table 2**  
Confusion matrix to evaluate accuracy of different methods.

Prediction used different methods						
	Detected faces		Facial features		Frontalized faces	
Normal	Normal	Acromegaly	Normal	Acromegaly	Normal	Acromegaly
Acromegaly	Acromegaly	Normal	Acromegaly	Normal	Acromegaly	Normal
LM	102	26	64	64	102	26
	86	28	103	11	100	14
KNN	102	26	113	15	119	9
	97	17	103	11	101	13
SVM	108	20	123	5	115	13
	91	23	98	16	88	26
RT	102	26	101	27	111	17
	97	17	98	16	103	11
CNN	111	17	–	–	123	5
	104	10	–	–	104	10
Ensemble MLs	–	–	–	–	123	5
	–	–	–	–	109	5
Specialists in pituitary disease (average) <sup>a</sup>	–	–	–	–	118	10
	–	–	–	–	83	31
Primary care doctors (average) <sup>a</sup>	–	–	–	–	109	19
	–	–	–	–	78	36

Confusion matrix in this table shows the details about the predictions and observations, which can evaluate the accuracy of classification based on different methods.

<sup>a</sup> The values for doctors are calculated based on doctors' diagnosis records on original facial images.

#### 4.3. Method Outline

To automatically detect acromegaly from facial photographs, we firstly detected, cropped and resized the original photographs, and then extracted the key facial landmarks. Lastly, we used frontalization to improve the diagnostic performance. Based on the detected facial photographs, extracted facial landmarks locations, as well as the synthesized frontal faces, several machine learning methods were used to automatically identify acromegaly. General framework of our method to automatically detect acromegaly from facial photographs using machine learning methods was shown in Fig. 3: the most left part showed the training processing, which included face detection and normalization from original input images, facial feature extraction and landmark localization, and face frontalization to improve the accuracy of acromegaly diagnosis; the middle left part showed the model training processing, which included Generalized Linear Models (LM), K-nearest neighbors (KNN), Support Vector Machines (SVM), Forests of randomized trees (RT), Convolutional Neural Network (CNN), and Ensemble Method (EM); the middle right part showed the testing processing with the trained models; the most right part showed the evaluation

between the ground truth, computers and doctors. Flowchart of our proposed facial photograph diagnosis system is shown in Fig. 4: the upper figures showed processing of the facial photograph: the first step is to detect faces from the original images, and then extract and localize the features, and the last step is to frontalize faces to improve the accuracy of acromegaly diagnosis; the below figures showed the overview of the inputs to machine learning methods.

#### 4.4. Image Pre-Processing

##### 4.4.1. Face Detection and Normalization

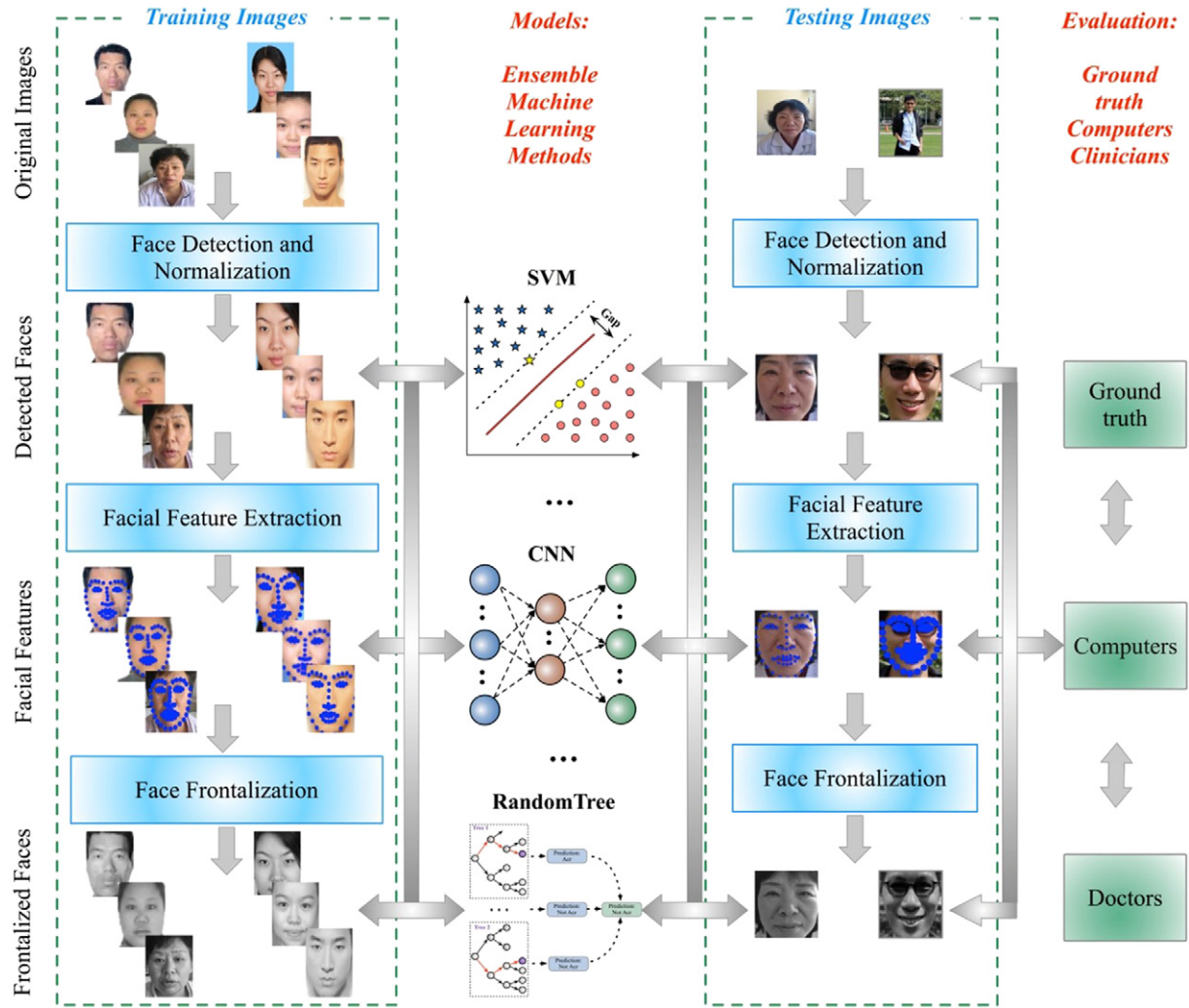
To detect acromegaly using automatic face-recognition technology, we firstly needed to delineate where a face was located in the photo. Therefore, face detection locating a face and getting the bounding rectangle square was the first step. Following the first semi-automated facial recognition programs developed in 1960, the rapid object detection technology using a boosted cascade of simple features was proposed in 2001 (Tikoo and Malik, 2016), which has been proved effective to obtain a high performance given enough labeled training

**Table 3**  
Specificity and sensitivity to evaluate accuracy of different methods.

	Detected faces			Facial features			Frontalized faces		
	NPV	Specificity	F1-score	NPV	Specificity	F1-score	NPV	Specificity	F1-score
Normal	PPV	Sensitivity	F1-score	PPV	Sensitivity	F1-score	PPV	Sensitivity	F1-score
Acromegaly	0.80	0.80	0.80	0.86	0.50	0.63	0.84	0.80	0.82
LM	0.75	0.75	0.75	0.61	0.90	0.73	0.85	0.88	0.86
KNN	0.87	0.80	0.83	0.91	0.88	0.89	0.93	0.93	0.93
	0.77	0.85	0.81	0.86	0.90	0.88	0.89	0.89	0.89
SVM	0.84	0.84	0.84	0.88	0.96	0.92	0.75	0.90	0.82
	0.80	0.80	0.80	0.95	0.86	0.90	0.90	0.77	0.83
RT	0.87	0.80	0.83	0.86	0.79	0.82	0.88	0.87	0.87
	0.77	0.85	0.81	0.78	0.86	0.82	0.89	0.90	0.89
CNN	0.87	0.87	0.87	–	–	–	0.92	0.96	0.94
	0.88	0.91	0.89	–	–	–	0.96	0.91	0.93
Ensemble MLs	–	–	–	–	–	–	0.95	0.96	0.95
	–	–	–	–	–	–	0.96	0.96	0.96
Specialists in pituitary disease (average) <sup>a</sup>	–	–	–	–	–	–	0.77	0.92	0.84
	–	–	–	–	–	–	0.90	0.73	0.87
Primary care doctors (average) <sup>a</sup>	–	–	–	–	–	–	0.70	0.85	0.77
	–	–	–	–	–	–	0.83	0.68	0.75

Abbreviations: Sensitivity refers to the classifier's ability to correctly detect acromegaly patients who do have the condition; Specificity relates to the test's ability to correctly reject healthy patients without acromegaly condition. NPV = negative predictive value; PPV = positive predictive value; LM = linear model; KNN = k-nearest neighbors; SVM = support vector machine; RT = forests of randomized tree; CNN = convolutional neural network; ML = machine learning.

<sup>a</sup> The values for doctors are calculated based on doctors' diagnosis records on original facial images.



**Fig. 3.** General framework of our method to automatically detect acromegaly from Facial Photographs using Machine Learning Methods: the most left part showed the training processing, which included face detection and normalization from original input images, facial feature extraction and landmark localization, and face frontalization to improve the accuracy of acromegaly diagnosis; the middle left part showed the model training processing, which included Generalized Linear Models (LM), K-nearest neighbors (KNN), Support Vector Machines (SVM), Forests of randomized trees (RT), Convolutional Neural Network (CNN), and Ensemble Method (EM); the middle right part showed the testing processing with the trained models; the most right part showed the evaluation between the ground truth, computers and doctors.

data. Herein, we used OpenCV Cascade Classifier with a Haar Cascade to detect a face from the original photographs and get the bounding rectangle box. Then we cropped and resized all the detected bounding boxes to the same pixel dimensions of 100 \* 100 pixels, as shown in Fig. 4.

#### 4.4.2. Facial Landmark Localization

Once we had the bounding box delineating a face, the next step was to find the locations of the key facial landmarks which are the potential clinical indicators for acromegaly. There has been a lot of effective algorithms as well as high quality open source libraries to extract the key facial landmarks, such as Dlib and CLM. Herein, we used Dlib library, which was a collection of algorithms in machine learning, computer vision and image processing, to extract 64 facial landmarks which included the nose tip, eyes corners, chins, mouth corners, nostril corners, and so forth, as shown in Figs. 3 and 4.

#### 4.4.3. Face Frontalization

Frontalization is the process of reconstructing frontal facing views which can potentially improve the performance of facial landmarks localization and face-recognition. In this study, we used the method proposed by Christosto et al. to produce frontalized views (Sagonas et al.,

2015). Facial feature points extracted in the above steps were used to align the image with a reference face to get the initial frontalized face. If some facial features were poorly visible, the final frontalized face would be synthesized from corresponding symmetric sides.

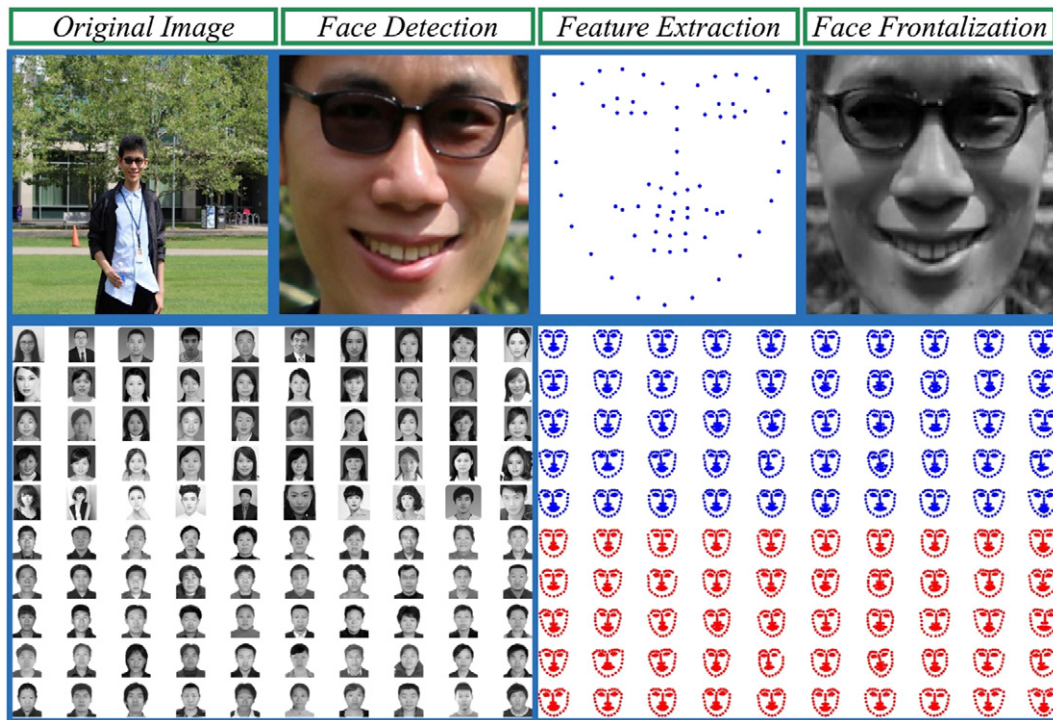
#### 4.5. Algorithm Development

##### 4.5.1. Generalized Linear Models (LM)

Logistic regression, a simple linear model, was firstly tried. In the logistic regression model, the probabilities described the possible outcome of a photograph which was modeled using a logistic function. The output was modeled as a linear combination of input facial information, and the objective function to be minimized with L2 regularization was defined as:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log \left( \exp \left( -y_i (X_i^T w + c) \right) + 1 \right)$$

where  $X$  is the input data,  $y$  is the true output,  $w$  is the weight matrix,  $c$  is the bias,  $C$  is the coefficient.



**Fig. 4.** Flowchart of our proposed facial photograph diagnosis system: the upper figures showed processing of the facial photograph: the first step is to detect faces from the original images, and then extract and localize the features, and the last step is to frontalize faces to improve the accuracy of acromegaly diagnosis; the below figures showed the overview of the inputs to machine learning methods.

#### 4.5.2. K-Near Neighbors (KNN)

Neighbors-based learning was a type of instance-based methods, which did not aim to build an internal model, but simply stored instances of previous learned images. Classification was then obtained from a majority vote of the nearest neighbors: a final outcome was then assigned to the class which had the most representatives for the new input. Herein, we used KNN algorithm to find the  $k$  nearest instances in the training images and computed the mean outcome as the final prediction. The distance between two images was computed using Euclidean distance defined as:

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

where  $q$  and  $p$  are the vectors of the two input images.

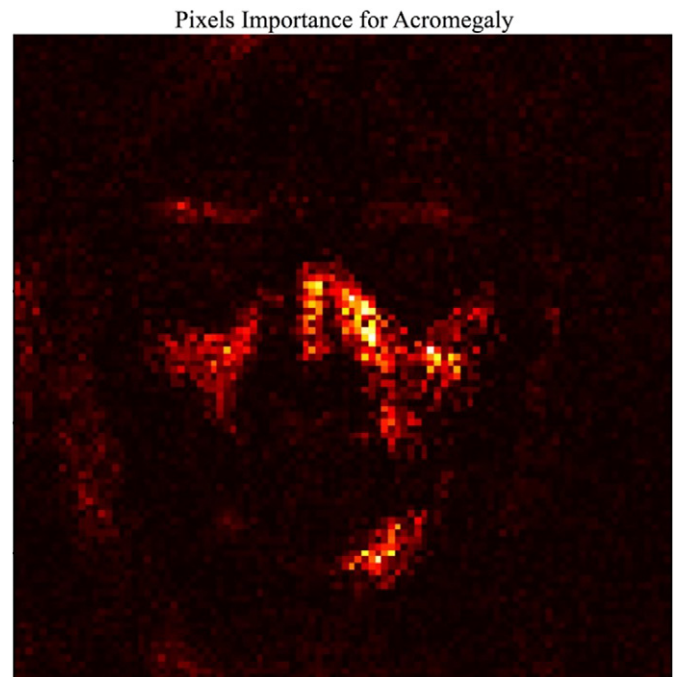
#### 4.5.3. Support Vector Machines (SVM)

SVM was a supervised classifier with a separating hyperplane. Based on a set of training faces marked as normal or acromegaly, our goal was to train a SVM model which was a representation of the labeled photographs that could be divided by a clear gap as wide as possible. New photographs were then mapped into that same space and predicted to a binary category based on which side of the gap they fall. Herein, we performed a grid search using 10-fold cross validation of training images to evaluate different models and thus selected the best hyperparameters for our SVM model. The best hyperparameters set tuned by 10-fold cross validation was linear kernel,  $C$  of 1 and gamma of 0.001.

#### 4.5.4. Forests of Randomized Trees (RT)

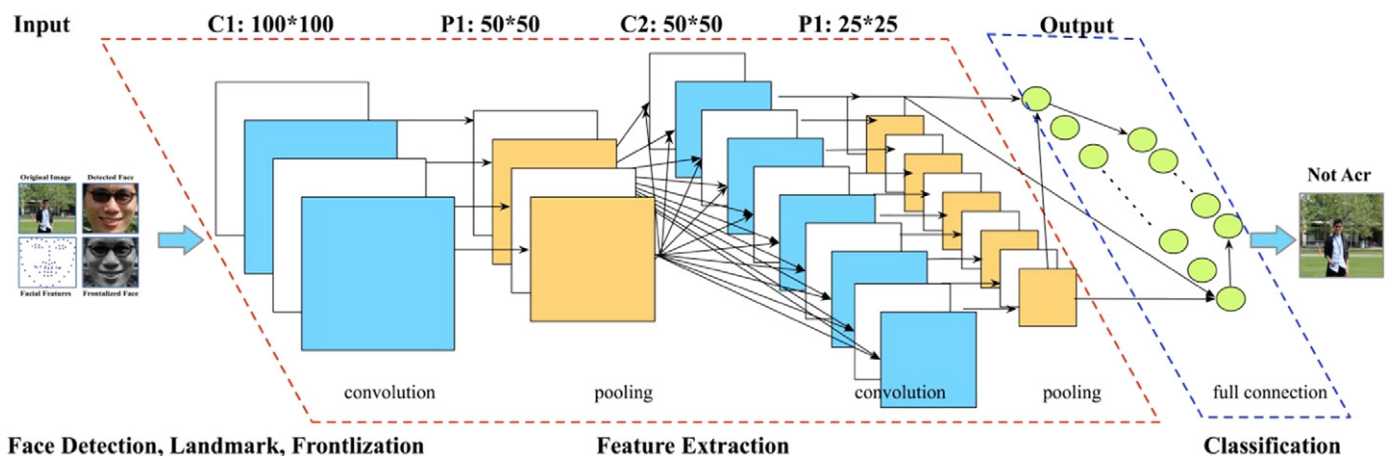
RT was also used to train our model, which was based on perturb-and-combine techniques specifically designed for trees to improve the generalization by introducing randomness in the classifier construction.

Besides, we used RT to evaluate the importance of the pixels in the acromegaly detection, which was shown as Fig. 5: the hotter the pixel, the more important.



**Fig. 5.** Pixels importance ranking of facial photograph for acromegaly learned automatically from the forests of randomized trees: the hotter the pixel (brighter and more yellow), the more important for marking the typical or unique facial features of acromegaly and the higher the pixel ranking is.





**Fig. 6.** Framework of Convolutional Neural Network used for acromegaly diagnosis: the typical types of layers included the Convolutional Layers, Pooling Layers, and Fully-Connected Layers to output the binary classification of acromegaly or non-acromegaly.

#### 4.5.5. Convolutional Neural Network (CNN)

CNN was a powerful artificial neural network method, which could preserve the spatial structure between pixels by learning the internal feature representations using small filters across the whole images, allowing for the patterns in the images to be shifted or translated in the scene but still detectable by the network. Recent research has shown that CNN can achieve state-of-the-art results on many difficult computer vision and natural language processing tasks (Seeliger et al., 2017). There were some typical types of layers for CNN: Convolutional layers, Pooling layers and Fully-Connected layers. Convolutional layers were comprised of filters and feature maps to extract features even from distorted images. Pooling layers following convolutional layers were to downsample, compress and generalize feature representations in this way to reduce the overfitting of the training images. Fully connected layers were the normal flat feedforward neural network layer, which might include a nonlinear activation function to output a binary classification.

As shown in Fig. 6, each image in our dataset was firstly cropped and resized into the same size of 100\*100 pixels. The convolutional layer was the core building block of a CNN consisting of a set of learnable filters with a small receptive field through the full depth of the input volume. Here we used Keras, a high-level neural networks library written in python and capable of running on top of TensorFlow (Chollet, 2016). The model included 2 Convolutional with activation function of RELU, Dropout with rate of 0.2, Convolutional and MaxPooling layers, which were repeated for 3 times with 32, 64, and 128 feature maps. Finally, an additional and larger Dense layer was used at the output end of the network to convert the feature maps to binary classification. The loss function was Binary Cross Entropy, the optimizer was ADAM, and metrics included accuracy, precision and recall. The model was fit over 200 epochs with updates every 10 images.

#### Conflicts of Interest

The authors declare that they have no conflicts of interest.

#### Author Contributions

Y.K. conceived and designed the study. X.K., S.G., L.S. and N.H. collected, analyzed and interpreted the data. All authors participated in the drafting, review, and approval of the report and in the decision to submit for publication.

#### Competing Interests

The authors declare that they have no competing interests.

#### Data and Materials Availability

All participants gave separate written informed consent for publication of this article, the copies of which are all available at any time for review.

#### Acknowledgements

We would like to thank our colleagues at the Department of Neurosurgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College.

#### References

- Chanson, P., et al., 1989. Heart failure responding to octreotide in patient with acromegaly. *Lancet* 1, 1263–1264.
- Chollet, F., 2016. Keras Documentation. available at: <https://keras.io/>.
- Esteva, A., et al., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
- Gencurk, B., Nabiyev, V.V., Ustubioglu, A., Ketenci, S., 2013. 2013 36th International Conference on Telecommunications and Signal Processing (TSP). pp. 817–821.
- Gulshan, V., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410.
- Hamet, P., Tremblay, J., 2017. Artificial intelligence in medicine. *Metabolism* 69S, S36–S40.
- Kanakasabapathy, M.K., et al., 2017. An automated smartphone-based diagnostic assay for point-of-care semen analysis. *Sci. Transl. Med.* 9. <https://doi.org/10.1126/scitranslmed.aai7863>.
- Katznelson, L., et al., 2014. Acromegaly: an endocrine society clinical practice guideline. *J. Clin. Endocrinol. Metab.* 99, 3933–3951.
- Kruszka, P., et al., 2017. 22q11.2 deletion syndrome in diverse populations. *Am. J. Med. Genet. A* 173, 879–888.
- Laurence, Katznelson, Laws Edward, R., Shlomo, Melmed, et al., 2014. Acromegaly: an endocrine society clinical practice guideline. *J. Clin. Endocrinol. Metab.* 99 (3933–3951).
- Learned-Miller, E., et al., 2006. Detecting acromegaly: screening for disease with a morphable model. *Med. Image Comput. Comput. Assist. Interv.* 9, 495–503.
- Melmed, S., 2006. Medical progress: acromegaly. *N. Engl. J. Med.* 355, 2558–2573.
- Melmed, S., et al., 1990. *N. Engl. J. Med.* 322, 966–977.
- Miller, R.E., Learned-Miller, E.G., Trainer, P., Paisley, A., Blanz, V., 2011. Early diagnosis of acromegaly: computers vs clinicians. *Clin. Endocrinol.* 75, 226–231.
- Obermeyer, Z., Emanuel, E.J., 2016. Predicting the future - big data, machine learning, and clinical medicine. *N. Engl. J. Med.* 375, 1216–1219.
- Reddy, R., Hope, S., Wass, J., 2010. Acromegaly. *BMJ* 341, c4189.
- Ribeiro-Oliveira Jr., A., Barkan, A., 2012. The changing face of acromegaly—advances in diagnosis and treatment. *Nat. Rev. Endocrinol.* 8, 605–611.
- Sagonas, C., Panagakis, Y., Zafeiriou, S., Pantic, M., 2015. Face Frontalization for Alignment and Recognition. *arXiv [cs.CV]*. available at: <http://arxiv.org/abs/1502.00852>.



- Schneider, H.J., Sievers, C., Saller, B., Wittchen, H.U., Stalla, G.K., 2008. High prevalence of biochemical acromegaly in primary care patients with elevated IGF-1 levels. *Clin. Endocrinol.* 69, 432–435.
- Schneider, H.J., et al., 2011. A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. *J. Clin. Endocrinol. Metab.* 96, 2074–2080.
- Seeliger, K., Fritsche, M., Güçlü, U., et al., 2017. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*. pii: S1053-8119(17)30586-4. <https://doi.org/10.1016/j.neuroimage.2017.07.018> [Epub ahead of print].
- Tikoo, S., Malik, N., 2016. Detection, segmentation and recognition of face and its features using neural network. *J. Biosens. Bioelectron.* 7. <https://doi.org/10.4172/2155-6210.1000210>.
- Utiger, R.D., 2000. Treatment of acromegaly. *N. Engl. J. Med.* 342, 1210–1211.
- Xie, D., Liang, L., Jin, L., Xu, J., Li, M., 2015. SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception. *arXiv [cs.CV]*. available at. <http://arxiv.org/abs/1511.02459>.