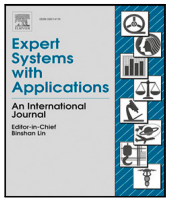
Expert Systems With Applications 252 (2024) 123881

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Simulated multimodal deep facial diagnosis

Bo Jin [a,*], Nuno Gonçalves [a,b], Leandro Cruz [a,c], Iurii Medvedev [a], Yuanyu Yu [d], Jiujiang Wang [d]

[a] Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
[b] Portuguese Mint and Official Printing Office, 1000-042 Lisbon, Portugal
[c] Align Technology, Inc., San Jose, CA 95134, United States
[d] Neijiang City Engineering Technology Research Center of Neurological Disease Information Interference, School of Artificial Intelligence, Neijiang Normal University, Neijiang 641100, China

ARTICLE INFO

Keywords:
Deep facial diagnosis
Simulated multimodal
Face depth estimation
Facial phenotypes
Condition-specific faces
Bilinear

ABSTRACT

Facial phenotypes are extensively studied in medical and biological research, serving as critical markers that potentially indicate underlying genetic traits or medical conditions. With the recent advancements in big data, algorithms, and hardware, deep facial diagnosis, which employs deep learning techniques to systematically examine facial phenotypes and identify signs of certain diseases or medical conditions, has attracted significant attention and research, gradually emerging as a promising tool in precision medicine. Primarily limited by the scarcity of data for training facial diagnosis models, the accuracy of facial diagnosis for various conditions remains low up to now. In the past decade, RGB-D cameras, measuring depth information along with standard RGB capabilities, have proven superior in processing spatial details with more stability and accuracy. Motivated by the facts mentioned above, in this paper, we propose a Simulated Multimodal Framework, which effectively improves the computer-aided facial diagnosis performance of state-of-the-art models in experiments under different conditions. The underlying principle is to leverage the simulated depth by generative models to improve the performance of RGB image recognition. Furthermore, as a rapid and non-invasive tool for disease screening and detection, our proposal demonstrated an average accuracy improvement of over 20% compared to practicing physicians in the study.

1. Introduction

Facial phenotypes are a distinctive set of observable facial characteristics or traits that can be attributed to specific genetic or environmental factors. The study of these phenotypes offers invaluable insights into understanding various health conditions, genetic predispositions, and even potential future risks. Recent advancements in medical and computational research have enabled a deeper analysis into the subtle nuances of facial phenotypes, making it possible to predict or diagnose certain conditions based solely on facial features. This method of prediction or diagnosis based on facial attributes has been historically referred to as facial diagnosis, and it continues to be widely practiced in many regions even today.

The history of facial diagnosis can be traced back to the *Huangdi Neijing* (Unschuld, 2003), which is one of the earliest and most important classics of Chinese medicine. It is believed to have been written around 2,500 years ago. According to the *Huangdi Neijing*, the human face can reflect information about body organs, blood circulation, and overall vitality. By analyzing facial features such as shape, size, position, and proportion, as well as skin color, texture, and wrinkles, practitioners can assess an individual's health condition. Up to today, this practice remains widely used in modern medical practice, both in China and elsewhere in the world. Contemporary research substantiates that specific facial features can indicate certain diseases or health conditions (Fanghänel et al., 2006; Gurovich et al., 2019; Jin et al., 2020). However, a challenge in facial diagnosis is that it often requires doctors to have substantial practical experience.

Even today, in numerous rural and underserved areas, limited access to medical resources makes it difficult for people to receive timely medical examinations, often resulting in treatment delays. Even in metropolitan areas, challenges persist, such as high costs, long wait times at hospitals, shortage of specialist doctors and doctor-patient conflicts that can lead to medical disputes.

Computer-aided facial diagnosis refers to the use of computer algorithms, often incorporating artificial intelligence (AI) and machine learning techniques, to analyze facial features and identify health concerns. It could allow for quick and non-invasive disease screening and

* Corresponding author.
E-mail addresses: jin.bo@isr.uc.pt (B. Jin), nunogon@deec.uc.pt (N. Gonçalves), lmvcruz@gmail.com (L. Cruz), iurii.medvedev@isr.uc.pt (I. Medvedev), yuyuanyu@gmail.com (Y. Yu), tswangjade@163.com (J. Wang).

https://doi.org/10.1016/j.eswa.2024.123881
Received 13 October 2023; Received in revised form 14 March 2024; Accepted 29 March 2024
Available online 4 April 2024

detection. Its convenience and non-contact characteristics make it an ideal tool for telemedicine. However, facial diagnosis is a challenging task for computers, which is not surprising. Due to advancements in hardware, big data, and algorithms, deep learning technology achieved breakthrough progress after 2006, leading to significant achievements in areas such as image recognition, speech recognition, and natural language processing (Jiang et al., 2023; LeCun et al., 2015; Li et al., 2023; Zhang et al., 2020). Deep facial diagnosis is to perform facial diagnosis by using deep learning models. However, the challenges of facial diagnosis tasks still remain, as the data for training deep learning models in facial diagnosis are scarce. So, the research problem is how to obtain a high accuracy in facial diagnosis with limited training data.

An RGB camera, an indispensable vision-based measuring instrument (Shirmohammadi & Ferrero, 2014), operates by processing light signals and independently capturing the red, green, and blue (RGB) channels and then combining them to generate full-color images. However, relying solely on RGB cameras often fails to meet the demand for precise depth perception and 3D object recognition. The RGB-D camera not only captures the visible spectrum to generate standard RGB color channels but also adds a depth channel to measure the distance between the camera and objects within its field of view. Over the past decade, it has been proven that the integration of this depth sensing significantly enhances the robustness and accuracy of imaging, particularly in terms of 3D modeling and spatial analysis. It is conceived that, without requiring additional training image data, generative models can be utilized to simulate depth maps from RGB images, the spatial information of which can improve recognition accuracy. To further make effective use of the simulated depth maps, a Simulated Multimodal Framework for facial diagnosis is proposed.

In this study, we have investigated the following six specific conditions and a healthy control group to effectively validate our findings. Prevalence and incidence are two important indicators used to describe the epidemiology of diseases. Prevalence primarily focuses on the extent to which a disease is present in a population, while Incidence concentrates on the number of new cases that occur. Both of these indicators play a significant role in epidemiological research and the development of public health policies.

Prevalence is the proportion of total cases of a disease or condition in a population at a specific time. It is calculated by dividing the number of cases of a disease or condition in a population by the total number of individuals in that population, as demonstrated in Eq. (1). Prevalence provides a snapshot of how common a disease or condition is in a population at a given time, which includes both new and existing cases of a disease or condition.

$$\text{Prevalence} = \frac{\text{Number of existing cases of a disease}}{\text{Total population}} \tag{1}$$

Incidence refers to the number of new cases of a disease or condition that develop in a population over a specific period of time. It is calculated by dividing the number of new cases of a disease or condition in a population by the total number of individuals at risk in that population, as demonstrated in Eq. (2). Incidence provides information on how quickly a disease or condition is spreading in a population. Incidence includes only new cases of a disease or condition that occurred during the specific period of time and does not include existing cases.

$$\text{Incidence} = \frac{\text{Number of new cases of a disease}}{\text{Number of individuals at risk}} \tag{2}$$

1.1. Acromegaly

Acromegaly is a hormone disorder caused by excessive secretion of growth hormone by the pituitary gland in adulthood, which will lead to abnormal hyperplasia or hypertrophy of organs. A survey shows that the prevalence rate of acromegaly ranges from 2.8 to 13.7 per 100,000 individuals approximately, and the annual incidence rate ranges from 0.2 to 1.1 per 100,000 individuals approximately (Lavrentaki et al., 2017). Acromegaly is not easily noticed by patients for a short period of time, and is often mistaken for a phenomenon of weight gain or normal aging. Acromegaly and related complications such as high blood pressure, diabetes, and heart disease seriously affect patient health, quality of life and longevity. Studies show that if the patients with acromegaly do not receive treatment, the average remaining life is only about 10 years. However, if they receive treatment, their life expectancy will be no different from that of ordinary people (Ho, 2011). Therefore, early diagnosis and treatment are necessary. Acromegaly could cause gradual facial changes. Symptoms of acromegaly that probably appear on the patients' face include a prominent lower jaw, prominent brow bones, an enlarged nose, thickened lips, and wider spacing between teeth, which is shown as Fig. 1(a).

1.2. Down syndrome

Down syndrome (DS) is a genetic disorder caused by trisomy of chromosome 21. Most patients with Down syndrome have physical and intellectual disabilities. Proper care can improve the quality of life for patients with Down syndrome. The estimated prevalence of DS approximately ranges from 136.6 to 142.9 per 100,000 individuals (Canfield et al., 2006; Freeman et al., 2007). According to the World Health Organization (Rodrigues et al., 2019), the incidence of DS approximately ranges from 90.9 to 100 per 100,000 live births worldwide. Symptoms of Down syndrome, which may appear on the patients' face, include small palpebral fissures, wide-set eyes, a low nasal bridge, low-set ears, and more, as illustrated in Fig. 1(b).

1.3. Facial nerve paralysis

Facial nerve paralysis, resulting from a dysfunction of the facial nerve, leads to a loss of control over facial muscles for smiling, blinking, and other facial movements on the affected side. Common causes of facial paralysis include facial nerve infection or inflammation, head trauma, and head or neck tumors. The prevalence of facial nerve paralysis ranges from 11.5 to 40.2 per 100,000 individuals (Kim et al., 2019), and the annual incidence of facial paralysis ranges from 15 to 30 per 100 000 individuals approximately (Tiemstra & Khatkhate, 2007). Facial nerve paralysis may cause numerous complications, including irreversible facial nerve damage, abnormal regeneration of nerve fibers, and partial or complete blindness in eyes that cannot be closed (Coulson et al., 2004). Symptoms of facial nerve paralysis, which likely appear on the patients' face, include paralysis of facial expression muscles on the affected side, disappearance of forehead wrinkles, flattened nasolabial folds, and drooping corners of the mouth, as illustrated in Fig. 1(c).

1.4. Leprosy

Leprosy, also referred to as Hansen's disease, is an infectious condition caused by the bacteria Mycobacterium leprae. It primarily affects the skin, nerves, and mucous membranes. Early detection and treatment are essential, as untreated leprosy can lead to nerve damage, muscle weakness, and eyesight problems. According to the World Health Organization, the incidence of leprosy approximately ranges from 2.5 to 3.2 per 100,000 individuals, and the prevalence of leprosy approximately ranges from 2.2 to 2.7 per 100,000 individuals (World Health Organization et al., 2014, 2020). Leprosy can manifest symptoms on a patient's face, including hair loss, granulomas, pale areas of skin, eye damage, and facial disfigurement, such as the loss of the nose, as illustrated in Fig. 1(d).

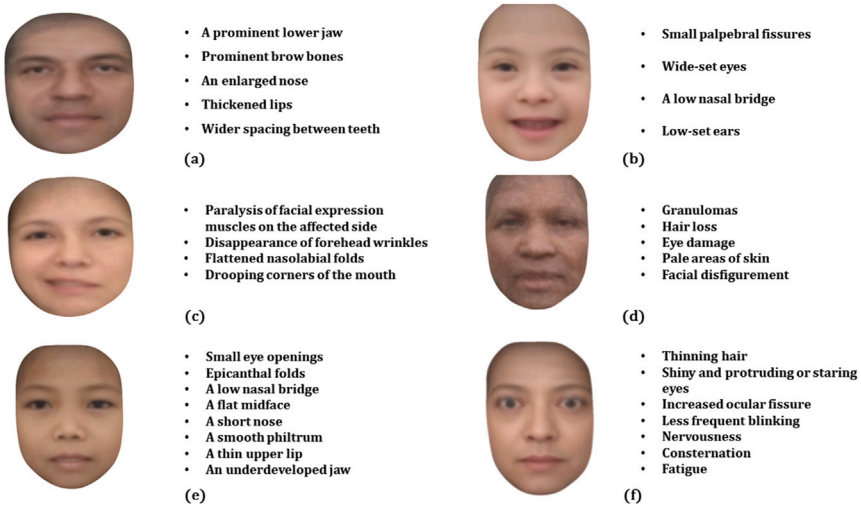


Fig. 1. Condition-specific faces in this study: (a) Acromegaly, (b) Down syndrome, (c) Facial Nerve Paralysis, (d) Leprosy, (e) Thalassemia, (f) Hyperthyroidism.

1.5. Thalassemia

Thalassemia, one of the most prevalent inherited blood conditions globally, is a hereditary disorder caused by irregular hemoglobin production, where hemoglobin consists of two alpha and two beta chains. Different types of globin gene deletions or defects result in the corresponding inhibition of globin chain synthesis. Based on this fact, thalassemia is primarily divided into two types: α and β. The global incidence of thalassemia approximately ranges from 0.74 to 39.79 per 100,000 individuals (Lee et al., 2022; Modell & Darlison, 2008), while the prevalence of thalassemia varies from 2,500 to 15,000 per 100,000 individuals approximately (Hoffman et al., 2018).

Early diagnosis of thalassemia is crucial, as without consistent treatment, it can be fatal in early childhood. Medical research (Alhaija et al., 2002) indicates that thalassemia can lead to facial bone deformities. Facial symptoms of thalassemia include small eye openings, epicanthal folds, a low nasal bridge, a short nose, a smooth philtrum, a flat midface, a thin upper lip, and an underdeveloped jaw, as illustrated in Fig. 1(e).

1.6. Hyperthyroidism

Hyperthyroidism is a prevalent endocrine disorder resulting from an overproduction of the thyroid hormones T3 and T4, which regulate the body's metabolism through various mechanisms. The incidence of hyperthyroidism approximately ranges from 50 to 1300 per 100,000 individuals (Muñoz-Ortiz et al., 2020), and the average prevalence of hyperthyroidism approximately ranges from 800 to 1300 per 100,000 individuals (Manifold, 2005).

If left untreated, hyperthyroidism can lead to severe complications, potentially endangering the patient's life. Distinct facial features associated with hyperthyroidism include shiny and protruding or staring eyes, less frequent blinking, increased ocular fissure, nervousness, consternation, fatigue, and thinning hair as illustrated in Fig. 1(f).

Fig. 2 summarizes the prevalence of the six aforementioned condition categories. Fig. 3 summarizes the incidence of the six aforementioned condition categories.

In this article, our contributions for facial diagnosis could be summarized as follows:

1. We propose a Simulated Multimodal Framework designed to enhance the performance of computer-aided facial diagnosis of six conditions: acromegaly, Down syndrome, facial nerve paralysis, leprosy, thalassemia, and hyperthyroidism.
2. In an implementation of the Simulated Multimodal Framework, we employ the D+GAN model for depth map generation, wavelet soft-threshold fusion for image fusion, and introduce fine-grained image classification to extract features. Comparative experiments are carried out using two state-of-the-art models, Insightface and FaceNet, as baseline models. Given these tools and methods, weighted majority voting is employed in the final stage, where the importance of each model's prediction is directly influenced by its accuracy performance on the training set.
3. Experimental results with different settings demonstrate that the proposed Simulated Multimodal Framework can significantly enhance the facial diagnosis performance of advanced models with high probability. The bilinear models, which adopt the concept of fine-grained classification, also outperform their counterparts without such implementation. The findings further suggest that the estimated depth can contribute to the improvement of 2D facial diagnosis. These findings are reproducible.

The remainder of this article is structured as follows: Chapter 2 provides a review of influential relevant literature. Chapter 3 details our proposed methodologies and their respective implementations. Chapter 4 primarily presents and summarizes the experimental results under various conditions. Chapter 5 mainly interprets the findings within the context of existing studies, explores their practical significance and limitations. Finally, Chapter 6 draws conclusions and charts potential avenues for future research.

2. Related work

In this chapter, we primarily review some classic research on computer-aided facial diagnosis, which is summarized in Table 1.

Schneider et al. performed detection of acromegaly through face classification based on texture and geometry similarity (Schneider et al., 2011). Their dataset includes face images of 57 patients with acromegaly. They claimed to have achieved an accuracy of 81.9%.

Zhao et al. proposed using ICA to identify anatomical facial landmarks to differentiate between individuals with Down syndrome and the healthy population (Zhao, Okada, et al., 2014). Their dataset includes 50 face images of patients with Down syndrome. They claimed to have achieved an accuracy of 0.967 and a F1 score of 0.956.

Zhao et al. proposed identifying patients with Down syndrome by ensembling the outputs of multiple different classifiers (Zhao, Werghi, et al., 2014). Their dataset includes 50 face images of patients with Down syndrome. They claimed to have achieved an accuracy of 0.967.

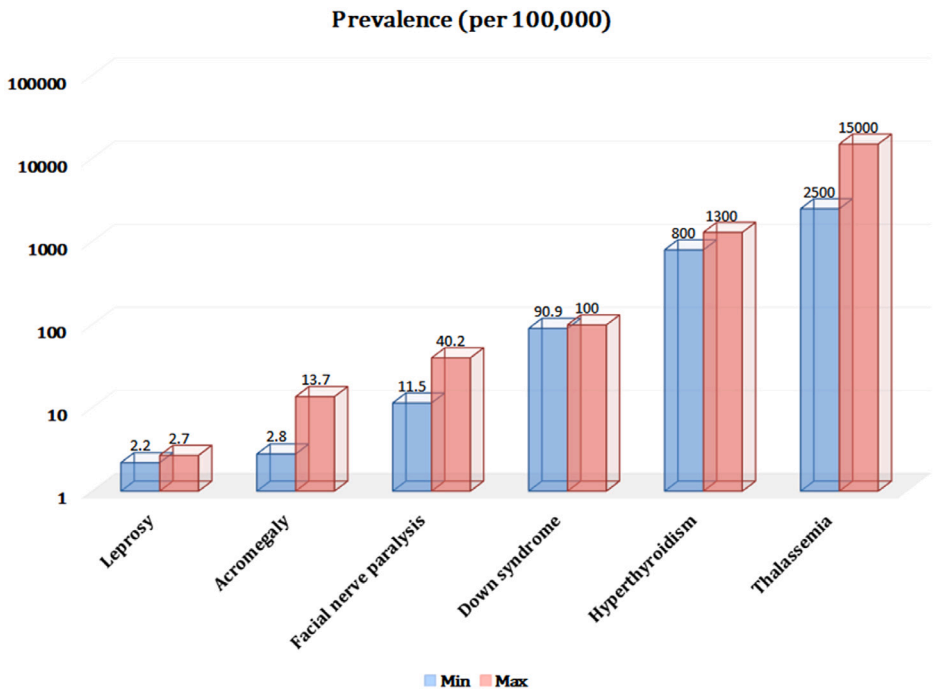


Fig. 2. Prevalence of the six conditions used for the study.

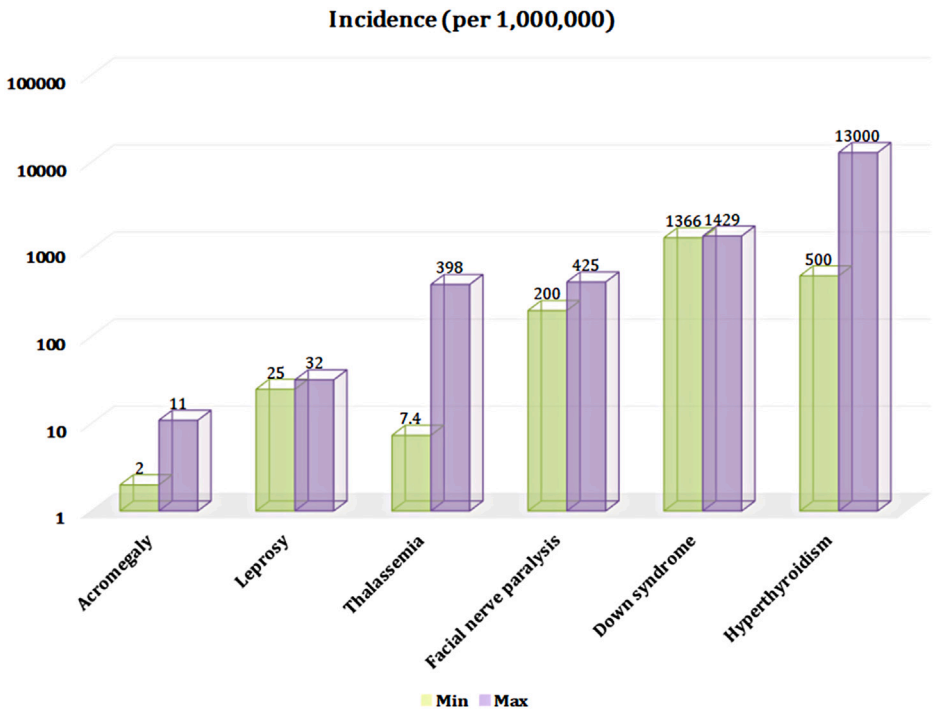


Fig. 3. Incidence of the six conditions used for the study.

Kong et al. adopted a series of basic estimators, including the k-Nearest Neighbors (KNN), Generalized Linear Model (GLM), Support Vector Machines (SVM), Random Forests (RF), and CNN. By employing a voting method, they combined the predictions of these models to detect acromegaly from facial photographs (Kong et al., 2018). Their dataset includes 641 face images of patients with acromegaly. They claimed to have achieved a sensitivity of 96% and a specificity of 96%.

Umeda-Kameyama et al. conducted a study utilizing five deep learning models and two optimizers to differentiate between dementia and non-dementia facial images (Umeda-Kameyama et al., 2021). Their dataset includes 124 face images of patients with cognitive impairment and 250 face images of cognitively sound participants. They reported that the Xception AI system achieved a notable accuracy of 92.56% along with an AUC for the ROC curve of 0.9717. Additionally, they found a significant and negative correlation between the Face AI score and MMSE scores ($r = -0.599$, $p < 0.0001$), indicating the potential of face recognition as a non-invasive tool for dementia screening.

Boehringer et al. performed principal component analysis and linear discriminant analysis for a computer-based diagnosis among the 10 syndromes (Boehringer et al., 2006). Their dataset includes 147

Table 1
Summary of classical studies on facial diagnosis.

Research	Condition	No. of DSF images per category	Cls. problem	Method
Schneider et al. (2011)	Acromegaly	57	2D Binary	Texture and geometry
Zhao, Okada, et al. (2014)	Down syndrome	50	2D Binary	ICA
Zhao, Werghi, et al. (2014)	Down syndrome	50	2D Binary	Ensemble learning
Kong et al. (2018)	Acromegaly	641	2D Binary	Ensemble learning
Umeda-Kameyama et al. (2021)	Alzheimer's disease	124	2D Binary	Deep learning model
Boehringer et al. (2006)	10 syndromes	15	2D Multi-class	LDA
Shukla et al. (2017)	6 disorders	188	2D Multi-class	DCNN
Gurovich et al. (2019)	216 syndromes	81	2D Multi-class	DCNN
Jin et al. (2020)	4 conditions	70	2D Multi-class	Deep transfer learning
Porras et al. (2021)	128 conditions	11	2D Multi-class	Deep neural networks
Hallgrímsson et al. (2020)	396 syndromes	8	3D Multi-class	Parametric methods and machine learning
Bannister et al. (2022)	47 syndromes	100	3D Multi-class	Normalizing flows

facial images with 10 syndromes, which means the average number of disease-specific face images for each category is approximately 15. They claimed to have achieved an accuracy of 75.7% for 10-class classification.

Shukla et al. used deep convolutional neural network to detect 6 disorders (Shukla et al., 2017). Their dataset includes 1126 facial images with 6 disorders, which means the average number of disease-specific face images for each category is approximately 188. They claimed to have achieved an accuracy of 48% for 6-class classification and an accuracy of 98.80% for binary classification.

Gurovich et al. introduced a facial image analysis framework called DeepGestalt, which employs computer vision and deep learning algorithms to quantify similarities to hundreds of syndromes (Gurovich et al., 2019). Their dataset includes 17106 images with 216 different syndromes, which means the average number of disease-specific face images for each category is approximately 79. They claimed to have achieved 61.3~68.7% top-1 accuracy and 89.4~90.6% top-10 accuracy in identifying the correct syndrome on hundreds of images.

Jin et al. proposed using deep transfer learning from face recognition to facial diagnosis, named *Deep Facial Diagnosis* (Jin et al., 2020). Their dataset includes 280 images with 4 different diseases, which means the average number of disease-specific face images for each category is 70. They claimed to have achieved an overall top-1 accuracy of over 90%.

Porras et al. screened children for genetic syndromes by using deep neural networks and facial statistical shape models (Porras et al., 2021). Their dataset includes 1,400 children images with 128 genetic conditions, which means the average number of disease-specific face images for each category is approximately 11. They claimed to have achieved an accuracy of 88% for the detection of a genetic syndrome.

Compared to two-dimensional images, three-dimensional images contain information about the spatial relationships between objects. In light of this, some researchers have started to explore facial diagnosis using three-dimensional facial images.

Hallgrímsson et al. conducted binary classification on 3D human face images using both parametric methods and machine learning techniques (Hallgrímsson et al., 2020). Their dataset includes 3327 images with 396 different syndromes, which means the average number of disease-specific face images for each category is approximately 8. They claimed to have achieved balanced accuracy was 73% and mean sensitivity 49%.

Bannister et al. performed 3D facial surface modeling using deep learning and performed 3D facial diagnosis (Bannister et al., 2022). Their dataset includes 4700 scans with 47 different syndromes, which means the average number of disease-specific face images for each category is approximately 100. They claimed to have achieved overall top-1 accuracy of 71%, and a mean sensitivity of 43% across all syndrome classes.

3. Materials and methods

In this study, we introduce a Simulated Multimodal Framework for facial diagnosis, which is illustrated in Fig. 4 and represents an enhancement of the previous work. In the following subsections, we introduce each module in the order of processing flow.

3.1. Preprocessing

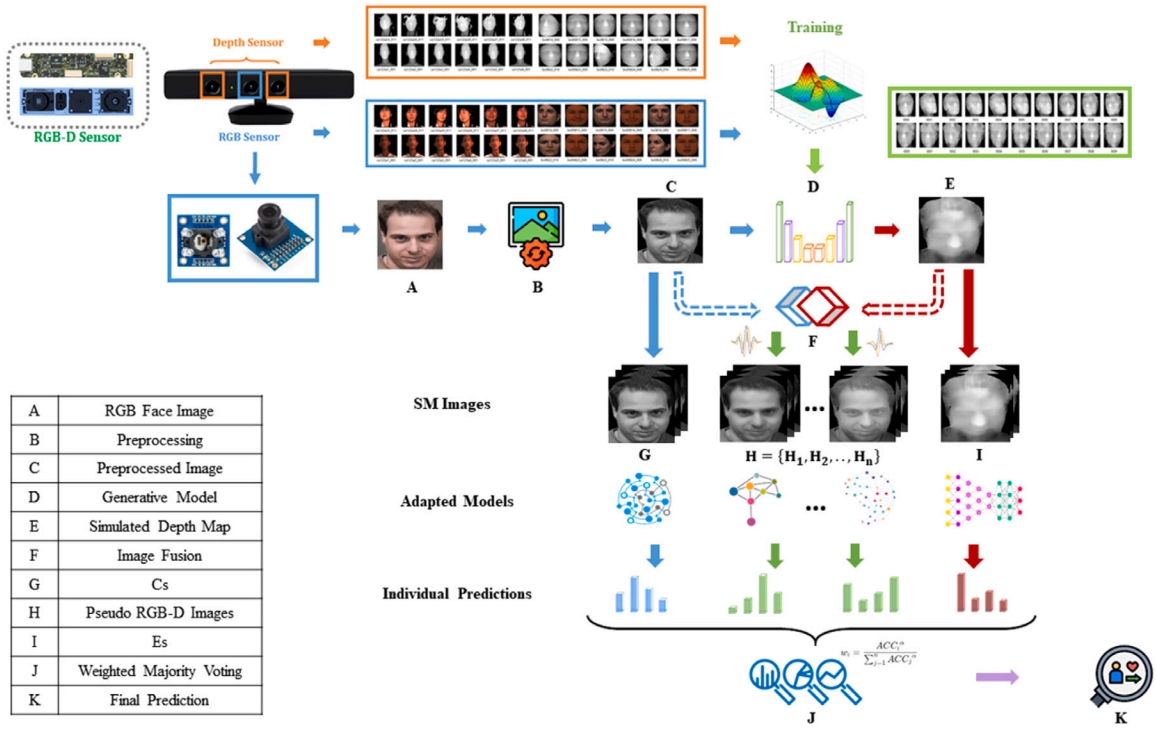
In real-world image analysis, diverse backgrounds can impede algorithm performance. To ensure uniformity with training image pairs, derived from 3D data, a process to eliminate non-facial backgrounds is applied. Our preprocessing approach employs Otsu's method (Otsu, 1975) to get a threshold for image binarization, the 8-connected labeling algorithm for facial detection, and an opening operation for image refinement. The result is an image with the background removed, focusing solely on facial features. The pseudo-code for removing the background is depicted as follows:

```
Function RemoveBg(Img): OutImg
Input: Img
Output: OutImg
Begin
 Thr ← Otsu(Img)
 BinImg ← Binarize(Img, Thr)
 LabImg ← Label(BinImg, 8-Conn)
 FaceImg ← BlackBg(LabImg)
 OutImg ← OpenOp(FaceImg)
 return OutImg
End
```

After removing the background of the image, we carry out image normalization to ensure that different features can be compared on the same scale.

3.2. Depth-generating model

Generative Adversarial Networks (GANs), originally proposed by Goodfellow et al., use random noise vectors to generate fake images through a two-component system: a generator and a discriminator (Goodfellow et al., 2014). The Conditional Generative Adversarial Network (cGAN), developed by Mirza and Osindero, further extends this model to generate images from random noise vectors based on



A	RGB Face Image
B	Preprocessing
C	Preprocessed Image
D	Generative Model
E	Simulated Depth Map
F	Image Fusion
G	Cs
H	Pseudo RGB-D Images
I	Es
J	Weighted Majority Voting
K	Final Prediction

Fig. 4. Simulated multimodal deep facial diagnosis framework.

specific conditions (Mirza & Osindero, 2014). Taking this a step further, ACGAN, proposed by Odena, Olah, and Shlens, enhances the discriminator part of the cGAN model to categorize the class of the input image (Odena et al., 2017). The Pix2Pix model by Isola et al., a specialized version of cGAN, applies 2D images as input conditions for image-to-image translation (Isola et al., 2017).

In our objective to efficiently simulate corresponding depth from RGB face images, we synthesize the aforementioned network architectures and advanced techniques, leading to the proposal of Depth Plus Generative Adversarial Network (D+GAN) (Jin et al., 2022). Distinct from prior models, the generator in D+GAN employs condition images and their associated labels to produce synthetic images. Concurrently, its discriminator not only verifies if the input is a genuine sample corresponding to the conditional image but also identifies the multiple categories to which the sample belongs. Fig. 5 illustrates the core architectures of cGAN, ACGAN, Pix2Pix, and D+GAN, succinctly highlighting the differences in the primary structures of D+GAN and other related GANs.

3.2.1. D+GAN

To fully leverage the attribute information of faces, we choose the D+GAN model as an instance of the depth generation model in the Simulated Multimodal Framework for facial diagnosis in this case.

The generator (G) of D+GAN takes as input an RGB image of dimensions 256 × 256, paired with facial attribute labels comprising gender, age, and race. It outputs a depth map of identical dimensions, thereby achieving an image-to-image mapping. The discriminator (D) assesses the depth map's quality. In the design, the discriminator receives a depth map of dimensions 256 × 256, along with its associated labels, and the output comprises four scalar values, each indicating the probability associated with authenticity, age, gender, and race. Specifically, age is categorized into three groups: under 18 years, 19–39 years, and 40 years or elder. Gender categories are divided into male and female, while racial categories are broadly divided into three groups: White, Asian, and Black.

In detail, the structure of the generator adopts a U-shaped design (Ronneberger et al., 2015), composed of an encoder that extracts features and a decoder that restores the original size of the image. Skip connections are employed between the encoder and decoder for resolving the vanishing gradient problem. The loss function for the generator, denoted as L_G, encompasses four components. It is expressed as:

$$L_G = \lambda_1 L_{S,G} + \lambda_2 L_{R,G} + \lambda_3 L_{C,G} + \lambda_4 L_{W,G} \tag{3}$$

where

$$L_{S,G} = -\mathbb{E}_{X \in P_{dat}(X)}[\log D_1(G(X), X)] \tag{4}$$

$$L_{R,G} = -\mathbb{E}_{Y \in P_{dat}(Y), X \in P_{dat}(X)}\left[\|Y - G(X)\|_2\right] \tag{5}$$

$$L_{C,G} = \sum_{i=2}^{4} \mathbb{E}_{X \in P_{dat}(X)}[\log P(D_i = c_i | G(X))] \tag{6}$$

$$L_{W,G} = \frac{1}{2}\|W\|^2 \tag{7}$$

In the aforementioned equations: X denotes the RGB facial image intended for translation, Y denotes the real depth image serving as the conditional image, P_{dat} represents the corresponding probability distribution, D_1 represents the primary discriminator's output, which determines authenticity, D_i represents the ith classifier, $G(x)$ denotes the generated image, and c_i denotes the corresponding category of the ith classifier. The objective of the loss function $L_{S,G}$ is to enable the generator to produce samples that can deceive the discriminator. The objective of the loss function $L_{R,G}$ is to ensure high similarity between the generator's output images and the condition images. The objective of the loss function $L_{C,G}$ is to ensure the detail of the generated images, allowing them to be correctly classified by the discriminator. The objective of the loss function $L_{W,G}$ which is a weight regularization term is to avoid overfitting.

The discriminator of D+GAN consists of four sub-networks. To focus on the local details of the generated images, these sub-networks are implemented using a fully convolutional network. Each sub-network takes the output from the intermediate node as input and performs its individual classification. The first sub-network is responsible for

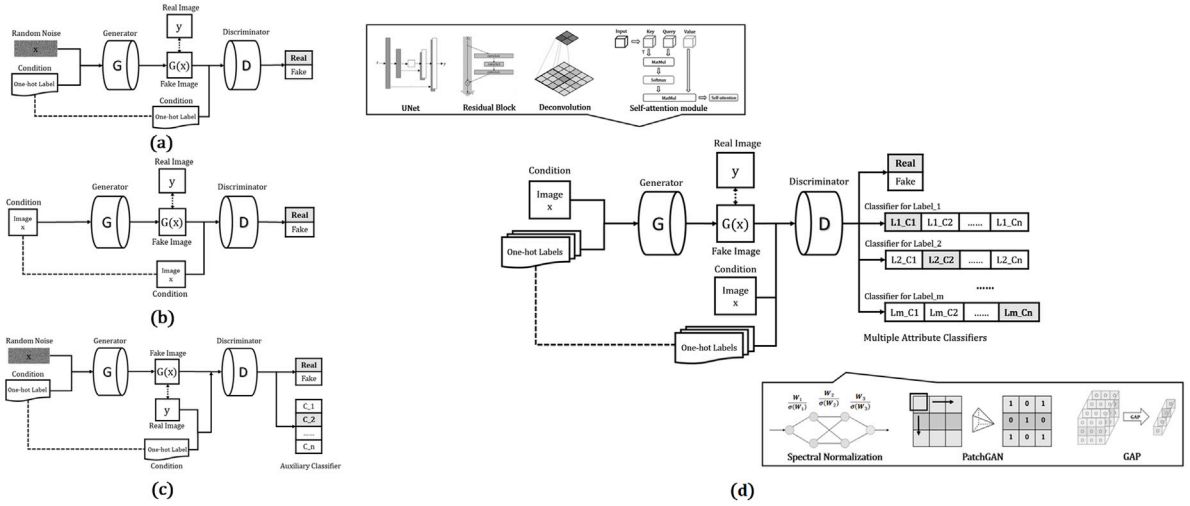


Fig. 5. Related classic GAN architectures: (a) cGAN, (b) Pix2Pix, (c) ACGAN, and (d) D+GAN.

determining if the depth map is generated, essentially functioning as a binary classifier. The remaining sub-networks are responsible for identifying facial attributes. The loss function for the discriminator, denoted as L_D, consists of two parts. It is expressed as:

$$L_D = \lambda_1 L_{S,D} + \lambda_2 L_{C,D} \tag{8}$$

where

$$\begin{aligned} L_{S,D} = {} & \mathbb{E}_{X\in P_{dat}(X), Y\in P_{dat}(Y)}\left[log\, D_1(X,Y)\right] \\ & + \mathbb{E}_{X\in P_{dat}(X)}\left[log(1 - D_1(G(X),X))\right] \end{aligned} \tag{9}$$

$$\begin{aligned} L_{C,D} = {} & \sum_{i=2}^{4} \mathbb{E}_{X\in P_{dat}(X)}[\log P(D_i = c_i|G(X))] \\ & + \mathbb{E}_{Y\in P_{dat}(Y)}[\log P(D_i = c_i|Y)] \end{aligned} \tag{10}$$

The objective of the loss function $L_{S,D}$, derived from the conventional GAN, is to differentiate between real samples and those that are generated. The objective of the loss function $L_{C,D}$, is to ensure the model correctly classifies facial attributes.

Residual block. It facilitates an easier optimization of the neural network (He et al., 2016). Within the residual block, the mapping is transformed from $F(x)$ to $F(x) + x$ through the utilization of skip connections. Instead of the original UNet design, residual blocks are employed at the junction between encoder and decoder.

Self-attention module. It allows the model to weigh and consider different parts of the input when producing an output, rather than treating all parts of the input equally (Zhang et al., 2019). Within the self-attention module, the input feature X is converted into Query (Q), Key (K), and Value (V) representations via distinct matrix multiplication operations, with their channel sizes adjusted accordingly. The module then computes attention weights between Q and K, applies them to V, resulting in a new input representation that emphasizes key features. The process can be expressed as:

$$Q = W_Q X \tag{11}$$

$$K = W_K X \tag{12}$$

$$V = W_V X \tag{13}$$

$$\text{Attention}(Q,K,V) = \text{Softmax}(QK^T)V \tag{14}$$

In the design of both the generator and discriminator, following certain higher-level convolutional layers, a self-attention module is incorporated.

Table 2
List of technologies involved.

D+GAN	Technology	Number
Generator	Deconvolution	4
	Residual Block	10
	Self-attention module	10
Discriminator	PatchGAN	4
	Spectral Normalization	10
	Self-attention module	2
	Global Average Pooling	4

Table 3
Parameter setting.

Parameter	Value
Input Image Size	256 × 256
Batch Size	4
Epochs	20
Optimizer	Adadelta (Zeiler, 2012)
Learning Rate	0.2
Learning Rate Half Life (Batches)	5000

Spectral normalization. It is aimed at controlling the Lipschitz constant of network layers in order to stabilize model training and enhance generalization performance (Miyato et al., 2018). Specifically, given a weight matrix W, the spectrally normalized version of W is computed as follows:

$$\hat{W} = \frac{W}{\sigma(W)} \tag{15}$$

where $\sigma(W)$ denotes the maximum singular value of the given W, and $\hat{W}$ is the normalized weight matrix. It is used in the discriminator in our design.

Table 2 shows the applied technologies list in D+GAN. Table 3 shows the parameters for training. Fig. 6 shows a typical loss curve for successful training of the generator, which converges before 16 epochs.

3.2.2. Dataset

In this study, colored images and their corresponding depth maps from the Bosphorus (Savran et al., 2008) and CASIA 3D Face Database (CASIA, 2004) are used to train the GAN, amounting to 9290 pairs in total. While, Binghamton University 3D Facial Expression (BU-3DFE) Database (Yin et al., 2006) was excluded from the training process. The Bosphorus 3D Face Database was captured using the Inspeck Mega Capturor II 3D, featuring sensor resolutions of 0.3 mm (x and y axes)

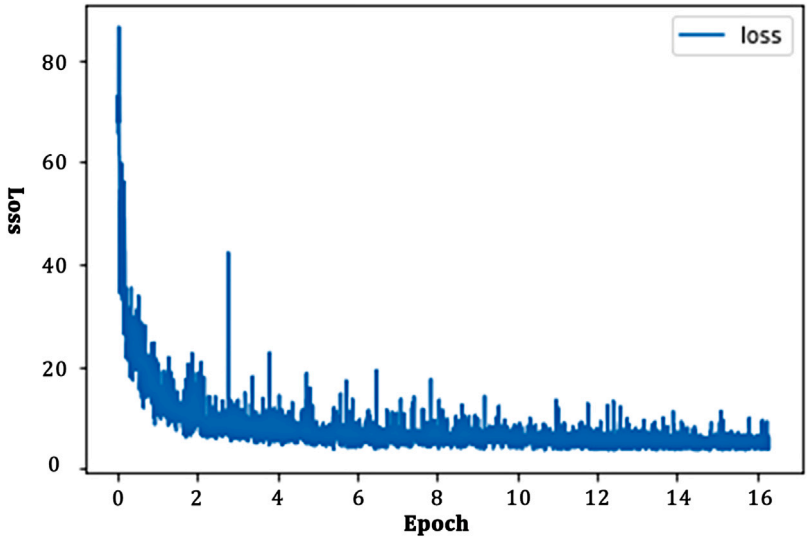


Fig. 6. Generator training loss curve in the experiment.

and 0.4 mm (z axis), along with color resolution of 1600 × 1200 pixels. The CASIA 3D Face Database was captured using the Minolta Vivid 910, featuring accuracies of ±0.22 mm in the X axis, ±0.16 mm in the Y axis, and ±0.1 mm in the Z axis to the Z reference plane, along with color data resolution of 640 × 480 pixels. The BU-3DFE database was captured using the 3DMD digitizer, featuring texture images of approximately 1300 by 900 pixels and model resolutions ranging from 20,000 to 35,000 polygons, dependent on the size of the subject's face.

Figs. 7, 8, and 9 illustrate some example RGB images along with their corresponding depth maps, which serve as the ground truth and are transformed from the aforementioned datasets. Furthermore, the simulated pseudo-depth samples and the local Structural Similarity Index Measure (SSIM) Map (Wang et al., 2004) comparing the pseudo-depth to the ground truth are also included in the aforementioned figures. The Structural Similarity Index Measure (SSIM) evaluates the perceptual quality of images and videos by considering changes in structure, luminance, and contrast, aiming to emulate the human visual system's sensitivity to these variations in image attributes. Within the local SSIM map, which pixel value ranges from 0 to 1, areas with higher SSIM values are highlighted in red, representing regions consistent with the reference image. In contrast, areas with lower SSIM values are marked in blue, indicating discrepancies from the reference image.

3.3. Image fusion

We propose a wavelet soft-thresholding-based approach for image fusion, which exhibits robustness against noise. The procedure is as follows:

First, each image that is to be fused undergoes a multilevel two-dimensional wavelet decomposition. The input for this process is the image matrix and the wavelet function to be used. In this instance, a 4-level decomposition is performed using the Symlets 4 wavelet function. The outputs are a wavelet decomposition vector and a bookkeeping matrix, which contains the number of coefficients by level and orientation.

Second, we determine a threshold value using the formula 'thr ~ sqrt(2 * log(n))', where n signifies the number of input image pixels (Donoho, 1995). With this threshold, we distinguish between soft or hard thresholding and whether the approximation coefficients should be thresholded or not for different purposes.

Lastly, we perform two-dimensional coefficient soft thresholding. This process takes as input the type of coefficients, the wavelet decomposition vector, the bookkeeping matrix, the detail levels to be thresholded, and the thresholds derived in the second step. The process determines whether soft or hard thresholding is applied.

The final result of these steps is an image that effectively combines the features of the original images, leading to a more comprehensive representation of the image characteristics.

The pseudo-code for the wavelet soft-thresholding image fusion is depicted as follows:

```
Function WST Fusion2D(Img1, Img2): FusedImage
Input: Img1, Img2
Output: FusedImage
Begin
 // Perform wavelet 2D decomposition
 C1, S1 ← WaveletDecomposition(Img1)
 C2, S2 ← WaveletDecomposition(Img2)
 // Compute threshold
 thr ← sigma * sqrt(2 * log(numel(Img)))
 // Perform soft-thresholding
 C1 ← SoftThreshold(C1, thr1)
 C2 ← SoftThreshold(C2, thr2)
 // Combine coefficients
 Cf ← CombineCoefficients(C1, C2)
 // Wavelet reconstruction
 FusedImage ← WaveletReconstruction(Cf)
 return FusedImage
End
```

3.4. Feature extractor

For feature extraction and classification, we initially fine-tune the pre-trained models of FaceNet (Schroff et al., 2015) and InsightFace (Deng et al., 2019). FaceNet in the study contains triplet loss function alongside the Inception-ResNet architecture. The Inception-ResNet combines the Inception model's capacity for capturing multi-scale image features with ResNet's residual learning framework to efficiently handle deep networks. This integration enables FaceNet to generate highly discriminative embeddings for faces, ensuring close proximity of embeddings from the same individual and distinct separation for those from different individuals. InsightFace in the study contains the ArcFace loss function alongside the IResNet network architecture. The ArcFace loss optimizes the angular distance within the feature space, significantly enhancing the accuracy and robustness of face recognition. Meanwhile, the IResNet architecture offers an efficient pathway for

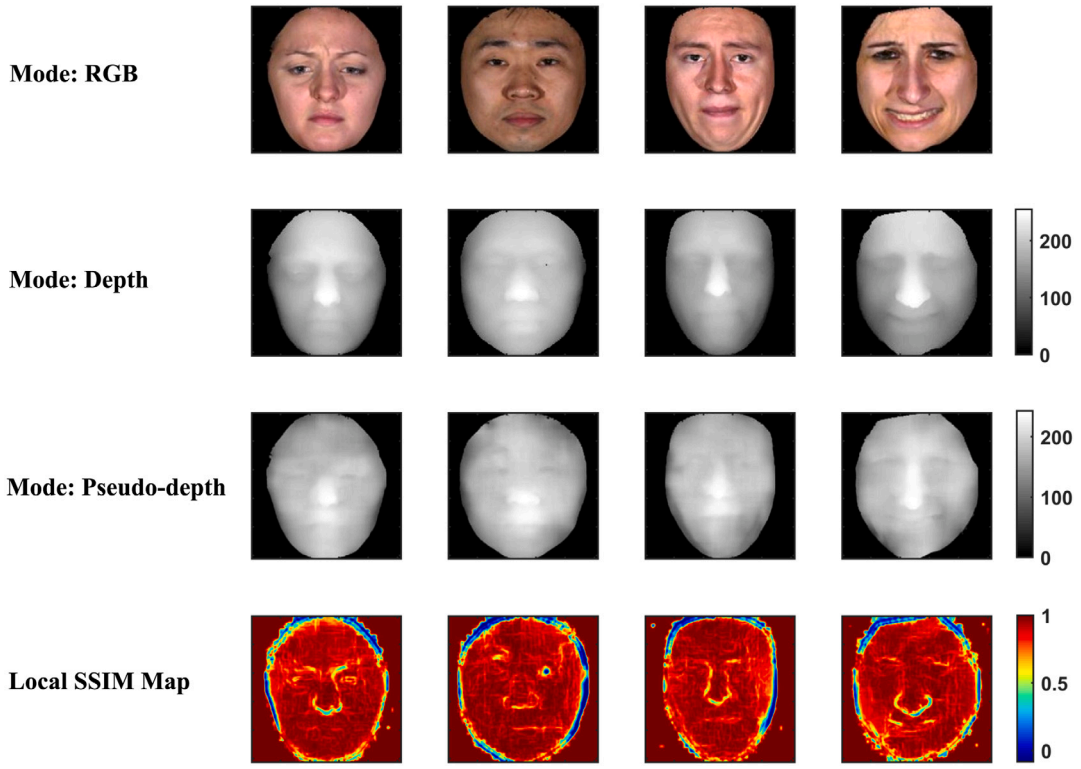


Fig. 7. Image samples and their related results from BU-3DFE Database.

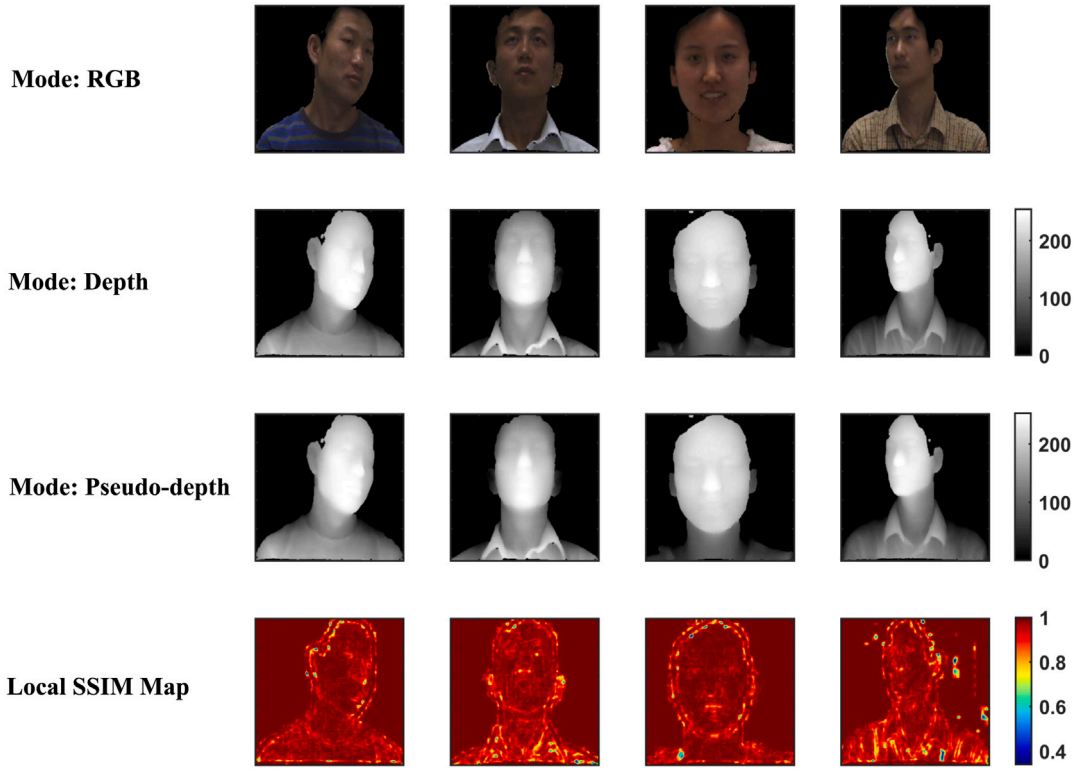


Fig. 8. Image samples and their related results from CASIA 3D Face Database.

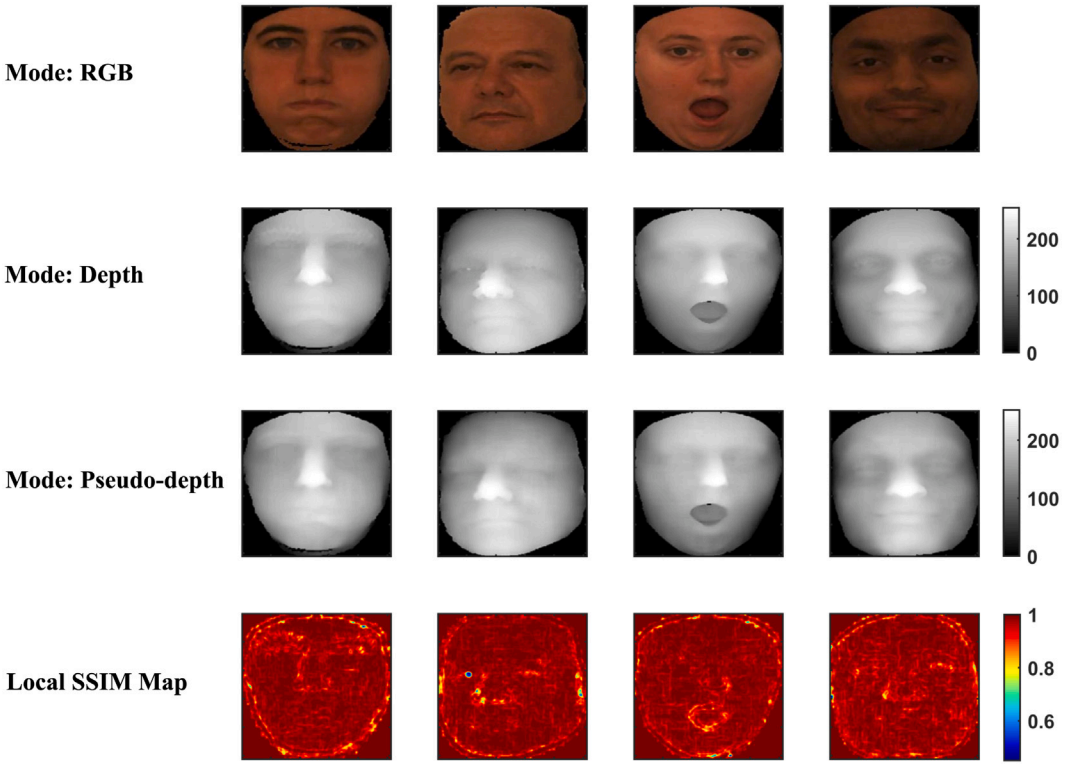


Fig. 9. Image samples and their related results from Bosphorus 3D Face Database.

information flow within deep learning models, supporting large-scale face recognition tasks.

FaceNet models are pre-trained with the CASIA-WebFace and VGG-Face2 datasets. For simplicity, FaceNet (CASIA-Webface) is denoted as A1, while FaceNet (VGG-Face2) is represented as B1 in the Figs. 13, 14, 15 . Meanwhile, The employed InsightFace model includes two structures: InsightFace: IResNet34 and InsightFace: IResNet100, both pre-trained with the MS1MV2 dataset. For simplicity, InsightFace: IResNet34 is denoted as C1, while InsightFace: IResNet100 is represented as D1.

Fine-grained classification is applicable to classification tasks characterized by substantial intra-class differences and minor inter-class differences. Inspired by the concept of fine-grained classification, we introduce a bilinear operation (Lin et al., 2015) into both InsightFace and FaceNet processing models, as illustrated in Fig. 10. The mathematical process can be represented by following equations:

$$Bi\,(l, I, u, v) = u^T(l, I)v(l, I) \tag{16}$$

where Bi represents the bilinear feature combination, l denotes location, I is the input image, and u and v are two feature functions.

$$\psi\,(I) = \sum_l Bi(l, I, u, v) \tag{17}$$

where ψ represents the feature map for the entire image.

$$x = vec(\psi(I)) \tag{18}$$

$$y = sign(x)\sqrt{|x|} \tag{19}$$

$$z = \frac{y}{\|y\|_2} \tag{20}$$

where z represents the final fused feature utilized for classification. The bilinear forms of models A1, B1, C1, D1 are represented as A2, B2, C2, D2, respectively.

3.5. Disease-specific faces 2 database

The Disease-specific Faces 2 (DSF2) database (Jin, 2023), which was released on IEEE DataPort, includes six condition-specific faces and health controls. Six conditions are acromegaly, facial nerve paralysis, Down syndrome, leprosy, thalassemia and hyperthyroidism, which is aforementioned in Chapter 1. The DSF2 dataset used in the experiment consists of condition-specific face images with diagnostic results, which were sourced from medical websites, forums, professional medical publications, and healthcare institutions. These results were further reviewed by qualified doctors to create labels for supervised learning of the model. Moreover, there is currently no evidence to suggest that the patient depicted in the photograph was suffering from two or more of the six diseases at the time the photograph was taken.

Informed consent has been obtained from the individuals in the dataset, except for those who are deceased, public figures, and others whose images have been publicly released in the media. To protect patient privacy, it is essential to de-identify condition-specific face image data. This entails removing all information that could potentially be used to identify an individual, such as names, birthdates, and medical record numbers. Furthermore, in order to protect patient privacy, the direct publication of any images from the dataset in any media or publications is not permitted.

The number of faces of each class is 85. There are a total of 595 images in the dataset. The proportions of age, gender, and ethnicity within the dataset are approximately represented in Fig. 11.

4. Experiments

In this implementation, we utilize the D+GAN to generate pseudo-depth maps, and apply two different strategies for conducting the wavelet soft-thresholding image fusion aforementioned on these pseudo-depth maps.

Strategy 1 employs the mean of wavelet coefficients from two images for both the low-frequency and high-frequency components. Applying this strategy yields Pseudo RGB-D_1 images. Strategy 2 involves

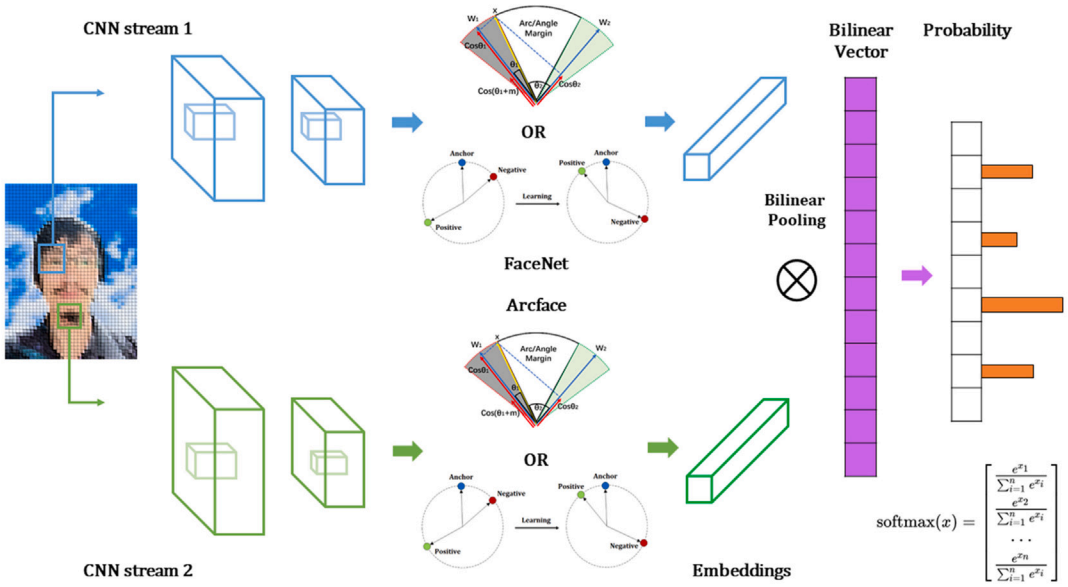


Fig. 10. Bilinear model for fine-grained facial diagnosis.

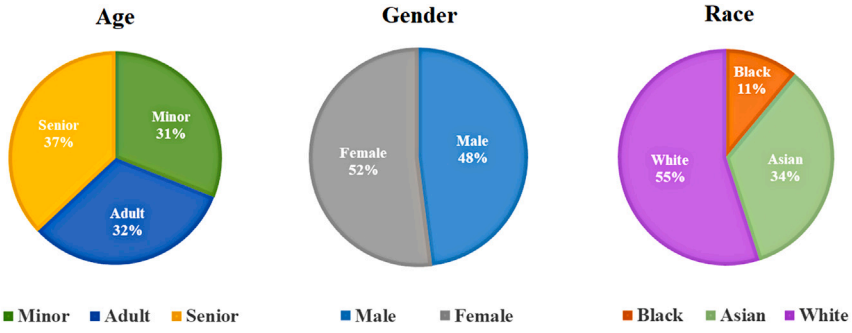


Fig. 11. The proportions of age, gender, and race of DSF2 dataset.

applying the wavelet coefficients with larger absolute values from the two images for the high-frequency components, and utilizing the mean of the wavelet coefficients from both images for the low-frequency components. Applying this strategy yields Pseudo RGB-D_2 images.

In the implementation, we have four modes of images which are RGB images, Pseudo RGB-D_1 images, Pseudo RGB-D_2 images and pseudo-depth images to perform training and predicting respectively. Four modes of image examples are displayed in Fig. 12.

In our study, we primarily focus on two advanced frameworks in the field of face recognition: InsightFace and FaceNet. By training with a low learning rate, we adapt the pre-trained model through the adoption of adaptive average pooling and a custom fully connected layer. Features are then processed via a bilinear transformation to extract rich representations, culminating in the classification of images into predefined categories using a series of dense layers. This method effectively combines deep learning and bilinear techniques, delivering superior performance in fine-grained classification tasks.

For comparison, all models were trained for 150 epochs at a low learning rate, achieving convergence of their loss functions. The final prediction results are obtained by weighted majority voting of the predictions from each model. The prediction weights assigned are positively correlated with the accuracy of the models on the training set.

Accuracy (ACC), represented as Eq. (21), is a performance metric in the context of binary classification problems that measures the proportion of true results (both true positives and true negatives) in the total dataset in the context of binary classification problem. It reflects the overall effectiveness of a model in correctly identifying both classes of an outcome, making it a straightforward indicator of a model's predictive precision and reliability.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

where TP means True Positives, TN means True Negatives, FP means False Positives, and FN means False Negatives. In practical applications, the formula for calculating multi-class accuracy can be simplified to:

$$ACC = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{22}$$

where TP_i denotes the number of true positives for class i, C represents the total number of classes, and N is the total number of samples.

For evaluation, in addition to accuracy being of significance for facial diagnosis, Matthews Correlation Coefficient is selected as an alternative indicator. The Matthews Correlation Coefficient (MCC) (Matthews, 1975) is a widely used metric for evaluating the performance of classification models, including multi-class classification tasks. It takes into account the confusion matrix to provide a comprehensive assessment of classification accuracy. The MCC ranges from −1

to 1, with −1 indicating a completely incorrect classification, 1 indicating a perfect classification, and 0 signifying a random classification. For binary classification, the MCC is calculated using the formula:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (23)$$

where TP means True Positives, TN means True Negatives, FP means False Positives, and FN means False Negatives. For multiclass classification problems, the MCC can be calculated by treating each class as binary (i.e., class i versus the rest) and averaging the MCCs for each binary problem.

For a more comprehensive assessment, three different cases are performed. Case 1 uses 45 images per category for a total of 315 images for training, and 40 images per category for a total of 280 images for testing. Case 2 uses 50 images per category for a total of 350 images for training, and 35 images per category for a total of 245 images for testing. Case 3 uses 55 images per category for a total of 385 images for training, and 30 images per category for a total of 210 images for testing.

The experimental results for the three cases are listed in Table 4, Tables 5 and 6. Fig. 13, Figs. 14 and 15 provide a clear visualization of the changes between the RGB mode and the Simulated Multimodal (SM) mode across models A1-D2. The red line, representing the SM mode, is generally positioned above the blue line that denotes the RGB mode. From the tables, it is observed that for the SM mode enhancing the RGB mode, out of the 24 experiments conducted, only one case did not show any improvement, which resulted in an effectiveness rate of 95.83%. In these 24 experiments, the ACC improved by an average of approximately 6.22%, while the MCC improved by an average of about 8.67%.

From the tables, it is observed that in terms of the improvement effect of the SM mode and the bilinear structure model on the RGB mode and non-bilinear structure models, only one out of 12 experiments showed no improvement, leading to an effectiveness rate of 91.67%. Moreover, in these 12 experiments, the ACC improved by an average of approximately 19.97%, and the MCC improved by an average of about 25.50%.

In the three experimental cases (see Fig. 16), the best-performing models demonstrate relatively high accuracy in identifying Down syndrome-specific and leprosy-specific faces, with comparatively few misclassifications. Facial paralysis-specific faces and those from the healthy control group are consistently subject to higher misclassification rates across models, likely due to significant feature expression similarities between these two categories. A commonality across the best-performing models is to confuse thalassemia-specific faces with those of the healthy control group, highlighting the challenges in differentiating between these two categories.

5. Discussion

In the field of facial diagnosis, the amount of data used in various studies varies greatly. In most cases, no more than 100 facial images are available for each disease category. Many studies do not clearly specify the number of images used for training and testing. Furthermore, the majority of the datasets are private and not publicly accessible. For binary classification tasks, the models reported in the literature generally perform well. However, for multi-classification tasks, there is a substantial discrepancy in the publicly reported recognition results, with accuracy rates ranging from 48% to 93%. Due to these factors, there are doubts surrounding the findings of many research studies, yet it is not possible to verify them. In recent years, some researchers have begun to utilize 3D DSF data in the hope of achieving more accurate results. The author believes that given the current scenario, even 2D DSF data is scarce, let alone the 3D DSF data, making it unsuitable for widespread application.

In the field of facial diagnosis, due to the scarcity of training data, we first proposed and applied transfer learning from facial recognition tasks and achieved good results (Jin et al., 2020). Facial recognition is a relatively mature research field, and many models have reached recognition accuracies of over 99% in various datasets, leaving limited room for improvement. Inspired by depth estimation (Jin et al., 2021), we utilized pseudo-depth to enhance facial recognition performance with a limited number of training images. In the expanded facial diagnosis task dataset, the recognition task is more difficult, and using only pseudo-depth does not guarantee improved results in every experiment. Similarly, studies have found that estimated depth does not always yield performance improvements in object detection (Cetinkaya et al., 2022). Therefore, we introduced the concept of fine-grained classification and employed a bilinear model structure. In combination with pseudo-depth, facial diagnosis performance is improved in most cases. However, the improvement still has a certain degree of probability. Based on this, we proposed the Simulated Multimodal Deep Facial Diagnosis, using processing models with the same structure for recognition comparison, to increase the likelihood of improvement. The improvement reflects the feature complementarity of simulated different modality features within the Simulated Multimodal Framework. The whole research results are reproducible.

5.1. Computational complexity

In comparison to the RGB computer-aided facial diagnosis, in the framework presented in this paper, the generation of synthetic facial depth images using D+GAN requires an additional computational demand of approximately 21.6G MACs during the forward propagation of the neural network. The computational complexity for both wavelet transform and inverse wavelet transform is generally $O(Nlog(N))$, where N represents the length of the input data. Thus, in the implementation described, the wavelet soft-thresholding image fusion algorithm requires an additional approximately 1.57M MACs. The computational complexity of bilinear pooling is typically $O(N^2)$. Thus, when comparing with ordinary face processing models, the use of a bilinear model for fine-grained classification requires an additional 131K MACs approximately. In summary, an additional approaching 100G of MACs is typically entirely feasible on modern computing systems and deep learning inference hardware, especially for offline processing or computations conducted in data centers. Disease screening and detection do not necessitate real-time processing. Nevertheless, for developers, it is crucial to balance the benefits of increased accuracy with the associated computational costs to ensure efficient and practical deployment across various platforms.

5.2. AI vs. human doctors

To compare the accuracy of facial diagnosis between artificial intelligence and human doctors, we engaged 15 practicing physicians from public hospitals in China to perform the prediction on the DSF2 dataset. These practicing physicians achieved an average accuracy of 59.1% on this dataset, which is at least 5% and 15% lower than the best models in RGB and SM modes, respectively. Moreover, the evaluation time taken by the physicians was invariably much longer than that of the AI models.

6. Conclusion

Deep facial diagnosis enables rapid, non-invasive disease screening and detection, which could benefit human beings and reduce the burden on the health system. In this paper, in order to leverage the estimated depth features more effectively, we propose the Simulated Multimodal Deep Facial Diagnosis. Based on facial depth estimation, our improvement introduces both early and late fusion strategies for optimized training and prediction, and by employing weighted majority

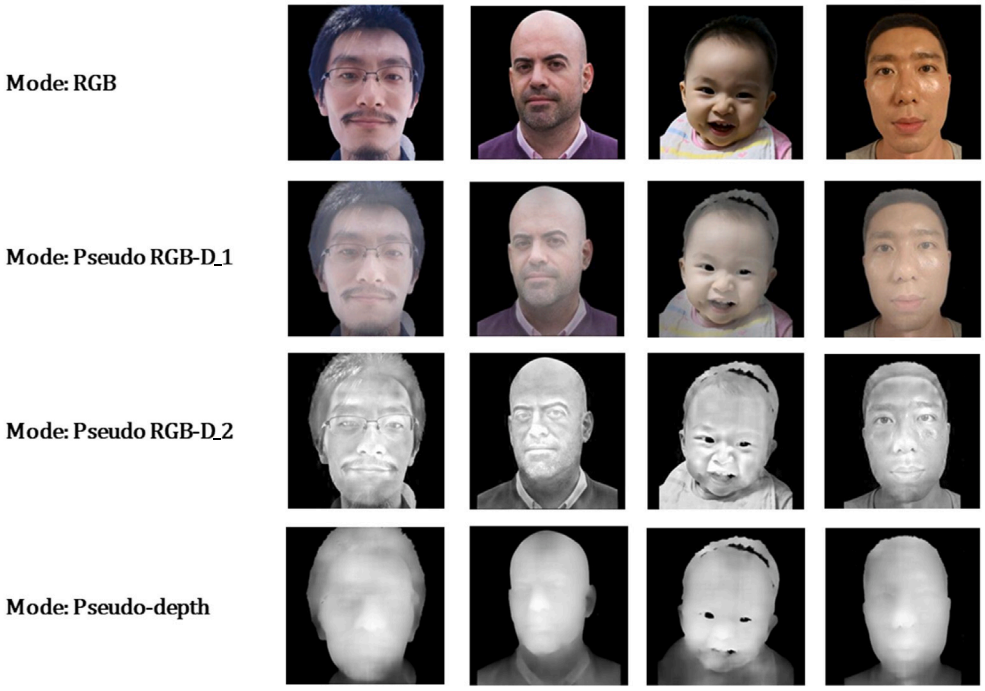


Fig. 12. Simulated multimodal image samples in DSF2 dataset.

Fig. 13. Case 1 difference area diagram (Left: ACC, Right: MCC).

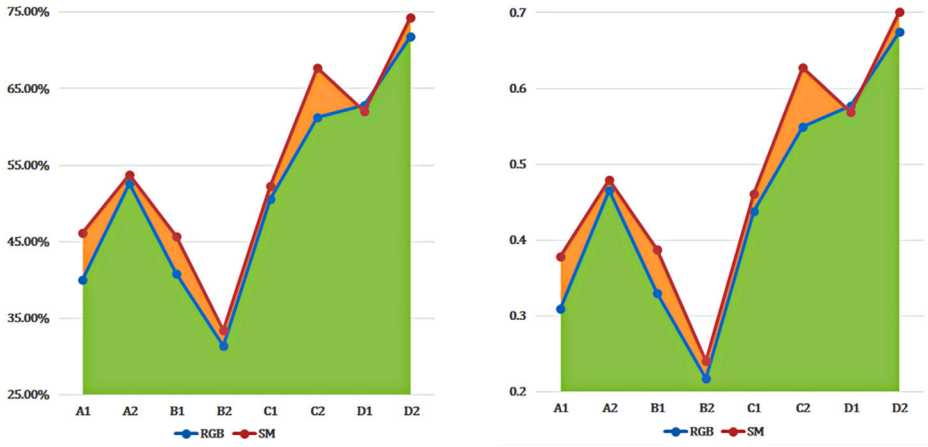


Fig. 14. Case 2 difference area diagram (Left: ACC, Right: MCC).

Table 4
Comparison results of models in Case 1.

Train-test split ratio	Model	Evaluation metrics	Mode	
			RGB	SM
315:280	FaceNet (CASIA-Webface)	ACC (%)	41.07 ± 0.002	41.43 ± 0.002
		MCC	0.352 ± 0.0004	0.355 ± 0.0003
	Bilinear FaceNet (CASIA-Webface)	ACC (%)	40.71 ± 0.005	45.36 ± 0.003
		MCC	0.323 ± 0.0005	0.374 ± 0.0001
	FaceNet (VGG-Face2)	ACC (%)	38.21 ± 0.005	39.64 ± 0.003
		MCC	0.322 ± 0.0005	0.332 ± 0.0001
	Bilinear FaceNet (VGG-Face2)	ACC (%)	45.36 ± 0.003	46.79 ± 0.005
		MCC	0.373 ± 0.0004	0.392 ± 0.0001
	InsightFace: IResNet34 (MS1MV2)	ACC (%)	55.71 ± 0.005	56.07 ± 0.002
		MCC	0.497 ± 0.0005	0.500 ± 0.0002
	Bilinear InsightFace: IResNet34 (MS1MV2)	ACC (%)	59.64 ± 0.003	65.71 ± 0.005
		MCC	0.532 ± 0.0001	0.601 ± 0.0005
	InsightFace: IResNet100 (MS1MV2)	ACC (%)	55.71 ± 0.005	60.00 ± 0.000
		MCC	0.492 ± 0.0003	0.543 ± 0.0002
	Bilinear InsightFace: IResNet100 (MS1MV2)	ACC (%)	71.07 ± 0.002	**72.14 ± 0.003**
		MCC	0.667 ± 0.0005	**0.678 ± 0.0001**

Table 5
Comparison results of models in Case 2.

Train-test split ratio	Model	Evaluation metrics	Mode	
			RGB	SM
350:245	FaceNet (CASIA-Webface)	ACC (%)	40.00 ± 0.000	46.11 ± 0.003
		MCC	0.309 ± 0.0001	0.379 ± 0.0001
	Bilinear FaceNet (CASIA-Webface)	ACC (%)	52.65 ± 0.004	53.88 ± 0.003
		MCC	0.466 ± 0.0001	0.479 ± 0.0002
	FaceNet (VGG-Face2)	ACC (%)	40.82 ± 0.004	45.71 ± 0.005
		MCC	0.331 ± 0.0004	0.387 ± 0.0004
	Bilinear FaceNet (VGG-Face2)	ACC (%)	31.43 ± 0.002	33.47 ± 0.001
		MCC	0.218 ± 0.0002	0.241 ± 0.0004
	InsightFace: IResNet34 (MS1MV2)	ACC (%)	50.61 ± 0.003	52.24 ± 0.005
		MCC	0.438 ± 0.0002	0.461 ± 0.0003
	Bilinear InsightFace: IResNet34 (MS1MV2)	ACC (%)	61.22 ± 0.005	67.76 ± 0.005
		MCC	0.550 ± 0.0001	0.627 ± 0.0005
	InsightFace: IResNet100 (MS1MV2)	ACC (%)	62.86 ± 0.003	62.04 ± 0.001
		MCC	0.578 ± 0.0005	0.569 ± 0.0004
	Bilinear InsightFace: IResNet100 (MS1MV2)	ACC (%)	71.84 ± 0.004	**74.29 ± 0.005**
		MCC	0.675 ± 0.0003	**0.701 ± 0.0001**

Table 6
Comparison results of models in Case 3.

Train-test split ratio	Model	Evaluation metrics	Mode	
			RGB	SM
385:210	FaceNet (CASIA-Webface)	ACC (%)	37.62 ± 0.001	43.81 ± 0.001
		MCC	0.298 ± 0.0005	0.372 ± 0.0003
	Bilinear FaceNet (CASIA-Webface)	ACC (%)	51.90 ± 0.001	58.57 ± 0.002
		MCC	0.459 ± 0.0004	0.525 ± 0.0005
	FaceNet (VGG-Face2)	ACC (%)	38.57 ± 0.002	41.90 ± 0.001
		MCC	0.292 ± 0.0004	0.334 ± 0.0001
	Bilinear FaceNet (VGG-Face2)	ACC (%)	41.90 ± 0.005	46.19 ± 0.001
		MCC	0.331 ± 0.0001	0.385 ± 0.0005
	InsightFace: IResNet34 (MS1MV2)	ACC (%)	58.10 ± 0.001	59.52 ± 0.003
		MCC	0.517 ± 0.0003	0.533 ± 0.0005
	Bilinear InsightFace: IResNet34 (MS1MV2)	ACC (%)	60.95 ± 0.003	64.76 ± 0.002
		MCC	0.546 ± 0.0004	0.591 ± 0.0005
	InsightFace: IResNet100 (MS1MV2)	ACC (%)	71.90 ± 0.005	72.38 ± 0.001
		MCC	0.675 ± 0.0004	0.681 ± 0.0002
	Bilinear InsightFace: IResNet100 (MS1MV2)	ACC (%)	74.29 ± 0.005	**74.76 ± 0.002**
		MCC	0.702 ± 0.0003	**0.706 ± 0.0001**

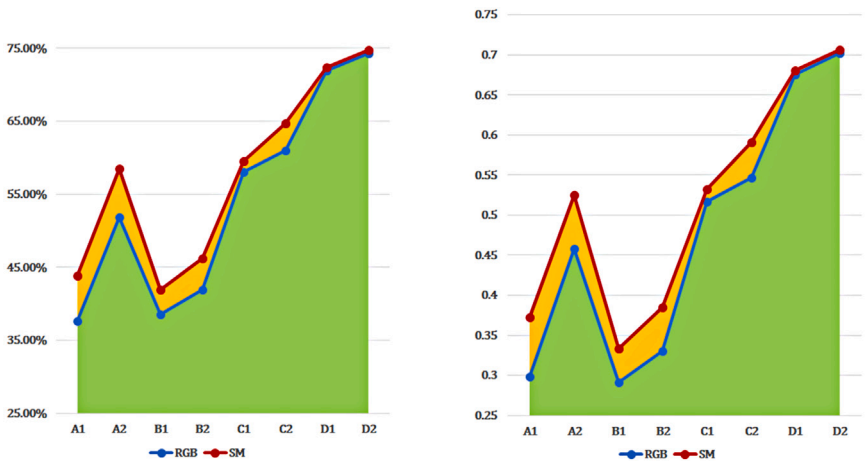


Fig. 15. Case 3 difference area diagram (Left: ACC, Right: MCC).

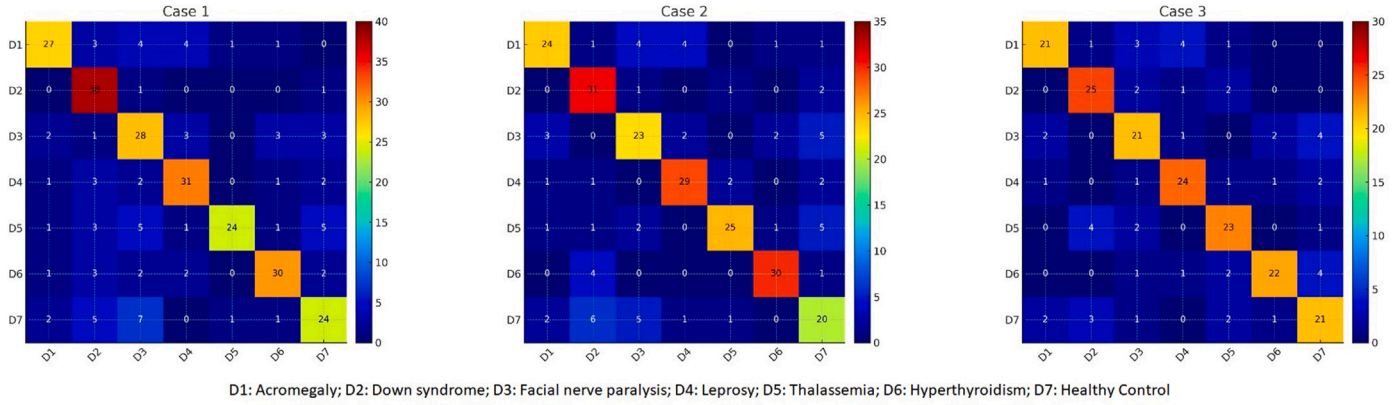


Fig. 16. Confusion Matrix Examples for Case Studies.

voting, we ultimately achieve promising results. Under this framework, we have retrained advanced pre-trained face recognition models using bilinear operations to adapt them for facial diagnosis tasks, a critical advantage in a context where training data is typically limited. Experimental results show that this approach significantly improves the performance of RGB facial diagnosis with a high probability.

In future work, we plan to collect more real-world data for training and testing facial diagnosis models and prepare the necessary software and hardware for practical applications in society. While the primary focus of this work was computer-aided facial diagnosis, we see potential for the Simulated Multimodal Framework in other fields, including autonomous driving and target tracking. Our future work will involve cross-disciplinary collaboration to explore these potential applications.

CRediT authorship contribution statement

Bo Jin: Conceptualization, Methodology, Supervision, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Nuno Gonçalves:** Data curation, Writing – review & editing, Validation, Visualization. **Leandro Cruz:** Visualization, Investigation, Data curation. **Iurii Medvedev:** Validation, Software, Formal analysis. **Yuanyu Yu:** Resources, Data curation, Visualization. **Jiujiang Wang:** Resources, Data curation, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the Fundação para a Ciência e a Tecnologia (FCT) under the Project UIDB/00048/2020. We extend our gratitude to Jiangsu Second Chinese Medicine Hospital, Zhongda Hospital affiliated to Southeast University, and Xiangya Hospital of Central South University, for their support in this study.

References

Alhaija, E. S. A., Hattab, F. N., & Al-Omari, M. A. (2002). Cephalometric measurements and facial deformities in subjects with β-thalassaemia major. *The European Journal of Orthodontics*, *24*(1), 9–19.

Bannister, J. J., Wilms, M., Aponte, J. D., Katz, D. C., Klein, O. D., Bernier, F. P. J., Spritz, R. A., Hallgrímsson, B., & Forkert, N. D. (2022). A deep invertible 3-D facial shape model for interpretable genetic syndrome diagnosis. *IEEE Journal of Biomedical and Health Informatics*, *26*(7), 3229–3239. http://dx.doi.org/10.1109/JBHI.2022.3164848.

Boehringer, S., Vollmar, T., Tasse, C., Wurtz, R. P., Gillessen-Kaesbach, G., Horsthemke, B., & Wieczorek, D. (2006). Syndrome identification based on 2D analysis software. *European Journal of Human Genetics*, *14*(10), 1082–1089. http://dx.doi.org/10.1038/sj.ejhg.5201673.

Canfield, M. A., Ramadhani, T. A., Yuskiv, N., & Davidoff, M. J. (2006). Improved national prevalence estimates for 18 selected major birth defects-United States, 1999–2001. *JAMA : The Journal of the American Medical Association*, [ISSN: 0098-7484] *295*(6), 618–620. http://dx.doi.org/10.1001/jama.295.6.618.

CASIA (2004). CASIA-3D face V1. http://biometrics.idealtest.org/.

Cetinkaya, B., Kalkan, S., & Akbas, E. (2022). Does depth estimation help object detection? *Image and Vision Computing*, *122*, Article 104427. http://dx.doi.org/10.1016/j.imavis.2022.104427.

Coulson, S. E., O'Dwyer, N. J., Adams, R. D., & Croxson, G. R. (2004). Expression of emotion and quality of life after facial nerve paralysis. *Otology & Neurotology*, *25*(6), 1014–1019. http://dx.doi.org/10.1097/00129492-200411000-00026.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690–4699). http://dx.doi.org/10.1109/CVPR.2019.00482.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory*, *41*(3), 613–627. http://dx.doi.org/10.1109/18.382009.

Fanghänel, J., Gedrange, T., & Proff, P. (2006). The face-physiognomic expressiveness and human identity. *Annals of Anatomy-Anatomischer Anzeiger*, *188*(3), 261–266. http://dx.doi.org/10.1016/j.aanat.2005.11.013.

Freeman, S. B., Allen, E. G., Oxford-Wright, C. L., Tinker, S. W., Druschel, C., Hobbs, C. A., O'Leary, L. A., Romitti, P. A., Royle, M. H., & Torfs, C. P. (2007). The national down syndrome project: Design and implementation. *Public Health Reports*, *122*(1), 62–72. http://dx.doi.org/10.1177/003335490712200109.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems: vol. 27*.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P. M., Kamphausen, S. B., & Zenker, M. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, *25*(1), 60–64. http://dx.doi.org/10.1038/s41591-018-0279-0.

Hallgrímsson, B., Aponte, J. D., Katz, D. C., Bannister, J. J., Riccardi, S. L., Mahasuwan, N., McInnes, B. L., Ferrara, T. M., Lipman, D. M., & Neves, A. B. (2020). Automated syndrome diagnosis by three-dimensional facial imaging. *Genetics in Medicine*, *22*(10), 1682–1693. http://dx.doi.org/10.1038/s41436-020-0845-y.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Ho, K. K. (2011). *Growth hormone related diseases and therapy*. Springer, http://dx.doi.org/10.1007/978-1-60761-317-6.

Hoffman, R., Benz, E. J., Silberstein, L. E., Heslop, H. E., Weitz, J. I., Anastasi, J., Salama, M. E., & Abutalib, S. A. (2018). Hematology (seventh edition), (7th ed). (pp. 546–570.e10). Elsevier, ISBN: 978-0-323-35762-3, http://dx.doi.org/10.1016/B978-0-323-35762-3.00040-8.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134). http://dx.doi.org/10.1109/CVPR.2017.632.

Jiang, Z., Jin, B., & Song, Y. (2023). A novel pet trajectory prediction method for intelligent plant cultivation robot. *IEEE Sensors Letters*, *7*(2), 1–4. http://dx.doi.org/10.1109/LSENS.2023.3238468.

Jin, B. (2023). Disease-specific faces 2. http://dx.doi.org/10.21227/zqra-nh98, IEEE Dataport.

Jin, B., Cruz, L., & Gonçalves, N. (2020). Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis. *IEEE Access*, *8*, 123649–123661. http://dx.doi.org/10.1109/ACCESS.2020.3005687.

Jin, B., Cruz, L., & Gonçalves, N. (2021). Face depth prediction by the scene depth. In *2021 IEEE/ACIS 19th international conference on computer and information science* (pp. 42–48). IEEE, http://dx.doi.org/10.1109/ICIS51600.2021.9516598.

Jin, B., Cruz, L., & Gonçalves, N. (2022). Pseudo RGB-D face recognition. *IEEE Sensors Journal*, *22*(22), 21780–21794. http://dx.doi.org/10.1109/JSEN.2022.3197235.

Kim, H.-S., Jung, J., Dong, S. H., Kim, S. H., Jung, S. Y., & Yeo, S. G. (2019). Association between high neutrophil to lymphocyte ratio and delayed recovery from bell's palsy. *Clin Exp Otorhinolaryngol*, *12*(3), 261–266. http://dx.doi.org/10.21053/ceo.2018.01018.

Kong, X., Gong, S., Su, L., Howard, N., & Kong, Y. (2018). Automatic detection of acromegaly from facial photographs using machine learning methods. *EBioMedicine*, *27*, 94–102. http://dx.doi.org/10.1016/j.ebiom.2017.12.015.

Lavrentaki, A., Paluzzi, A., Wass, J. A., & Karavitaki, N. (2017). Epidemiology of acromegaly: Review of population studies. *Pituitary*, *20*, 4–9. http://dx.doi.org/10.1007/s11102-016-0754-x.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Lee, J.-S., Rhee, T.-M., Jeon, K., Cho, Y., Lee, S.-W., Han, K.-D., Seong, M.-W., Park, S.-S., & Lee, Y. K. (2022). Epidemiologic trends of thalassemia, 2006–2018: A nationwide population-based study. *Journal of Clinical Medicine*, *11*(9), 2289. http://dx.doi.org/10.3390/jcm11092289.

Li, Q., Ma, L., Jiang, Z., Li, M., & Jin, B. (2023). TECMH: Transformer-based cross-modal hashing for fine-grained image-text retrieval. *Computers, Materials & Continua*, [ISSN: 1546-2226] *75*(2), 3713–3728. http://dx.doi.org/10.32604/cmc.2023.037463, http://www.techscience.com/cmc/v75n2/52124.

Lin, T.-Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 1449–1457). http://dx.doi.org/10.1109/ICCV.2015.170.

Manifold, A. (2005). Hyperthyroidism, thyroid storm, and graves' disease. *E-medicine*, *4*, 1–18.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451. http://dx.doi.org/10.1016/0005-2795(75)90109-9.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

Modell, B., & Darlison, M. (2008). Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization*, *86*(6), 480–487. http://dx.doi.org/10.2471/blt.06.036673.

Muñoz-Ortiz, J., Sierra-Cote, M. C., Zapata-Bravo, E., Valenzuela-Vallejo, L., Marin-Noriega, M. A., Uribe-Reina, P., Terreros-Dorado, J. P., Gómez-Suarez, M., Arteaga-Rivera, K., & De-La-Torre, A. (2020). Prevalence of hyperthyroidism, hypothyroidism, and euthyroidism in thyroid eye disease: A systematic review of the literature. *Systematic Reviews*, *9*(1), 1–12. http://dx.doi.org/10.1186/s13643-020-01459-7.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning* (pp. 2642–2651). PMLR, https://proceedings.mlr.press/v70/odena17a.html.

Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, *11*(285-296), 23–27. http://dx.doi.org/10.1109/TSMC.1979.4310076.

Porras, A. R., Rosenbaum, K., Tor-Diez, C., Summar, M., & Linguraru, M. G. (2021). Development and evaluation of a machine learning-based point-of-care screening tool for genetic syndromes in children: a multinational retrospective study. *The Lancet Digital Health*, *3*(10), e635–e643. http://dx.doi.org/10.1016/S2589-7500(21)00137-0.

Rodrigues, M., Nunes, J., Figueiredo, S., Martins de Campos, A., & Geraldo, A. F. (2019). Neuroimaging assessment in down syndrome: A pictorial review. *Insights into Imaging*, *10*, 1–13. http://dx.doi.org/10.1186/s13244-019-0729-3.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234–241). Springer, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., & Akarun, L. (2008). Bosphorus database for 3D face analysis. In *European workshop on biometrics and identity management* (pp. 47–56). Springer, http://dx.doi.org/10.1007/978-3-540-89991-4_6.

Schneider, H. J., Kosilek, R. P., Günther, M., Roemmler, J., Stalla, G. K., Sievers, C., Reincke, M., Schopohl, J., & Würtz, R. P. (2011). A novel approach to the detection of acromegaly: accuracy of diagnosis by automatic face classification. *The Journal of Clinical Endocrinology & Metabolism*, *96*(7), 2074–2080. http://dx.doi.org/10.1210/jc.2011-0237.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823). http://dx.doi.org/10.1109/CVPR.2015.7298682.

Shirmohammadi, S., & Ferrero, A. (2014). Camera as the instrument: The rising trend of vision based measurement. *IEEE Instrumentation & Measurement Magazine*, *17*(3), 41–47. http://dx.doi.org/10.1109/MIM.2014.6825388.

Shukla, P., Gupta, T., Saini, A., Singh, P., & Balasubramanian, R. (2017). A deep learning frame-work for recognizing developmental disorders. In *2017 IEEE winter conference on applications of computer vision* (pp. 705–714). http://dx.doi.org/10.1109/WACV.2017.84.

Tiemstra, J. D., & Khatkhate, N. (2007). Bell's palsy: Diagnosis and management. *American Family Physician*, *76*(7), 997–1002, PMID: 17956069.

Umeda-Kameyama, Y., Kameyama, M., Tanaka, T., Son, B.-K., Kojima, T., Fukasawa, M., Iizuka, T., Ogawa, S., Iijima, K., & Akishita, M. (2021). Screening of Alzheimer's disease by facial complexion using artificial intelligence. *Aging (Albany NY)*, *13*(2), 1765. http://dx.doi.org/10.18632/aging.202545.

Unschuld, P. U. (2003). *Huang Di Nei Jing Su Wen: Nature, knowledge, imagery in an ancient Chinese medical text: With an appendix: The doctrine of the five periods and six Qi in the Huang Di Nei Jing Su Wen*. Univ of California Press, http://www.jstor.org/stable/10.1525/j.ctt1ppf4w.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. http://dx.doi.org/10.1109/TIP.2003.819861.

World Health Organization (2014). Global leprosy update, 2013; reducing disease burden. *Weekly Epidemiological Record=Relevé épidémiologique hebdomadaire*, *89*(36), 389–400, https://iris.who.int/handle/10665/242259.

World Health Organization (2020). Global leprosy (hansen disease) update, 2019: time to step-up prevention initiatives. *The Weekly Epidemiological Record*, *95*(36), 417–440.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition* (pp. 211–216). IEEE, http://dx.doi.org/10.1109/FGR.2006.6.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Zhang, X., Chen, F., Wang, C., Tao, M., & Jiang, G.-P. (2020). SiENet: Siamese expansion network for image extrapolation. *IEEE Signal Processing Letters*, *27*, 1590–1594. http://dx.doi.org/10.1109/LSP.2020.3019705.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354–7363). PMLR, https://proceedings.mlr.press/v97/zhang19d.html.

Zhao, Q., Okada, K., Rosenbaum, K., Kehoe, L., Zand, D. J., Sze, R., Summar, M., & Linguraru, M. G. (2014). Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ICA. *Medical Image Analysis*, *18*(5), 699–710. http://dx.doi.org/10.1016/j.media.2014.04.002.

Zhao, Q., Werghi, N., Okada, K., Rosenbaum, K., Summar, M., & Linguraru, M. G. (2014). Ensemble learning for the detection of facial dysmorphology. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society* (pp. 754–757). IEEE, http://dx.doi.org/10.1109/EMBC.2014.6943700.