Journal of
*Personalized
Medicine*

MDPI

Article

Facial Recognition Intensity in Disease Diagnosis Using Automatic Facial Recognition

Danning Wu [1,2], Shi Chen [2], Yuelun Zhang [3], Huabing Zhang [2], Qing Wang [4], Jianqiang Li [5], Yibo Fu [1], Shirui Wang [2], Hongbo Yang [2], Hanze Du [2], Huijuan Zhu [2], Hui Pan [6,*] and Zhen Shen [7,8,*]

[1] Eight-Year Program of Clinical Medicine, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100730, China; danie_wu@student.pumc.edu.cn (D.W.); pumcfyb@163.com (Y.F.)
[2] Department of Endocrinology, Key Laboratory of Endocrinology of National Health Commission, Translation Medicine Centre, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100730, China; cs0083@126.com (S.C.); huabingzhangchn@163.com (H.Z.); wangsr13@126.com (S.W.); yanghb@pumch.cn (H.Y.); vespasian_du@126.com (H.D.); shengxin2004@163.com (H.Z.)
[3] Medical Research Center, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100730, China; yuelunzhang@outlook.com
[4] Department of Automation, Tsinghua University, Beijing 100084, China; qing.wang@tsinghua.edu.cn
[5] School of Software Engineering, Beijing University of Technology, Beijing 100124, China; lijianqiang@bjut.edu.cn
[6] Key Laboratory of Endocrinology of National Health Commission, Department of Endocrinology, State Key Laboratory of Complex Severe and Rare Diseases Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China
[7] State Key Laboratory for Management and Control of Complex Systems, Beijing Engineering Research Center of Intelligent Systems and Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[8] Qingdao Academy of Intelligent Industries, Qingdao 266109, China
* Correspondence: panhui20111111@163.com (H.P.); zhen.shen@ia.ac.cn (Z.S.)

Citation: Wu, D.; Chen, S.; Zhang, Y.; Zhang, H.; Wang, Q.; Li, J.; Fu, Y.; Wang, S.; Yang, H.; Du, H.; et al. Facial Recognition Intensity in Disease Diagnosis Using Automatic Facial Recognition. *J. Pers. Med.* **2021**, *11*, 1172. https://doi.org/10.3390/jpm11111172

Academic Editors: Youxin Wang and Ming Feng

Received: 10 October 2021
Accepted: 5 November 2021
Published: 10 November 2021

Abstract: Artificial intelligence (AI) technology is widely applied in different medical fields, including the diagnosis of various diseases on the basis of facial phenotypes, but there is no evaluation or quantitative synthesis regarding the performance of artificial intelligence. Here, for the first time, we summarized and quantitatively analyzed studies on the diagnosis of heterogeneous diseases on the basis on facial features. In pooled data from 20 systematically identified studies involving 7 single diseases and 12,557 subjects, quantitative random-effects models revealed a pooled sensitivity of 89% (95% CI 82% to 93%) and a pooled specificity of 92% (95% CI 87% to 95%). A new index, the facial recognition intensity (FRI), was established to describe the complexity of the association of diseases with facial phenotypes. Meta-regression revealed the important contribution of FRI to heterogeneous diagnostic accuracy ($p = 0.021$), and a similar result was found in subgroup analyses ($p = 0.003$). An appropriate increase in the training size and the use of deep learning models helped to improve the diagnostic accuracy for diseases with low FRI, although no statistically significant association was found between accuracy and photographic resolution, training size, AI architecture, and number of diseases. In addition, a novel hypothesis is proposed for universal rules in AI performance, providing a new idea that could be explored in other AI applications.

Keywords: artificial intelligence; computer-aided diagnosis; facial phenotypes; machine learning; complexity theory

1. Introduction

Many diseases display distinctive facial manifestations, especially endocrine diseases and genetic diseases, including monogenic disorders, chromosomal diseases, and thousands of rare diseases [1]. Recognition by the human eye often causes misjudgment and delays diagnosis due to inconspicuous early facial symptoms associated with these diseases, large individual facial differences, and lack of physicians' knowledge of rare diseases. With the development of artificial intelligence (AI) technology, AI methods have been widely applied in different fields [2–6]. Automatic image recognition based on AI could identify image features for the diagnosis and screening of various diseases, with satisfactory performance for the diagnosis of pulmonary nodules, tumors, fundus diseases, even COVID-19 [7–10]. Among these AI techniques, facial recognition based on artificial intelligence enables computers to detect underlying facial patterns and has played an important role in the diagnosis and screening of diseases with facial phenotypes or changes in recent years [11,12]. It is assumed that artificial intelligence could help to improve diagnostic accuracy and to avoid delayed diagnosis, leading to earlier intervention, conservation of social healthcare resources, and implementation of health policies in the future [12–14]. Different models and systems have been developed to provide possible improvement for diagnostic accuracy [15].

However, there remains a lack of exploration of the factors influencing AI performance or of universal rules to reduce heterogeneity [14]. As has been shown before, diagnostic accuracy of facial recognition for Turner syndrome tended to be lower than that of Down syndrome, although a larger sample size helped to improve it [16,17]. However, the heterogeneity of diseases and AI methods studied and the limited number of works on rare diseases makes it difficult to review and summarize individual studies in a unified manner. Since the complexity theory could be applied to quantitatively describe facial features, this theory needs to be developed to explore the universal rules determining the diagnostic performance of AI based on facial features for heterogeneous diseases.

This is the first study that conducted a systematic review and meta-analysis to summarize the data regarding the diagnosis of heterogeneous diseases on the basis of facial features and explored the universal rules governing the application of facial recognition based on AI in the field of medical diagnosis. We aimed to quantitatively analyze the diagnostic accuracy of facial recognition based on AI, as well as the factors influencing the diagnostic performance and to provide a potential reference for clinical practice. In addition, our study proposes a potential hypothesis for evaluating the performance of AI in other fields, such as image recognition based on AI, and provides a new idea for dealing with heterogeneity when reviewing and analyzing the performance of AI applications.

2. Materials and Methods

2.1. Study Identification and Selection

We searched Medline, PubMed, IEEE, Cochrane Library, EMBASE to identify potential eligible studies published from 1 January 2010 to 15 August 2021. The references of relevant publications were also checked manually. The detailed search strategy containing the index test (facial recognition) and the target condition (diagnosis) is shown in Supplementary Table S1.

Studies were included if they evaluated facial recognition by algorithms of artificial intelligence for the diagnosis of diseases based on facial phenotypes or deformities using photographs and provided sufficient information for quantitative data synthesis. Studies were excluded of they were reviews, lacked a control group, or identified more than one possible disease as a diagnostic result by facial recognition. The titles and the abstracts were screened by two reviewers independently (DW and SC), and the full texts of potentially eligible studies were further screened.

2.2. Data Extraction and Quality Assessment

The data obtained from each study included publication characteristics (authors and year of publication); characteristics of the targeted disease (number of diseases and specific facial features); characteristics of the sample set (data sources, age, sex, and resolution of photographs); characteristics of the index test (algorithms, and number of images used in model training); characteristics of the reference standard (diagnostic criteria); accuracy data (number of true positives, true negatives, false positives, and false negatives). Supplements in each study were also reviewed if available.

Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) was used to assess the risk of bias in patient selection, index test, reference standard, and flow and timing of the included studies. Publication bias was not assessed in our study because there is not a universally accepted method for the review of diagnostic studies to detect publication bias according to the Cochrane Handbook for Diagnostic Tests Review.

2.3. Definition and Calculation of FRI

We defined facial recognition intensity (FRI) as an index to describe the difference of facial features between a studied disease and healthy controls. FRI is calculated as shown in Equation (1) by multiplying the number of independent facial phenotypes of a disease and the maximum penetrance among these facial features.

$$\mathrm{FRI} = \mathrm{Nf} \times \mathrm{Pmax} \tag{1}$$

In Equation (1), Nf represents the number of facial phenotypes relevant to a disease, and Pmax is the maximum penetrance among these facial features, representing the percentage of individuals in a group of patients who exhibited a specific facial phenotype. The facial features and the penetrance of facial phenotypes were collected from the original articles and relevant reviews. If a facial phenotype was associated with a specific group of patients, penetrance was defined to be 100%. Since some of the facial phenotypes were correlated, such as small jaws and crowded teeth, associated phenotypes were counted only once to calculate FRI. For example, Down syndrome displayed nine independent facial phenotypes, and the maximum penetrance of these facial phenotypes was 100% [18]; hence, FRI of Down syndrome was calculated by multiplying 9 by 100%, resulting in 9. FRI was defined to summarize the common characteristics of objects, e.g., facial phenotypes in the presence of different diseases, and to minimize heterogeneity among objects analyzed by AI methods so to make them comparable in the subsequent analysis of performance of facial recognition based on AI for disease diagnosis.

2.4. Statistical Methods

Extracted two-by-two data are graphically shown in a forest plot with the point estimate of sensitivity and specificity and their 95% CIs. Considering the unclear and heterogeneous thresholds for diagnosing different disease with facial phenotypes by facial recognition methods, we used a quantitative random-effects model with bivariate mixed-effects binary regression to combine the sensitivity and specificity and to estimate the summary receiver operating characteristic (SROC) curve. The combined SROC curve and the optimum diagnostic threshold with 95% confidence region and 95% prediction region were plotted. Subgroup analyses and meta-regression were used to explore the heterogeneity between studies. Facial recognition intensity (FRI) and sample size of the training set were analyzed as covariates in meta-regression to explore quantitative relationships with diagnostic accuracy of facial recognition. The result of the meta-regression is shown in a bubble chart and demonstrates a fitting straight line. In addition to FRI and sample size of the training set, we also estimated the following covariates in subgroup analysis: resource of the control group, photo resolution, number of included diseases, and model of facial recognition. Covariates with statistically significant coefficients were regarded as a source of heterogeneity. The robustness of the main results was evaluated by sensitivity analyses. We explored the effect of excluding studies not reporting the model of facial recognition or gold standard of targeted conditions and those using internal validation to evaluate the models.

Data analysis for this paper was performed using Stata Statistical Software 16 (StataCorp., College Station, TX, USA) with two-tailed probability of Type I error of 0.05 ($\alpha = 0.05$).

3. Results

3.1. Systematic Review

Figure 1 shows the flow diagram for filtering articles. We identified 2534 records by electronic search and 29 by hand search. In total, 141 full-text articles were assessed for eligibility, and 20 studies in 14 publications met our criteria for inclusion. Ozdemir et al. [19] included three studies, and Basel-Vanagaite et al. [20], Gurovich et al. [2], Zhao et al. [17], and Saraydemir et al. [16] included two studies using different sample sets in one publication.

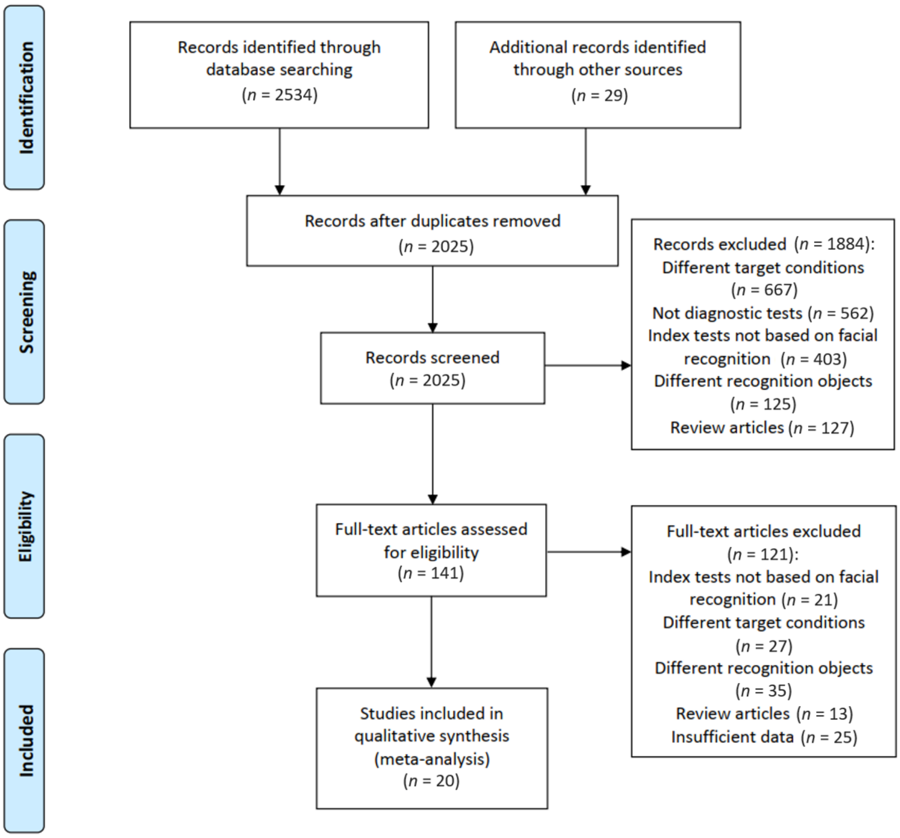
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 2534)
Additional records identified through other sources (n = 29)
Records after duplicates removed (n = 2025)
Records screened (n = 2025)
Records excluded (n = 1884): Different target conditions (n = 667) Not diagnostic tests (n = 562) Index tests not based on facial recognition (n = 403) Different recognition objects (n = 125) Review articles (n = 127)
Full-text articles assessed for eligibility (n = 141)
Full-text articles excluded (n = 121): Index tests not based on facial recognition (n = 21) Different target conditions (n = 27) Different recognition objects (n = 35) Review articles (n = 13) Insufficient data (n = 25)
Studies included in qualitative synthesis (meta-analysis) (n = 20)

Figure 1. Flow chart for study inclusion and exclusion. The titles and the abstracts were screened by two reviewers independently, and the full texts of potentially eligible studies were further screened.

The detailed characteristics of the eligible studies are shown in Supplementary Table S2. The total number of subjects tested in the included studies was 12,557. A single disease was targeted in 16 studies, including 3 studies on Cornelia de Lange syndrome [2,20], 2 on Turner syndrome [21,22], 3 on Down syndrome [16,17], 1 on Angelman syndrome [2], 4 on acromegaly [23–26], 2 on Cushing's syndrome [27,28], and 1 study on fetal alcohol spectrum disorders (FASD) [29], as multiple diseases were detected in 4 studies [17,19]. Nine studies used photographs from public databases and web pages [2,25,27], and 11 studies obtained their photographs in local hospitals [20–24]. Ten studies described the demographic characteristics of their study population, reporting a percentage of males ranging from 0 to 66.2% [16,17,21,22,24–26]. The diagnostic criteria of the targeted diseases were reported in 12 studies and included analysis of gene mutation [2,20] and karyotype [16,17,21,22], success of previous treatment [23], experts' opinions [26], diagnostic tests [24,27,29]. An internal validation set was used for evaluation of the model in 12 studies [16,17,19,21,26–29], and an external validation set was reported in 8 studies [2,20,22–25]. Nine studies included a healthy control group [2,17,19,20,22], and patients with other diseases were included in 11 studies as a control group [16,17,21,23–29]. Apart from 5 studies not reporting the used AI architecture [17,19,20,26,27], several types of machine learning models were applied in 15 studies, including 7 studies using algorithms of deep learning and neural network [2,20,22,28,29] or a combination of neural network and other models [24]. The following models were also reported: SVM [16,21,23], Haar cascade classifier [25], hierarchical decision tree [19], k-NN [16,19] and combination of conventional models [11]. Fourteen studies reported a resolution of photographs ranging from 100 × 100 to 1500 × 1000 pixels [2,16,17,19,21,22,24–26,28]. The number of photographs used to train the model was reported in 20 studies and ranged from 30 to 3465, whereas the number of photographs in the testing set ranged from 17 to 242 [2,16,17,19–29].

3.2. Risk of Bias Assessment of the Eligible Studies

Supplementary Tables S2 and S3 show the results of the risk of bias assessment of the included studies. Regarding patient selection, risk of bias was unclear in 4 studies due to the insufficient information describing the sampling method [2,20] and high in 16 studies with a case–control design [16,17,19,21–29]. With respect to the index test, facial recognition was based on artificial intelligence algorithms without knowledge of the clinical diagnosis in all studies. As for the reference standard, risk of bias was low in 15 studies [2,16,17,20–22,24,26–29] and unclear in 5 studies that did not report the reference standard or an interpretation [19,23,25]. In the domain of flow and timing, risk of bias was low in 16 studies [2,16,17,20–23,25–29], unclear in 3 studies that did not report the reception of the reference standard [19], and high in 1 study because not all patients were subjected to the two tests assessed in the study [24].

3.3. Meta-Analysis

Figure 2 shows the paired forest plot for sensitivity and specificity with the corresponding 95% CIs for each study. Eligible studies were further combined, and the summary receiver operating characteristic (SROC) curve is shown in Figure 3 with the 95% confidence region and 95% prediction region. We calculated the following summarized estimates using random-effects models with 95% confidence interval (CI): sensitivity 89% (95% CI 82% to 93%), specificity 92% (95% CI 87% to 95%), positive likelihood ratio 11.1 (95% CI 6.5 to 18.8), negative likelihood ratio 0.12 (95% CI 0.08 to 0.20), and diagnostic odds ratio (OR) 90 (95% CI 35 to 230).

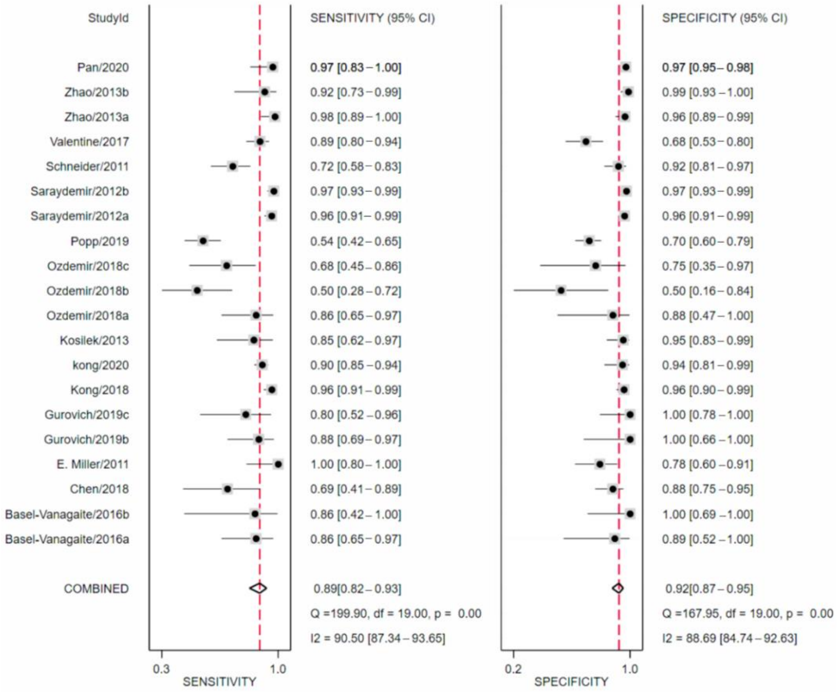

Figure 2. Forest plots of sensitivity and specificity in automatic diagnosis by facial recognition.

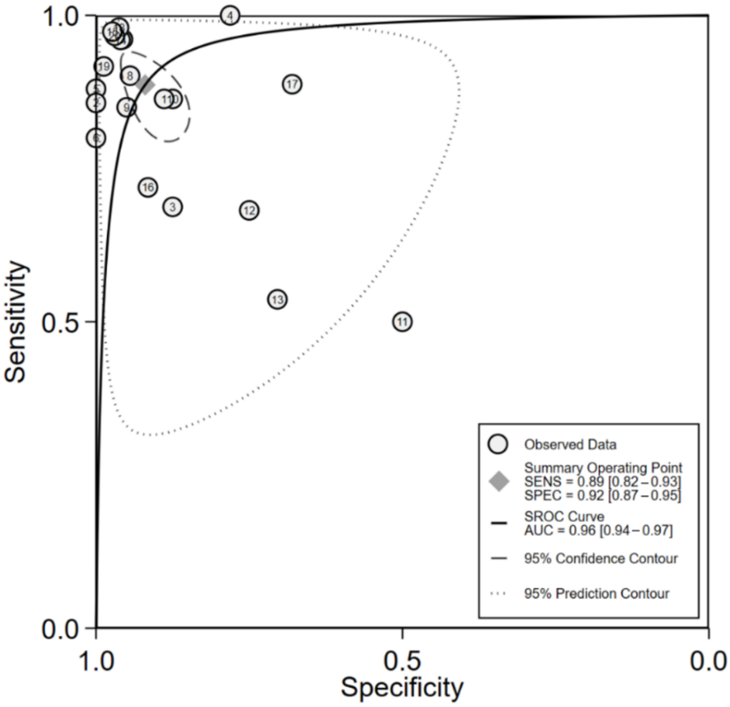

Figure 3. Summary receiver operating characteristics (SROC) curves of eligible studies. The dashed line indicates the 95% confidence region, and the dotted line indicates the 95% prediction region.

3.4. Sensitivity Analysis

After excluding eight studies that evaluated the models with an external validation set [2,20,22–25], pooled sensitivity was 86% (95% CI 75% to 93%), and specificity was 90% (95% CI 82% to 95%). After excluding studies with unclear models [17,19,20,26,27], pooled sensitivity was 90% (95% CI 83% to 94%), and specificity was 91% (95% CI 84% to 96%). After excluding studies with an unclear reference standard [17,20,25,28], pooled sensitivity was 89.0% (95% CI 82.0% to 94.0%), and specificity was 93.0% (95% CI 88.0% to 96.0%). Since these estimates were similar to the main results for the whole dataset, we did not find evidence that the overall combined estimates were influenced by external validation sets, unclear models, or unclear reference standards.

3.5. Evaluation of Facial Recognition Intensity (FRI)

Table 1 shows the prevalence, facial phenotypes of disease, and maximum penetrance of the phenotypes in the eligible studies. Among 16 studies targeting a single disease, Down syndrome showed 9 specific facial phenotypes, and the maximum penetrance of the facial phenotypes was 100% [18]; hence, the calculated FRI of Down syndrome was 9. As for Cornelia de Lange syndrome [2,20], it showed nine facial phenotypes, and the maximum penetrance was 82.7% according to the international consensus statement [30]. After calculation, FRI of Cornelia de Lange syndrome was 7.443. Angelman syndrome showed six facial features, with maximum penetrance of facial phenotypes of 100% and FRI of 8. Turner syndrome showed six facial phenotypes and the maximum penetrance of facial phenotypes was 56% [31]; therefore, FRI of Turner syndrome was 3.36. Fetal alcohol spectrum disorders (FASD) were associated with four facial phenotypes with maximum penetrance of 100% [29], resulting in FRI of 4.

Table 1. Assessment of facial recognition intensity (FRI) of diseases in the eligible studies.

Disease	Prevalence	Maximum Penetrance (Pmax)	Facial Phenotypes		Facial Recognition Intensity (FRI)
			Independent Facial Phenotypes	Number of Facial Phenotypes (Nf)	
Down syndrome [16,17]	1/300~1000	100%	Short face Upward slanting eyes Epicanthus Brushfield spots (white spots on the colored part of the eyes) Low-set ears Small ears Flattened nose Small mouth Protruding tongue	9	9
Acromegaly [23–26]	7/1000	100%	Forehead bulge Prominent jaw Prominent zygomatic arch Deep nasolabial folds Enlarged nose Enlarged brow Enlarged ear Enlarged lip	8	8
Cornelia de Lange Syndrome [2,20]	1/10,000~1/30,000	82.7%	Short face Small jaw Arched eyebrows Joined eyebrows Short nose Forward nostril Long philtrum Thin upper lip Upturned corners of the mouth	9	7.443
Angelman syndrome [2]	1/20,000~1/12,000	100%	Narrow bifrontal diameter Huge jaw Almond-shaped palpebral fissures Narrow nasal bridge Thin upper lip Protruding tongue	6	6
Cushing's syndrome [27,28]	4/100,000	100%	Red face Full moon face Acne Excessive hair Chemosis conjunctiva	5	5
Fetal alcohol spectrum disorders (FASDs) [29]	7.7/1000	100%	Small head Short palpebral fissures Smooth philtrum Thin vermilion border of the upper lip	4	4
Turner syndrome [21,22]	1/2500	56%	Small jaw Epicanthus Ptosis Ocular hypertelorism Low-set ears Multiple facial nevi	6	3.36

Among endocrine diseases, acromegaly showed eight facial phenotypes [28]. Since the maximum penetrance was 100%, FRI of acromegaly was 8. Cushing's syndrome showed five facial phenotypes and maximum penetrance of facial phenotypes of 100% [27,28], resulting in FRI of 5.

3.6. Effect of FRI on the Accuracy of Facial Recognition

Table 2 shows the results of random-effects model meta-regression analysis exploring the relationship between facial recognition intensity (FRI), sample size of the training set, and diagnostic accuracy of facial recognition. The coefficient of FRI in the model was 0.4868 (95% CI 0.0935 to 0.8800, p = 0.015), revealing a significant association with natural logarithms of OR of automatic diagnosis by facial recognition. Meanwhile, the sample size of the training set was not associated with diagnostic accuracy of facial recognition, indicating no significant contribution to the heterogeneity between studies.

Table 2. Meta-regression between FRI, sample size of the training set, and ln(OR) of automatic diagnosis by facial recognition. FRI = facial recognition intensity, OR = diagnostic odds ratio. FRI and sample size of the training set were analyzed as covariates in a meta-regression model to explore the heterogeneity between studies. Their coefficient and 95% confidence interval in the model are shown with two-tailed probability of type I error of 0.05 (α = 0.05).

Covariate	Coefficient [95 CI]	*p* Value
Facial recognition intensity (FRI)	0.4939 [0.0710,0.9169]	0.022
Sample size of the training set	0.0004 [−0.0006,0.0014]	0.467

Therefore, after excluding the sample size of the training set from the model, the relationship between facial recognition intensity and diagnostic accuracy of facial recognition was determined as shown in Figure 4. The model with FRI as a variable showed significant association with natural logarithms of OR of automatic diagnosis, with the coefficient of FRI corresponding to 0.4960 (95% CI 0. 0748 to 0.9171, p = 0.021), indicating that a larger FRI value of a disease was significantly associated with a higher diagnostic accuracy by facial recognition. The relationship between FRI value for a disease and diagnostic accuracy is shown in Equation (2):

$$\ln(\mathrm{OR}) = \ln[\mathrm{Se}\,\mathrm{Sp}/((1-\mathrm{Se})\times(1-\mathrm{Sp}))] = 0.4960 \times \mathrm{FRI} + 1.459 \quad (2)$$

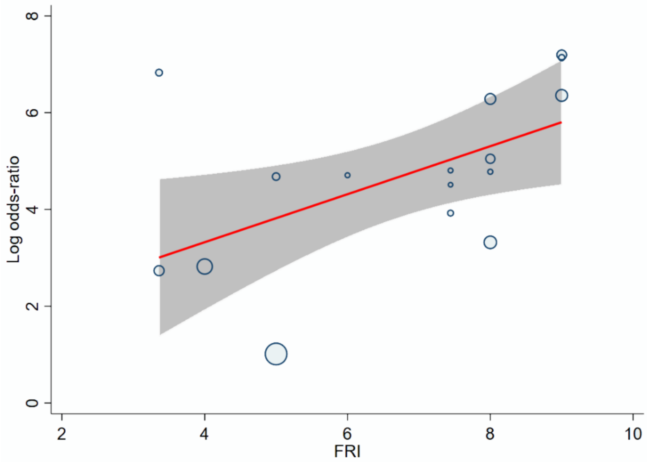

Figure 4. Bubble plots of meta-regression between FRI and ln(OR) of automatic diagnosis by facial recognition. FRI = facial recognition intensity, OR = diagnostic odds ratio. The straight line indicates linear prediction in the meta-regression model between FRI and diagnostic accuracy. The gray zone indicates the 95% confidence region, and the round bubbles represent the eligible studies. The size of the bubbles indicates the impact on the model.

According to Equation (2), Table 3 shows the quantitative association between FRI and accuracy of automatic diagnosis by facial recognition. When both sensitivity and specificity reached 85%, it was required that the FRI value of a disease reached 4.05. When sensitivity and specificity rose to 90%, FRI should correspondingly increase to 5.92. FRI needed to reach 8.93 to ensure that the sensitivity and specificity reached 95%.

Table 3. Association between FRI and accuracy of automatic diagnosis by facial recognition. FRI = facial recognition intensity, OR = diagnostic odds ratio. Quantitative relationship between FRI and diagnostic accuracy (including Figure 2. in meta-analysis. ln (OR) = ln [Se Sp/(1 − Se) (1 − Sp)] = 0.4951 × FRI + 1.46.

Sensitivity	Specificity	OR	ln(OR)	FRI
85%	85%	32.11	3.47	4.05
90%	85%	51.00	3.93	4.98
90%	90%	81.00	4.39	5.92
95%	90%	171.00	5.14	7.42
95%	95%	361.00	5.89	8.93

3.7. Effect of Sample Size of the Training Set and AI Model on the Accuracy of Facial Recognition

Table 4 lists the range of FRI, sample sizes of the training set, AI models, as well as relative median and range of diagnostic accuracy by facial recognition. As for the sample size of the training set, which ranged from 30 to 3465 in the eligible studies, it was shown that the diagnostic accuracy of diseases with FRI higher than 8 was greater than 0.95, even if the sample size of the training set was lower than 100, with the minimum sample size being 30. Diseases with FRI ranging from 6 to 8 showed relatively low diagnostic accuracy when the sample size of the training set was lower than 100, with the minimum sample size being 49, and the accuracy increased with the sample size. The minimum training size for diseases with FRI lower than 6 was 60, and a sample size greater than 1000 significantly improved the diagnostic accuracy of facial recognition, indicating that a modest increase in the sample size of the training set played an important role in improving the diagnostic accuracy of diseases with low FRI.

Table 4. Association between FRI, sample size of the training set, AI models, and accuracy of automatic diagnosis by facial recognition. FRI = facial recognition intensity, DL = deep learning. The diagnostic accuracy is shown as median (minimum, maximum).

FRI	Minimum Sample Size of Training Set	Range of Sample Size of Training Set	Range of Accuracies		AI Models	Range of Accuracies	
			Sensitivities	Specificities		Sensitivities	Specificities
>8	30	<100	0.967 (0.960~0.973)	0.967 (0.960~0.973)	Non-DL	0.973 (0.960~0.977)	0.962 (0.960~0.973)
		100~200	0.977	0.962			
6~8	49	<100	0.710	1.000	Non-DL	0.810 (0.719~0.901)	0.972 (0.944~1.000)
		100~1000	0.790 (0.719~0.860)	0.903 (0.890~0.915)	DL	0.860 (0.800~0.960)	1.000 (0.890~1.000)
		>1000	0.901 (0.800~0.960)	1.000 (0.944~1.000)			
<6	60	<100	0.769 (0.688~0.850)	0.913 (0.875~0.950)	Non-DL	0.688	0.875
		100~1000	0.714 (0.537~0.890)	0.697 (0.690~0.704)	DL	0.929 (0.890~0.967)	0.830 (0.690~0.970)
		>1000	0.967	0.970			

AI methods also showed a similar trend. Diagnostic accuracy of AI reached more than 0.95 with non-deep learning models for diseases with FRI higher than 8, and the application of deep learning models contributed to a higher sensitivity for diseases with lower FRI. Especially for diseases with FRI lower than 6, the median sensitivity improved from 0.688 to 0.929 by using deep learning models. However, the specificity was not influenced by the use of deep learning models.

3.8. Sources of Heterogeneity

Table 5 shows the detailed results of subgroup analyses exploring the potential source of between-study heterogeneity. Facial feature strength was significantly associated with diagnostic accuracy by facial recognition (p = 0.003). However, we found no association between facial recognition's accuracy and photographic resolution, sample size of training sets, model of machine learning, number of targeted diseases, and selection of the control group.

Table 5. Subgroup analyses for the accuracy of automatic diagnosis by facial recognition. Image resolution was calculated by multiplying column pixels by row pixels. If images of different resolution were used, the average resolution was calculated. The two-tailed probability of type I error was 0.05 (α = 0.05).

Subgroup Variables	Numbers of Eligible Studies	Sensitivity, % [95 Cl]	Specificity, % [95 Cl]	*p* for Interaction
Image resolution				0.415
<30,000 pixels	7	0.85 [0.73–0.97]	0.90 [0.82–0.98]	
≥30,000 pixels	7	0.90 [0.82–0.98]	0.94 [0.89–0.98]	
Sample size of training set				0.145
<1000	14	0.87 [0.80–0.93]	0.89 [0.84–0.95]	
≥1000	6	0.92 [0.86–0.99]	0.97 [0.93–1.00]	
Model/system of AI				0.802
Neural network	7	0.91 [0.83–0.99]	0.93 [0.85–1.00]	
Non-neural network	8	0.92 [0.86–0.97]	0.92 [0.86–0.98]	
Number of diseases				0.930
1	16	0.90 [0.86–0.95]	0.78 [0.60–0.97]	
>1	4	0.93 [0.89–0.97]	0.88 [0.74–1.00]	
Selection of control group				0.573
Healthy	9	0.85 [0.75–0.95]	0.94 [0.89–0.99]	
Other diseases	11	0.90 [0.84–0.96]	0.91 [0.86–0.97]	
Facial recognition intensity (FRI)				0.003
≤6	7	0.81 [0.71–0.90]	0.90 [0.83–0.96]	
>6	9	0.95 [0.92–0.98]	0.95 [0.91–0.98]	

4. Discussion

At present, artificial intelligence methods have been widely applied in different fields. However, studies exploring factors influencing the diagnostic accuracy of these methods, as well as systematic reviews and meta-analyses summarizing AI application in the diagnosis of heterogeneous diseases are still lacking. To our knowledge, this is the first study that fills this gap by summarizing heterogeneous studies on the automatic diagnosis of diseases on the basis of facial features and quantitatively analyzes the diagnostic capability of facial recognition based on AI. The review and meta-analysis were conducted strictly following the guidelines for diagnostic reviews [32]. Comprehensive and large-scale studies published so far were included, searched in both medical databases and engineering and technology databases. Representative and high-quality studies focused on different diseases using various known AI methods and were conducted in different countries. Our study summarized and quantitatively analyzed heterogeneous studies on the automatic diagnosis of different diseases based on facial features, showing a pooled sensitivity of 89% (95% CI 82% to 93%) and a specificity of 92% (95% CI 87% to 95%), similar to the results of previous meta-analyses on automatic image recognition for diabetic retinopathy screening [8,33,34], colorectal neoplasia, and breast cancer [35–38], indicating a promising diagnostic performance of facial recognition based on AI for heterogeneous diseases. A sensitivity analysis was conducted to evaluate the robustness of the results. The results were interpreted logically and adapted to clinical applications.

We propose a new index, facial feature intensity (FRI), to reflect the complexity of facial features associated with a targeted object. FRI was defined to minimize the heterogeneity across objects in AI applications and is calculated by multiplying the number of independent facial phenotypes by the maximum penetrance of these facial phenotypes. The number of details in facial features determines the complexity that distinguishes facial features of the targeted object from those of other objects, and the penetrance is the proportion of patients showing a certain complexity of facial features. Since FRI was revealed as the most important influencing factor for the diagnostic accuracy of facial recognition based on AI, the complexity of a targeted object plays the most important role in AI performance, rather than AI technology itself. According to Equation (2) in the meta-regression analysis, the expected accuracy of facial recognition for detecting a disease with the known FRI value could be predicted by calculation, which is of great clinical value.

The interactions between AI parameters and FRI were also taken into consideration, including sample size of the training set and AI architecture. The results revealed that, although larger training size and selection of deep-learning models did not contribute significantly to the heterogeneity between studies in either meta-regression or subgroup analysis, they showed a trend indicating improved diagnostic accuracy for diseases with lower FRI. An appropriate increase in the size of the training samples and the use of deep-learning models improved the accuracy of facial recognition, revealing that the improvement of AI parameters contributed to a better performance of AI for objects with low complexity. This finding is also supported by results on the detection of breast cancer, showing that increasing the training set size would not increase the diagnostic accuracy continuously [38]. Since the number of patients with rare diseases is limited, this finding is clinically significant as it indicates that the sample size of the training set can be within reasonable limits in AI applications. Moreover, the existing AI models have still to be improved to increase the diagnostic accuracy by facial recognition. Therefore, technology innovation is needed, and new AI methods might show better diagnostic accuracy by facial recognition.

Moreover, according to our findings, we propose a new hypothesis regarding AI application, that we named object's complexity theory (OCT) and that could be expanded to the application of AI technology in other fields. According to OCT, within the limits of a reasonable research design, the complexity of the targeted objects determines the complexity of AI processing and plays the most important role in AI performance, while improvement of AI parameters contributes to a better performance of AI for objects with low complexity. The hypothesis is consistent with existing evidence and is supported by previous theorems. According to the complexity theory proposed by J. Hartmanis and R. E. Stearns in 1965, the deep commonalities typical of complex systems determine the process of solving problems, which is relevant in diverse fields [39]. OCT represents the development and extension of the complexity theory regarding the performance of AI applications. According to the No Free Lunch Theorem (NFLT) for artificial intelligence proposed by David Wolpert and William Macready in 1996 [40] and optimized in 1997 [41], an algorithm performing well on a certain object paid with degraded performance on all remaining objects. If we use i to index the examined objects arbitrarily and O_i to represent an object, the NFLT is represented by Equation (3)

$$\sum_k f(O_k, a_i) = \sum_k f(O_k, a_j),\ \forall i, j \tag{3}$$

where a_i and a_j are algorithms, and $f(O_k, a_i)$ is the performance of a_i on the object O_k. The equation shows that the overall performances of all the algorithms were the same. The only way a strategy could outperform another is to specialize the structure of the specific object under consideration [42]. As for our hypothesis, OCT, based on the application in facial recognition, we can establish Equation (4), on the basis of NFLT:

$$f(O_k, a_i) = g(O_k, \mathrm{FRI}_k),\ \forall i,\ \text{if } \mathrm{FRI}_k \geq 6 \tag{4}$$

where $g(O_k, \mathrm{FRI}_k)$ is the performance of the algorithm a_i on the k-th object. The equation revealed that the structure of the object is reflected in the FRI. For objects with a large enough FRI, independently of the parameters of AI technology, the performances are more or less the same. The theory provides a new idea, suggesting that more indices for the evaluation of the complexity of targeted objects should be explored and developed in further studies to better determine AI performance in other fields.

Moreover, OCT and its application in facial recognition provide a new idea to deal with heterogeneity in studies and to evaluate the complexity of targeted objects. OCT should be applied and developed in further studies to determine AI performance in other fields. For image recognition based on AI, facial feature intensity (FRI) could also be converted into image feature intensity (IRI) to describe the characteristics of images related to more diseases. IRI might be the most important factor for AI performance within the limits of a reasonable sample size and of the study design. Previous studies have demonstrated that the image characteristics of diseases play an important role in the performance of image recognition by AI methods [43], including the automatic screening of pulmonary nodules [7,44,45], referable glaucomatous optic neuropathy (GON) [46], colorectal adenoma and polyps [47,48], which also indicates that IRI describes image characteristics of diseases and is critical for AI performance in automatic image recognition. As has been shown before for diabetic retinopathy screening, no statistically significant contribution to heterogeneous diagnostic accuracy has been demonstrated for sample size of the training sets and architecture of convolutional neural networks [34]. Therefore, the complexity theory explains the relationship between complexity of a disease and AI performance and should be extended to other AI applications.

There are some limitations in our study. First, the photographs overlapped in several studies using the same data sources, and it was difficult to eliminate this and evaluate its influence. Second, the risk of bias for the domain of patient selection was high or unclear in several studies. More than half of the studies had a case–control design, due to the limited number of patients with rare diseases. In addition, no traditional thresholds were mentioned in these studies, and we could only compare the sensitivity and specificity by finding the best cut-off point.

5. Conclusions

We quantitatively analyzed studies on the association of heterogeneous diseases with facial features and revealed the promising diagnostic performance of facial recognition based on AI in detecting diseases on the basis of facial features. A new index, facial feature intensity (FRI), was proposed to describe the complexity with facial features associated with different diseases, which was proved to be the most important factor influencing diagnostic accuracy by facial recognition. In addition, we explored the universal rules governing facial recognition based on AI in the field of medical diagnosis and provide a potential reference to solve practical problems in AI applications. An appropriate increase in training sample size and the use of deep learning models might play a role in improving the diagnostic accuracy for diseases with lower FRI. Our study firstly proposes a new hypothesis, the object's complexity theory (OCT), on the performance of AI and provides a new idea for dealing with heterogeneity when evaluating AI performance in other applications.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/jpm11111172/s1, Table S1: Search strategy (Medline), Table S2: Characteristics of included studies, Table S3: Methodological quality assessment of included studies using QUADAS-2.

Author Contributions: S.C. and D.W. contributed equally to this work. S.C., D.W., H.P. and Z.S. designed the study. D.W. and S.C. performed the literature search and appraised the articles. D.W. and Y.Z. performed the analysis with support from Y.F., S.W., H.Y. and H.D., S.C. and D.W. wrote the first draft. H.Z. (Huabing Zhang), Q.W., J.L., H.Z. (Huijuan Zhu), H.P. and Z.S. revised, edited, and finalized the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Beijing Municipal Natural Science Foundation (Grant No. 7192153) and the National Natural Science Foundation of China (Grants 61773382, 61872365, U1909218).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data from this study will be made available upon request from the authors.

Conflicts of Interest: The authors declare that they have no competing interest or personal relationships that could be perceived as prejudicing the impartiality of this study.

References

1. Hurst, A.C.E. Facial recognition software in clinical dysmorphology. *Curr. Opin. Pediatrics* **2018**, *30*, 701–706. [CrossRef]
2. Gurovich, Y.; Hanani, Y.; Bar, O.; Nadav, G.; Fleischer, N.; Gelbman, D.; Basel-Salmon, L.; Krawitz, P.M.; Kamphausen, S.B.; Zenker, M.; et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nat. Med.* **2019**, *25*, 60–64. [CrossRef]
3. Miller, D.D.; Brown, E.W. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am. J. Med.* **2018**, *131*, 129–133. [CrossRef]
4. Huang, S.; Yang, J.; Fong, S.; Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* **2020**, *471*, 61–71. [CrossRef]
5. Loftus, T.J.; Tighe, P.J.; Filiberto, A.C.; Efron, P.A.; Brakenridge, S.C.; Mohr, A.M.; Rashidi, P.; Upchurch, G.R., Jr.; Bihorac, A. Artificial Intelligence and Surgical Decision-making. *JAMA Surg.* **2020**, *155*, 148–158. [CrossRef]
6. Liu, G.; Wei, Y.; Xie, Y.; Li, J.; Qiao, L.; Yang, J.-J. A computer-aided system for ocular myasthenia gravis diagnosis. *Tsinghua Sci. Technol.* **2021**, *26*, 749–758. [CrossRef]
7. Zheng, G.; Han, G.; Soomro, N.Q. An inception module CNN classifiers fusion method on pulmonary nodule diagnosis by signs. *Tsinghua Sci. Technol.* **2020**, *25*, 368–383. [CrossRef]
8. Kaushik, H.; Singh, D.; Kaur, M.; Alshazly, H.; Zaguia, A.; Hamam, H. Diabetic Retinopathy Diagnosis from Fundus Images Using Stacked Generalization of Deep Models. *IEEE Access* **2021**, *9*, 108276–108292. [CrossRef]
9. Alshazly, H.; Linse, C.; Abdalla, M.; Barth, E.; Martinetz, T. COVID-Nets: Deep CNN architectures for detecting COVID-19 using chest CT scans. *PeerJ Comput. Sci.* **2021**, *7*, e655. [CrossRef]
10. Alshazly, H.; Linse, C.; Barth, E.; Martinetz, T. Explainable COVID-19 Detection Using Chest CT Scans and Deep Learning. *Sensors* **2021**, *21*, 455. [CrossRef]
11. Hong, N.; Park, H.; Rhee, Y. Machine Learning Applications in Endocrinology and Metabolism Research: An Overview. *Endocrinol. Metab.* **2020**, *35*, 71–84. [CrossRef] [PubMed]
12. Marwaha, A.; Chitayat, D.; Meyn, M.S.; Mendoza-Londono, R.; Chad, L. The point-of-care use of a facial phenotyping tool in the genetics clinic: Enhancing diagnosis and education with machine learning. *Am. J. Med. Genet. Part A* **2021**, *185*, 1151–1158. [CrossRef] [PubMed]
13. Elmas, M.; Gogus, B. Success of Face Analysis Technology in Rare Genetic Diseases Diagnosed by Whole-Exome Sequencing: A Single-Center Experience. *Mol. Syndromol.* **2020**, *11*, 4–14. [CrossRef]
14. Dias, R.; Torkamani, A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med.* **2019**, *11*, 70. [CrossRef] [PubMed]
15. Saraydemir, S.; Taşpınar, N.; Eroğul, O.; Kayserili, H.; Dinçkan, N. Down syndrome diagnosis based on Gabor Wavelet Transform. *J. Med. Syst.* **2012**, *36*, 3205–3213. [CrossRef] [PubMed]
16. Zhao, X.; Wang, Z.; Gao, L.; Li, Y.; Wang, S. Incremental face clustering with optimal summary learning via graph convolutional network. *Tsinghua Sci. Technol.* **2021**, *26*, 536–547. [CrossRef]
17. Zhao, Q.; Rosenbaum, K.; Okada, K.; Zand, D.J.; Sze, R.; Summar, M.; Linguraru, M.G. Automated Down syndrome detection using facial photographs. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 3670–3673. [CrossRef]
18. Devlin, L.; Morrison, P.J. Accuracy of the clinical diagnosis of Down syndrome. *Ulst. Med. J.* **2004**, *73*, 4–12.
19. Özdemir, M.E.; Telatar, Z.; Eroğul, O.; Tunca, Y. Classifying dysmorphic syndromes by using artificial neural network based hierarchical decision tree. *Australas. Phys. Eng. Sci. Med.* **2018**, *41*, 451–461. [CrossRef]
20. Basel-Vanagaite, L.; Wolf, L.; Orin, M.; Larizza, L.; Gervasini, C.; Krantz, I.D.; Deardoff, M.A. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. *Clin. Genet.* **2016**, *89*, 557–563. [CrossRef]
21. Chen, S.; Pan, Z.X.; Zhu, H.J.; Wang, Q.; Yang, J.J.; Lei, Y.; Li, J.Q.; Pan, H. Development of a computer-aided tool for the pattern recognition of facial features in diagnosing Turner syndrome: Comparison of diagnostic accuracy with clinical workers. *Sci. Rep.* **2018**, *8*, 9317. [CrossRef] [PubMed]
22. Pan, Z.; Shen, Z.; Zhu, H.; Bao, Y.; Liang, S.; Wang, S.; Li, X.; Niu, L.; Dong, X.; Shang, X.; et al. Clinical application of an automatic facial recognition system based on deep learning for diagnosis of Turner syndrome. *Endocrine* **2020**, *72*, 865–873. [CrossRef] [PubMed]

23. Miller, R.E.; Learned-Miller, E.G.; Trainer, P.; Paisley, A.; Blanz, V. Early diagnosis of acromegaly: Computers vs clinicians. *Clin. Endocrinol.* **2011**, *75*, 226–231. [CrossRef]
24. Kong, X.; Gong, S.; Su, L.; Howard, N.; Kong, Y. Automatic Detection of Acromegaly from Facial Photographs Using Machine Learning Methods. *EBioMedicine* **2018**, *27*, 94–102. [CrossRef]
25. Kong, Y.; Kong, X.; He, C.; Liu, C.; Wang, L.; Su, L.; Gao, J.; Guo, Q.; Cheng, R. Constructing an automatic diagnosis and severity-classification model for acromegaly using facial photographs by deep learning. *J. Hematol. Oncol.* **2020**, *13*, 88. [CrossRef] [PubMed]
26. Schneider, H.J.; Kosilek, R.P.; Günther, M.; Roemmler, J.; Stalla, G.K.; Sievers, C.; Reincke, M.; Schopohl, J.; Würtz, R.P. A novel approach to the detection of acromegaly: Accuracy of diagnosis by automatic face classification. *J. Clin. Endocrinol. Metab.* **2011**, *96*, 2074–2080. [CrossRef]
27. Kosilek, R.P.; Schopohl, J.; Grunke, M.; Reincke, M.; Dimopoulou, C.; Stalla, G.K.; Würtz, R.P.; Lammert, A.; Günther, M.; Schneider, H.J. Automatic face classification of Cushing's syndrome in women—A novel screening approach. *Exp. Clin. Endocrinol. Diabetes* **2013**, *121*, 561–564. [CrossRef]
28. Popp, K.H.; Kosilek, R.P.; Frohner, R.; Stalla, G.K.; Athanasoulia-Kaspar, A.; Berr, C.; Zopp, S.; Reincke, M.; Witt, M.; Würtz, R.P.; et al. Computer Vision Technology in the Differential Diagnosis of Cushing's Syndrome. *Exp. Clin. Endocrinol. Diabetes* **2019**, *127*, 685–690. [CrossRef]
29. Valentine, M.; Bihm, D.C.J.; Wolf, L.; Hoyme, H.E.; May, P.A.; Buckley, D.; Kalberg, W.; Abdul-Rahman, O.A. Computer-Aided Recognition of Facial Attributes for Fetal Alcohol Spectrum Disorders. *Pediatrics* **2017**, *140*, e20162028. [CrossRef]
30. Kline, A.D.; Moss, J.F.; Selicorni, A.; Bisgaard, A.M.; Deardorff, M.A.; Gillett, P.M.; Ishman, S.L.; Kerr, L.M.; Levin, A.V.; Mulder, P.A.; et al. Diagnosis and management of Cornelia de Lange syndrome: First international consensus statement. *Nat. Rev. Genet.* **2018**, *19*, 649–666. [CrossRef]
31. Kruszka, P.; Addissie, Y.A.; Tekendo-Ngongang, C.; Jones, K.L.; Savage, S.K.; Gupta, N.; Sirisena, N.D.; Dissanayake, V.H.W.; Paththinige, C.S.; Aravena, T.; et al. Turner syndrome in diverse populations. *Am. J. Med. Genet. Part A* **2020**, *182*, 303–313. [CrossRef] [PubMed]
32. Higgins, J.P.T.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.J.; Welch, V.A. *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed.; John Wiley & Sons: Chichester, UK, 2019.
33. Wu, H.Q.; Shan, Y.X.; Wu, H.; Zhu, D.R.; Tao, H.M.; Wei, H.G.; Shen, X.Y.; Sang, A.M.; Dong, J.C. Computer aided diabetic retinopathy detection based on ophthalmic photography: A systematic review and Meta-analysis. *Int. J. Ophthal.* **2019**, *12*, 1908–1916. [CrossRef] [PubMed]
34. Wang, S.; Zhang, Y.; Lei, S.; Zhu, H.; Li, J.; Wang, Q.; Yang, J.; Chen, S.; Pan, H. Performance of deep neural network-based artificial intelligence method in diabetic retinopathy screening: A systematic review and meta-analysis of diagnostic test accuracy. *Eur. J. Endocrinol.* **2020**, *183*, 41–49. [CrossRef] [PubMed]
35. Posso, M.; Puig, T.; Carles, M.; Rué, M.; Canelo-Aybar, C.; Bonfill, X. Effectiveness and cost-effectiveness of double reading in digital mammography screening: A systematic review and meta-analysis. *Eur. J. Radiol.* **2017**, *96*, 40–49. [CrossRef]
36. Dorrius, M.D.; Jansen-van der Weide, M.C.; van Ooijen, P.M.; Pijnappel, R.M.; Oudkerk, M. Computer-aided detection in breast MRI: A systematic review and meta-analysis. *Eur. Radiol.* **2011**, *21*, 1600–1608. [CrossRef]
37. Hassan, C.; Spadaccini, M.; Iannone, A.; Maselli, R.; Jovani, M.; Chandrasekar, V.T.; Antonelli, G.; Yu, H.; Areia, M.; Dinis-Ribeiro, M.; et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: A systematic review and meta-analysis. *Gastrointest. Endosc.* **2021**, *93*, 77–85.e76. [CrossRef]
38. Hughes, K.S.; Zhou, J.; Bao, Y.; Singh, P.; Wang, J.; Yin, K. Natural language processing to facilitate breast cancer research and management. *Breast J.* **2020**, *26*, 92–99. [CrossRef]
39. Hartmanis, J.; Stearns, R.E. On the computational complexity of algorithms. *Trans. Am. Math. Soc.* **1965**, *117*, 285–306. [CrossRef]
40. Wolpert, D.H. The Lack of a Priori Distinctions between Learning Algorithms. *Neural Comput.* **1996**, *8*, 1341–1390. [CrossRef]
41. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
42. Ho, Y.C.; Pepyne, D.L. Simple explanation of the no-free-lunch theorem and its implications. *J. Optim. Theory Appl.* **2002**, *115*, 549–570. [CrossRef]
43. Tagliafico, A.S.; Piana, M.; Schenone, D.; Lai, R.; Massone, A.M.; Houssami, N. Overview of radiomics in breast cancer diagnosis and prognostication. *Breast* **2020**, *49*, 74–80. [CrossRef] [PubMed]
44. Gong, J.; Liu, J.; Hao, W.; Nie, S.; Wang, S.; Peng, W. Computer-aided diagnosis of ground-glass opacity pulmonary nodules using radiomic features analysis. *Phys. Med. Biol.* **2019**, *64*, 135015. [CrossRef]
45. Beig, N.; Khorrami, M.; Alilou, M.; Prasanna, P.; Braman, N.; Orooji, M.; Rakshit, S.; Bera, K.; Rajiah, P.; Ginsberg, J.; et al. Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. *Radiology* **2019**, *290*, 783–792. [CrossRef]
46. Phene, S.; Dunn, R.C.; Hammel, N.; Liu, Y.; Krause, J.; Kitade, N.; Schaekermann, M.; Sayres, R.; Wu, D.J.; Bora, A.; et al. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* **2019**, *126*, 1627–1639. [CrossRef]

47. Aziz, M.; Fatima, R.; Dong, C.; Lee-Smith, W.; Nawras, A. The impact of deep convolutional neural network-based artificial intelligence on colonoscopy outcomes: A systematic review with meta-analysis. *J. Gastroenterol. Hepatol.* **2020**, *35*, 1676–1683. [CrossRef]
48. Wang, P.; Berzin, T.M.; Glissen Brown, J.R.; Bharadwaj, S.; Becq, A.; Xiao, X.; Liu, P.; Li, L.; Song, Y.; Zhang, D.; et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut* **2019**, *68*, 1813–1819. [CrossRef]