ChromosomeNet : A massive dataset enabling benchmarking and building basedlines of clinical chromosome classification

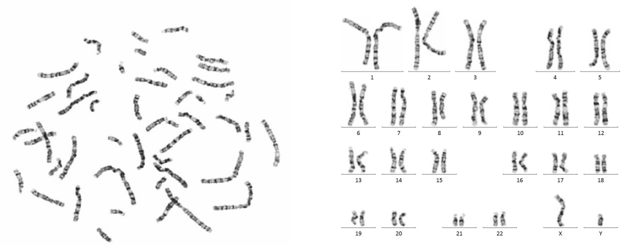
Chengchuang Lin*, Hanbiao Chen*, Jiesheng Huang, Jing Peng, Li Guo, Zhirong Yang, Jiahua Du, Shuangyin Li⋆, Aihua Yin⋆, Gansen Zhao⋆,

Abstract—Chromosome karyotyping analysis is a vital cytogenetics technique for diagnosing genetic and congenital malformations, analyzing gestational and implantation failures, etc. Since the chromosome classification as an essential stage in chromosome karyotype analysis is a highly time-consuming, tedious, and error-prone task, which requires a large amount of manual work of experienced cytogenetics experts. Many deep learning-based methods have been proposed to address the chromosome classification issues. However, two challenges still remain in current chromosome classification methods. First, most existing methods were developed by different private datasets, making these methods difficult to compare with each other on the same base. Second, due to the absence of reproducing details of most existing methods, these methods are difficult to be applied in clinical chromosome classification applications widely. To address the above challenges in the chromosome classification issue, this work builds and publishes a massive clinical dataset. This dataset enables the benchmarking and building chromosome classification baselines suitable for different scenarios. The massive clinical dataset consists of 126,453 privacy preserving G-band chromosome instances from 2,763 karyotypes of 408 individuals. To our best knowledge, it is the first work to collect, annotate, and release a publicly available clinical chromosome classification dataset whose data size scale is also over 120,000. Meanwhile, the experimental results show that the proposed dataset can boost performance of existing chromosome classification models at a varied range of degrees, with the highest accuracy improvement by 5.39 percentage points. Moreover, the best baseline with 99.33% accuracy reports state-of-the-art classification performance. The clinical dataset and state-of-the-art baselines can be found at https://github.com/CloudDataLab/BenchmarkForChromosomeClassification.

Index Terms—Chromosome Classification, Chromosome Karyotyping Analysis, Biomedical Image Processing, Clinical Dataset, Benchmark and Baselines, Deep Learning, Artificial Intelligence

1 Introduction

Chromosomal analysis provides valuable information on human chromosome abnormities and relationships of corresponding congenital genetic diseases. Chromosomal anomalies result in numerous genetic diseases and are responsible for gestational losses, implantation failures, and congenital malformations [1]. The chromosome karyotyping analysis, a most common and essential approach in chromosomal analysis, refers to the operations of segmenting chromosome instances from a given cell image and arranging these instances into the corresponding karyotype according to *International System for Human Cytogenomic Nomenclature* (ISCN) [2] criteria [3]. *Fig.1* presents an example of chromosome karyotyping analysis. The chromosome classification (one of the most vital tasks in chromosome karyotype analysis) is a highly time-consuming specialized task, dependent on the experience of chromosomal analysts and influenced by factors such as fatigue and decreased attention [4].



(a)

(b)

Fig. 1: An example of chromosome karyotyping analysis. The sub-figure (a) illustrates a G-band chromosome stained cell microphotograph. The sub-figure (b) depicts the corresponding chromosome karyotype.

Prior efforts [5–10] have contributed a lot to the task of chromosome classification. However, two challenges still remain in the chromosome classification task and its clinical applications. First, the above methods were built and evaluated on different private datasets with different sample volumes of different qualities, making these methods hard

- *C. Lin, J. Huang, G. Zhao, J. Peng, S. Li are with South China Normal University and Key Lab on Cloud Security and Assessment technology of Guangzhou, Guangzhou, China, 510631.*
- *A. Yin, H. Chen, and L. Guo are with Guangdong Women and Children Hospital, Guangzhou, China, 511400.*
- *J. Du is with Department of Information and General Technology, The Affiliated High School of South China Normal University, Guangzhou 510631, China.*
- *Z. Yang is with Norwegian University of Science and Technology, Trondheim 17491, Norway.*
- ** C. Lin and H. Chen contributed equally.*
- *⋆ S. Li, A. Yin and G. Zhao are corresponding authors.*

to compare fairly. Moreover, most of these private datasets are not available for peers or clinical applications. This leads to the difficulty of choosing a better method among the available methods. Second, most existing methods have not be presented with enough implementation details or executable pre-trained models in their papers. Even though the authors have claimed that their methods are good enough for their corresponding private datasets, peers and clinical chromosome analysts have difficulty in reproducing their methods or reported performance.

Solid experiments of deep learning-based medical image analysis methods require massive annotated data to prove the usefulness and effectiveness of proposed methods [11, 12]. Accordingly, the motivations of this work include three parts. The first one is to build a fair and objective performance evaluation system for chromosome classification models. The second one is to provide a large-scale clinical dataset, which facilitates peers to develop novel chromosome classification models or verify existing models. The last but not least motivation is to offer chromosome classification baselines for peers to compare their developing models or apply in clinical applications.

The constructed clinical dataset contains 126,453 privacy-removal G-band chromosome instances segmented from 2,763 chromosome karyotypes of 225 individuals, allowing different deep learning-based methods to be tested and validated on the same dataset. This work also provides different chromosome classification baselines to meet scenarios with different requirements for efficiencies and accuracies. The highest performance baseline achieves 99.33% classification $accuracy$, 99.32% precision, 99.29% F_1, 99.27% $sensitivity$, and 99.97% $specificity$ on the proposed clinical dataset, while the lightest baseline requires only 6.58 milliseconds to infer a sample. The highest performance baseline with 99.33% accuracy achieves a state-of-the-art result according to our best knowledge. Meanwhile, extensive experimental results have suggested that the proposed dataset can boost all reproducible deep learning-based chromosome classification methods, where the most significant classification accuracy improvement achieves 5.39 percentage points.

The rest of this paper is structured as follows. *Section 2* reviews the existing works on chromosome classification tasks and dataset advancements of medical image analysis based on deep learning techniques. *Section 3* introduces the proposed clinical chromosome classification benchmark dataset. *Section 4* describes various clinical chromosome classification baselines. *Section 5* reports the evaluations and discuss the experimental results. Finally, *Section 6* concludes this work.

2 Related Work

This section reviews existing deep learning-based methods for the task of clinical chromosome classification and various benchmarks or datasets for different medical image analysis scenes.

2.1 Methods for Chromosome classification

Inspired by the advance of Siamese Networks [13] in the image classification task, [14] introduced Siamese Networks for the chromosome classification task. This method achieved 84.5% classification accuracy on a private dataset consisting of 1,720 samples (including training and testing samples). [5] presented a method using crowdsourcing, preparation, and deep learning techniques to classify chromosomes. It claims an accuracy score of 86.7% on their private dataset with 9,600 chromosome images. [15] presented a two-stage chromosome classification method named Competitive SVM Teams (*CSVMTs*). Their results yielded 91.00% classification accuracy on a database consisting of 4,400 samples. [6] proposed a method based on attention sequence learning of chromosome bands (*Res-CRANN*) for the Q-band chromosome classification task. The experimental results evaluated on the BioImLab database achieved 91.94% classification accuracy. [7] proposed a specific CNNs (Convolutional Neural Networks) model to classify chromosomes automatically. Their proposed method was trained and tested on a private dataset containing 10,304 chromosome images and achieved an accuracy of 92.5%. [16] presented a CNNs (with six convolutional layers, three pooling layers, four dropout layers, and two fully connected layers) for the chromosome classification task. This classifier hits an accuracy of score 93.79% on a database containing 4,184 images. Its dataset is available on GitHub. [8] designed a Varifocal-Net to address the chromosome classification issue whose method includes a global-scale network (*G-Net*) and a local-scale network (*L-Net*). Evaluation results from 1,909 karyograms showed that their proposed Varifocal-Net achieved the highest accuracy per patient case of 98.9% accuracy for both type and polarity tasks. [17] provided two varieties of CNNs named *ChromNet1* and *ChromNet2* for the chromosome classification task. According to their evaluation results on the private dataset with 21,423 samples, *ChromNet2* achieved 91.3% classification accuracy. The experimental result of *ChromNet1* was not reported in their article. Inspired by the achievement of the Inception-ResNet [18] in ImageNet image classification challenge competition, a chromosome classification method named CIR-Net [9] was proposed for the chromosome classification task. CIR-Net obtained 95.98% classification accuracy on a public dataset consisting of 2,990 samples. Recently, a mixed classification classifier named MixedNet [3] was proposed based on prior method [19]. Unlike other methods, MixedNet is a 25-classes classifier that identified a given image sample into a corresponding category if it is a chromosome instance. Otherwise, it identifies the sample into a specific category indicating a chromosome cluster. The authors evaluated their proposed method on a private dataset with 10,856 samples and reported 99.53% accuracy. Moreover, [10] detect and identify chromosome instances from cell images using the DeepAcc model whose chromosome classification accuracy is 84.85% on a collected dataset with 3,390 giemsa-stained metaphase images.

According to the above researches on the issue of chromosome classification, two following challenges can be concluded. First of all, none of these methods were developed and evaluated on the same dataset. Most of these methods [3, 5, 7, 8, 14–17] are evaluated on private datasets that are not available to the public. [6] developed and evaluated their method on the BioImLab dataset that is a Q-band chromosome dataset. However, in clinical chromosome karyotyping analysis, G-band chromosomes are the most commonly used. Although methods [9, 16] have released their datasets, the volume of these datasets is far from enough to corroborate the usefulness and effectiveness of deep learning-based models. Secondly, the implementation and training of a deep learning-based model are critical to the performance of the given model. Even a hyperparameter change in the tuning process may bring a significant performance gap. Unfortunately, most of the above methods did not present complete implementation and training details of their proposed models, making it difficult for peers to reproduce their reported performance.

2.2 Datasets for Medical Image Analysis

Many medical benchmarks have been built to address various clinical issues. In polyp segmentation, [20] presented a polyp segmentation benchmark named Kvasir-SEG, an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated by a medical doctor then verified by an experienced gastroenterologist. Based on this dataset, a number of algorithms [21–24] have been developed, tested and applied in clinical polyp segmentation scenarios that has continuously improved polyp segmentation accuracy from 0.8180 to 0.898 mean dice. In lesion segmentation, [25] proposed a large, open-source dataset of stroke anatomical brain images and manual lesion segmentations (ATLAS) that attracts a surge of algorithms and deep learning-based models [26–30] to address the issue of segmenting out lesions. In less than 3 years, these algorithms have pushed the new SOTA from 0.3938 to 0.6331 segmentation precision.

Other benchmarks for different medical image analysis include mortality prediction based on ECG signal processing [31], retinal OCT disease classification [32], heart rate estimation [33], and chromosome instance segmentation dataset [34], etc. However, for clinical chromosome classification applications, there is still lacking a qualified peer-available benchmark with massive scale data for development, testing and comparison of different algorithms.

3 Dataset

3.1 Characteristics of the Proposed Dataset

According to the analysis of different existing medical datasets [20, 25, 31–33], a qualified chromosome classification dataset expects the following characteristics.

- Authenticity: All samples in the target dataset should come from clinical practices, and their amount distribution of all chromosome categories in the dataset is consistent with the clinical practices.
- Objectivity: The target dataset should be able to provide an objective evaluation base for existing chromosome classification algorithms.
- Effectiveness: Existing deep learning-based image classification models should be able to learn their convergent chromosome recognition abilities through the target dataset as baselines for chromosome classification tasks.
- Availability: The target dataset should be obtained by peers to verify existing different chromosome classification algorithms and develop more advanced algorithms.

To ensure the scientificity of the constructed dataset, the construction procedure of the proposed dataset includes four major stages: collection of stained cell images, selection, chromosome karyotyping analysis, and extraction of chromosome instances. Fig. 2 illustrates the overview procedure of the proposed dataset. The summary of the proposed dataset is as follows. The proposed clinical chromosome classification benchmark dataset contains 24 categories of G-band chromosome instances, including 22 categories of autosomes labeled to 0 to 21, and X, Y sex chromosome labeled to 22 and 23. All chromosome instances of the same category are organized in the same folder named by the given category. Accordingly, our proposed dataset has 24 folders. All chromosome instance images in our proposed dataset have a uniform resolution of 300 × 300. There are a total number of 126,453 chromosome instances in the proposed dataset, and Fig. 3 summarizes the total sample amount of each category.

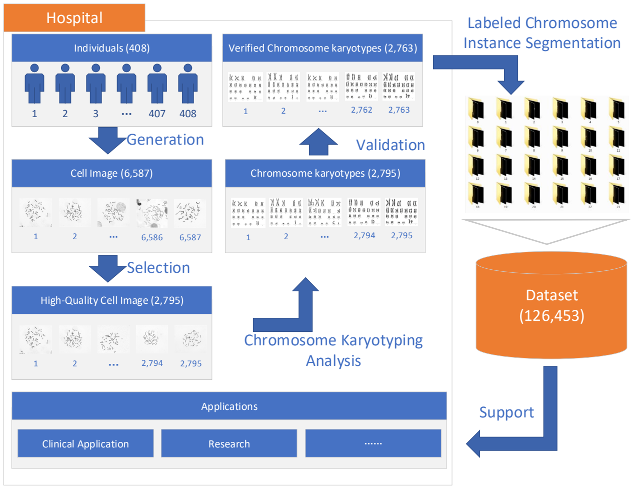


Fig. 2: The overview procedure of the proposed dataset.

3.2 Collection of Stained Cell Images

We collected 6,587 stained cell microphotograph images from Guangdong Women and Children Hospital of 408 individuals who did chromosome analysis in the same month. All individual-related privacy (including name and age) has been removed. All cell images have a standard resolution of 1360 × 1024 with 96 dpi. The individual id and serial number are automatically generated for identifying these cell images. For example, 20 cell images were collected from the peripheral blood of an individual whose generated ID is P10000. Then, these images are named as P10000.001.A.JPG to P10000.020.A.JPG, respectively.

3.3 Stained Cell Image Selection

In the clinical chromosomal analysis, cell images of the metaphase of cell division are automatically collected from

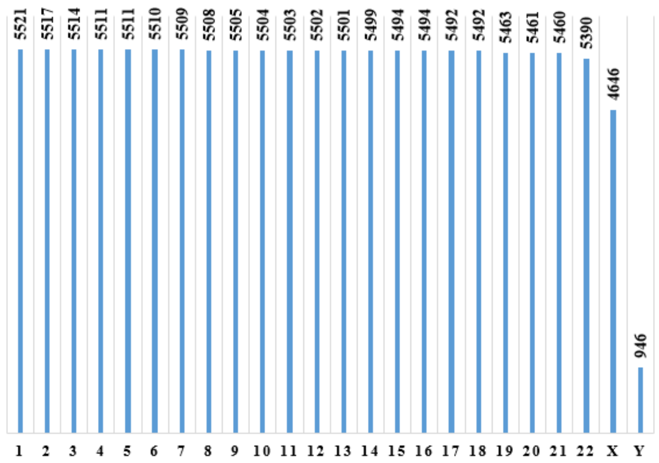


Fig. 3: The total number of clinical chromosome samples per category.

the microscopes. Accordingly, some images may have quality issues, such as cell impurities, noise, indistinct shooting. 16 experienced cytogeneticists experts of Guangdong Women and Children Hospital worked together to remove those low-quality cell images which are not suitable for chromosome karyotyping analysis. Finally, we got 2,795 high-quality cell images.

3.4 Chromosome Karyotyping Analysis

We organized 16 experienced cytogeneticists experts to do chromosome karyotyping analysis and got 2,795 karyotypes of corresponding cell images. As manually chromosome karyotyping analysis is an error-prone process influenced by factors such as fatigue and decrease of attention, each chromosome karyotype is validated and confirmed by another three cytogeneticists to ensure the correctness of chromosome karyotyping analysis. 32 error chromosome karyotypes are found in all karyotype analysis results. Finally, we obtain a total of 2,763 chromosome karyotypes.

3.5 Labeled Chromosome Instance Extraction

All chromosome instances in chromosome karyotypes have been correctly classified. Consequently, labeled chromosome instances can be extracted from chromosome karyotypes. In the chromosome instance extraction stage, firstly, we segmented individual chromosome instance from karyotypes and placed it in the center of a 300×300 blank image as a chromosome sample. Then, we saved the image into the corresponding folder. Finally, we built a dataset consisting of 24 folders with 126,543 chromosome samples. *Fig. 3* summarizes the total number of samples per category in the proposed dataset. As the X chromosome and Y chromosome consist of a pair of sex chromosomes, the sum of the numbers of X and Y chromosomes in the dataset is approximately equal to the sum of a pair of autosomes.

There are two considerations for obtaining chromosome instances from karyotypes instead of segmenting them from cell images directly. The chromosome karyotypes analyzed and verified by cytogenetics experts can minimize the risk of chromosome classification errors by comparing chromosome instances in a cell image to help inference their corresponding categories. Moreover, all chromosome instances obtained from the karyotypes have been adjusted into corrected polarity by the cytogenetics experts to ensure that all chromosomes in the proposed dataset are in the unified direction.

4 Baselines

TABLE 1: The baseline models for clinical chromosome classification.

Baseline	Base Model	Source	Flops	Params
baseline#1	MobileNetV2 [35]	CVPR@2018	0.312B	2.255M
baseline#2	DenseNet [36]	CVPR@2017	2.865B	6.98M
baseline#3	ResNet-50 [37]	CVPR@2016	4.110B	23.56M
baseline#4	ResNet-101 [37]	CVPR@2016	7.832B	42.55M
baseline#5	ResNet-152 [37]	CVPR@2016	11.557B	58.19M
baseline#6	ResNeXt-101 [38]	CVPR@2017	16.475B	86.79M
baseline#7	Res2Net-50 [39]	TPAMI@2019	4.281B	25.70M
baseline#8	ResNest-50 [40]	arXiv@2020	5.399B	25.48M

This paper presents multiple chromosome classification baselines based on deep learning image classification models to evaluate the proposed clinical chromosome classification dataset and provide off-the-shelf chromosome classifiers for clinical applications. This work comprehensively considers the inference speeds, computations, and performance of presented baselines to meet different clinical application demands. Therefore, a set of clinical chromosome classification baselines based on MobileNetV2 [35], ResNet-50/101/152 [37], DenseNet121 [36], ResNeXt101$(32 \times 8d)$ [38], Res2Net [39], and ResNest [40] image classification models are transferred and implemented in this work. Table 1 lists the base models of presented baselines for the clinical chromosome classification.

MobileNetV2 [35] is a neural network architecture that is specifically tailored for mobile and resource-constrained environments. The most significant advantage of MobileNetV2 is extremely fast running efficiency on compact devices such as cell phones.

DenseNet [36] shortens those connections between layers closing to input and layers closing to the output by linking each convolutional neural layer to every other layer in a feed-forward fashion. The advantages of DenseNet include vanishing-gradient alleviation, feature propagation strengthening, feature reusing, small-scale parameters, which make it easy to train.

The models of the ResNet [37] series are very classic and prevalent deep learning network models widely applied in various fields. According to the total number of neural layers, the ResNet series models include ResNet-50, ResNet-101, and ResNet-152, which means that the corresponding model has 50/101/152 layers of convolutional networks. As ResNet50 makes a reasonable tradeoff between calculations and performance, it is often taken as a baseline during new deep learning architecture development.

ResNeXt [38] is a simple, highly modularized network architecture for image classification. A ResNeXt model is built by repeating a building block that aggregates a set of transformations with the same topology. The ResNet model of the proposed baseline 6, specific to ResNeXt-101$(32 \times 8d)$, consists of 101 convolutional neural layers constructed by the building block with 32 parallel paths and eight channels. The most significant advantage of ResNeXt series models is superior performance, while the most apparent disadvantage is horrendous calculations caused by their massive parameters.

Res2Net [39] is a multi-scales features representation backbone that improves the building block of ResNet [37] by constructing hierarchical residual-like connections within one single residual block. The building block of Res2Net can be easily plugged into the state-of-the-art backbone models, e.g., ResNet [37] and ResNeXt [38]. Extensive ablation experiments on representative computer vision tasks have shown state-of-the-art results, demonstrating its excellent multi-scales features representation ability.

ResNeSt [40], a new variant of ResNet [37], is a modular *Split-Attention* block that enables attention across feature-map groups. Currently, extensive experiments showed that ResNeSt architecture outperforms other networks with similar model complexities on the ImageNet classification task.

5 Evaluations

This section presents evaluation details in six parts. This first part introduces evaluation objectives, while the second part provides the definitions of evaluated metrics. The third part describes the evaluation experimental settings and details. The last three parts are evaluation experimental results.

5.1 Evaluation Objectives

The experimental objectives are set to evaluate the scientificity of the proposed dataset for enabling benchmarking and building baselines of clinical chromosome classification, which is conducted by evaluating and verifying the objectivity and effectiveness of the proposed dataset and corresponding baselines. Accordingly, the experiment objectives include two aspects. The first experimental goal is to test chromosome classification performances of the presented reproducible baselines on the proposed clinical chromosome classification dataset. Meanwhile, these experiments are required to verify whether the chromosome classification performances of the presented baselines are superior to those reported by the existing chromosome classification methods. Secondly, these experiments are designed to test whether the proposed massive clinical dataset can boost performances of existing chromosome classification methods or not.

5.2 Evaluation Metrics

As clinical chromosome classification is a familiar classification problem, we can follow standard evaluation metrics to evaluate the performance of these baselines. Standard evaluated metrics for classification problems include $precision$, $accuracy$, F_1, $sensitivity$, and $specificity$. However, these metrics are designed for binary classification tasks, and clinical chromosome classification is a multi-class classification task, causing these metrics cannot be directly applied to evaluations of presented baselines. Therefore, to apply the above metrics to evaluations of presented baselines, the following four criteria should be defined first.

- True Positives(TP_j): Chromosome instances are classified as category j which actually belong to j.
- False Positives(FP_j): Chromosome instances are classified as category j which actually do not belong to j.
- False Negatives(FN_j): Chromosome instances are classified as category $k(\forall k \neq j)$ which belong to j.
- True Negatives(TN_j): Chromosome instances are classified as category $k(\forall k \neq j)$ which do not belong to j.

Based on four criteria of the TP_j, FP_j, FN_j, and TN_j, the evaluation metrics of proposed baselines can be calculated as follows.

$$precision_j = \frac{TP_j}{TP_j + FP_j} \tag{1}$$

$$sensitivity_j = \frac{TP_j}{TP_j + FN_j} \tag{2}$$

$$specificity_j = \frac{TN_j}{TP_j + TN_j} \tag{3}$$

$$F_1^j = \frac{2 \cdot precision_j \cdot recall_j}{precision_j + recall_j} \tag{4}$$

In the above definitions, N_{types} denotes the total categories of chromosome instances while N represents the total sample number of the testing dataset. $precision_j$, $sensitivity_j$, $specificity_j$, and f_beta_j denote the calculations of $precision$, $sensitivity$, $specificity$, and F_1^j of the category j. Accordingly, calculations of $precision$, $sensitivity$, $specificity$, and F_1 of presented baselines over testing samples classes can be defined as follows.

$$accuracy = \frac{1}{N} \sum_{j=1}^{N_{\text{types}}} TP_j \tag{5}$$

$$precision = \frac{1}{N_{\text{types}}} \sum_{j=1}^{N_{types}} precision_j \tag{6}$$

$$F_1 = \frac{1}{N_{\text{types}}} \sum_{j=1}^{N_{types}} F_1^j \tag{7}$$

$$sensitivity = \frac{1}{N_{\text{types}}} \sum_{j=1}^{N_{types}} sensitivity_j \tag{8}$$

$$specificity = \frac{1}{N_{\text{types}}} \sum_{j=1}^{N_{types}} specificity_j \tag{9}$$

Meanwhile, this work introduces a novel metric termed $InferenceTime$ to evaluate the running efficiency of a given baseline. It is calculated by the total consuming time of the given model for predicting 1000 samples consecutively. The smaller of the $InferenceTime$ metric value, the less time of the corresponding baseline consumes for a given model to predict a sample, which means higher running efficiency.

TABLE 2: Comparison results of presented baselines. These results are presented in terms of six evaluation metrics: average-precision of all testing sample ($\boldsymbol{precision}$), the average-accuracy of all testing samples ($\boldsymbol{accuracy}$), the average-fbeta of testing samples ($\boldsymbol{F_1}$), the average-sensitivity of all testing samples ($\boldsymbol{sensitivity}$), the average-specificity of all testing sample ($\boldsymbol{specificity}$), and the total consuming time for predicting per 1000 samples ($\boldsymbol{InferenceTime}$).

No.	Baseline	$\boldsymbol{precision}$	$\boldsymbol{accuracy}$	$\boldsymbol{F_1}$	$\boldsymbol{sensitivity}$	$\boldsymbol{specificity}$	$\boldsymbol{InferenceTime}$
1	baseline#1	97.54	97.58	99.20	98.07	97.12	$\mathbf{8.6}\boldsymbol{s}$
2	baseline#2	97.95	98.17	99.24	99.42	98.40	53.4s
3	baseline#3	98.34	98.62	99.60	98.98	98.29	14s
4	baseline#4	98.47	98.67	**99.67**	**99.49**	98.43	24.1s
5	baseline#5	98.39	98.51	99.63	99.38	98.89	29.6s
6	baseline#6	98.51	98.62	98.57	98.59	99.94	34.2s
7	baseline#7	99.28	99.27	99.21	99.16	**99.97**	52.4s
8	baseline#8	**99.32**	**99.33**	99.29	99.27	**99.97**	22.2s

TABLE 3: The classification accuracy of the presented baselines for each chromosome class.

Baseline	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**	**Avg**	**Max**
baseline#1	99.09	98.37	98.00	96.73	98.37	98.00	98.91	96.55	97.28	96.91	98.19	98.00	95.99	96.85	96.52	96.73	99.09	97.27	98.18	98.18	97.07	98.53	96.12	92.55	97.58	0
baseline#2	99.09	98.37	99.64	98.37	98.01	98.91	99.09	98.00	97.64	97.64	98.73	98.55	96.17	97.03	97.62	96.91	98.73	98.18	98.54	98.91	98.53	97.44	97.84	96.81	98.17	0
baseline#3	**99.64**	98.55	98.91	98.00	**99.46**	98.91	**99.64**	98.91	97.64	97.64	98.91	98.18	97.27	98.33	98.72	97.27	99.64	99.09	98.54	**99.64**	**99.45**	97.99	97.84	**98.94**	98.62	6
baseline#4	**99.64**	98.73	99.46	98.19	99.09	99.27	99.27	97.64	98.19	98.55	98.00	98.91	97.63	98.33	98.72	98.18	**99.82**	98.72	98.72	**99.64**	99.08	98.35	98.28	96.81	98.67	3
baseline#5	**99.64**	98.91	99.46	**99.09**	98.73	99.27	99.27	98.00	98.19	98.73	97.28	99.27	97.09	98.14	97.44	97.45	99.46	98.72	98.36	99.45	98.17	98.90	96.98	95.74	98.51	2
baseline#6	**99.64**	**99.73**	99.46	98.19	99.09	99.27	99.27	97.64	98.19	98.55	98.00	98.91	97.63	98.33	98.72	98.18	**99.82**	98.00	98.72	**99.64**	99.08	98.35	98.28	96.81	98.62	5
baseline#7	99.46	99.64	**99.64**	98.91	99.09	99.27	99.27	**99.27**	**99.00**	**99.18**	99.27	99.45	99.18	99.17	98.53	**99.27**	99.64	**99.64**	99.18	99.27	**99.45**	**99.36**	**99.78**	95.77	99.27	9
baseline#8	**99.64**	**99.73**	99.55	99.00	98.91	**99.55**	99.46	**99.27**	98.91	99.09	**99.46**	**99.55**	**99.36**	**99.26**	**99.08**	99.00	**99.82**	99.54	**99.36**	99.55	99.27	98.99	99.68	97.35	**99.33**	11

5.3 Evaluation Experimental Settings

The clinical benchmark dataset was divided into the training dataset (80%), validation dataset (10%), and testing dataset (10%) using random stratified sampling [41]. All baselines were trained on the same training and validation datasets and finally evaluated on the same testing dataset. All clinical chromosome classification baselines presented by this work have been developed under the *Pytorch* [42] development framework. All experiments were conducted under a CentOS OS workstation with Xeon(R) CPU E5-2620 v4 @ 2.1GHz, 64 GB of RAM, and 4 Nvidia GeForce RTX 2080Ti GPUs with 11,019MiB GPU memory.

The training processes of the proposed baselines are concluded as follows. First, all baseline models were pre-trained and transferred from ImageNet [43] classification. Second, this work utilized the One Cycle Learning (*1cycle LR*) [44] training methodology to gradually tune the proposed baselines on the proposed training dataset. Limited by the running memory capacity of GeForce RTX 2080Ti GPU, the batch size of all baselines was set to 32. The loss function is adapted by label-smoothing [18] cross-entropy loss with $\alpha = 0.1$. The learning rate lr is set to $1e^{-4}$. The hyperparameter of the max training epochs was set to 500, and an early-stopping [44] strategy was employed to terminate the training process when the validation loss does not descend in five consecutive epochs. Finally, the best training weights of the presented baselines were restored to the corresponding models when their training process has finished.

5.4 Evaluation Experimental Results of Different Baselines

The performance results of the presented baselines have been concluded in *Table 2*. The baseline#8 based on ResNeSt-50 [40] has obtained 99.32% classification precision and 99.33% classification accuracy, both of which are the best among all baselines. The baseline#4 based on ResNet-101 [37] has yielded a F_1 score of 99.67% and a sensitivity score of 99.49%, which exceeded other baselines. In the $specificity$ evaluation metric, baseline#7 based on Res2Net-50 [39] and #8 based on ResNeSt-50 [40] have surpassed all other baselines with a $specificity$ result of 99.97%. The $InferenceTime$ metric of baseline#1 based on MobileNetV2 [35] is 8.6 ms for per sample and beats all the other baselines. Accordingly, the comprehensive chromosome classification performances of baselines #4, #7, and #8 are pretty advanced, and there is no absolute optimal performance baseline that can beat others in all evaluation metrics.

According to the experimental results of the these baselines shown in *Table 2*, baseline#8 based on ResNeSt [40] model requires an average *inference time* of 22 ms for per sample, which is comparatively inefficient compared with baselines #1 and #3. However, baseline#8 has achieved the best clinical chromosome classification performance among all proposed baselines. Consequently, baseline#8 is suitable for those clinical chromosome classification scenarios requiring the most accurate classification performance.

The baseline#2 based on the MobileNetV2 [35] model is the most lightweight of all proposed baselines, making it capable of the highest inference efficiency. The drawback of baseline#2 is that the clinical chromosome classification performance is relatively weak among all baseline models. The baseline#3 based on the ResNet-50 [37] model has made a balanced trade-off between the classification performance and running efficiency with 98.62% classification accuracy and 14 ms for predicting per sample. Therefore, it is applicable to general clinical chromosome classification tasks.

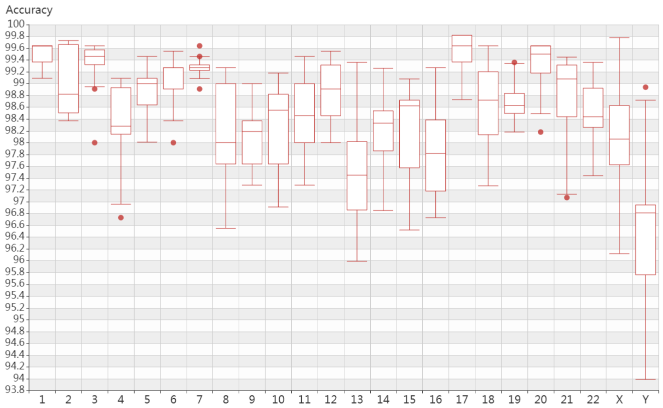


Fig. 4: The accuracy boxplot of all chromosome categories.

To evaluate the classification performance of different baselines on each category, this work has conducted further category-level experiments whose results have been concluded in *Table 3* and *Fig. 4*. According to the detailed experimental results shown in *Table 3*, different baselines perform slightly differently in different chromosome categories. The baseline#8 performed the most robust among all the baselines and achieved the highest classification accuracy among 11 chromosome categories, followed by baseline#7, which achieved the highest classification accuracy among 9 chromosome categories. Moreover, except for the highest accuracy of the Y chromosome of 98.94%, the highest classification accuracy rate of each category exceeds 99%, which demonstrates the advanced performance of proposed baselines. From the perspective of the accuracy of different baselines in each chromosome category (See boxplot in *Fig. 4*), the difficulties for correctly distinguishing different chromosomes are different. Autosomes 1, 3, 7, and 17 are the easiest to be classified correctly, while autosomes 4, 8, 9, 13, 16, sex chromosomes X and Y are comparatively difficult to be classified correctly. Nevertheless, for chromosome categories that are difficult to classify correctly, these baselines have achieved a good median accuracy with 96.8% or more.

TABLE 4: Quartile statistics results of various baselines.

	Min	$Q1$	$Medium$	$Q3$	IQR	Max
#1	92.55	96.73	97.64	98.24	1.51	99.09
#2	96.17	97.64	98.28	98.73	1.10	99.64
#3	97.27	98.00	98.82	99.18	1.18	99.64
#4	96.81	98.19	98.72	99.14	0.95	99.82
#5	95.74	97.87	98.73	99.27	1.41	99.64
#6	96.81	98.19	98.64	99.27	1.08	99.82
#7	95.77	99.18	99.27	99.45	0.28	99.79
#8	97.35	99.06	99.36	99.55	0.49	99.82

To evaluate the stabilities of various baselines, this work conducts a quartile [45] statistics analysis whose results

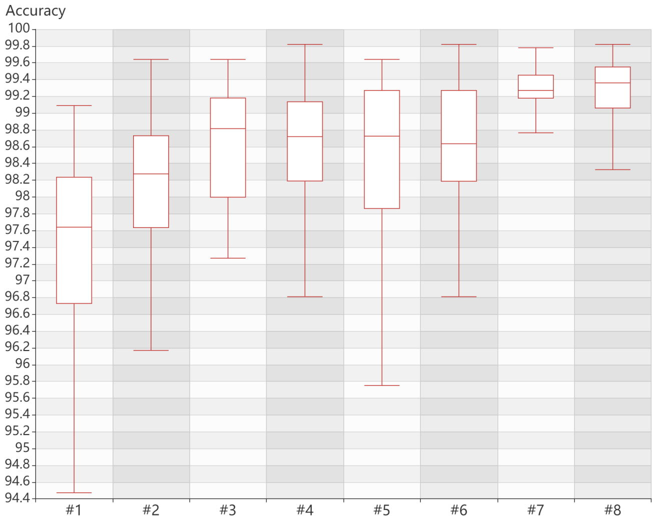


Fig. 5: The accuracy boxplot of various chromosome classification baselines.

have been concluded in *Table 4*. In *Table 4*, #1∼#8 refer to baseline#1 to baseline#8 while *min* and *max* refer to the lowest and largest data point excluding any outliers, respectively. $Q1$ means the first quartile (25^{th} percentile), which is also known as the lower quartile. $Medium$ represents the middle value of the dataset, which is also named $Q2$. $Q3$ denotes the 75^{th} percentile, which is also called the upper quartile. IQR, shorted for *interquartile range*, is the distance between the upper and lower quartiles, formalized as $IQR = Q3 - Q1$. In *Table 4*, the maximum value of IQR is 1.51 while the minimum value is 0.28, which are obtained by the baseline#1 based on MobileNetV2 [35] and baseline#7 based on Res2Net-50 [39], respectively. Therefore, the chromosome classification performance of baseline#7 is the most stable, while the chromosome classification performance of baseline#1 is relatively the worst. The baseline#8 based on ResNeSt-50 [40] has the highest medium value, indicating that the comprehensive chromosome classification performance of the model is the highest. Accordingly, for application scenarios that require strict stability, the baseline#7 based on Res2Net-50 [39] is recommended, and for scenarios that require higher comprehensive performance, the baseline#8 based on ResNeSt-50 [40] is suggested. *Fig. 5* utilizes a boxplot to show chromosome classification stabilities of different baseline models visually.

5.5 Comparisons with Existing Works

Some studies [5–9, 15] have made some contributions to the chromosome classification task. To compare the differences of these studies with the proposed benchmark in chromosome classification reported performance, we collect their results from the corresponding articles and conclude these results in *Table 5*.

According to the comparisons, the existing methods have achieved different classification accuracies on their respective datasets which contain different amounts of chromosome samples. For example, MixedNet [3] reported chromosome classification accuracy of 98.73%. The VarifocalNet [8] published 98.90% chromosome classification accuracy on

TABLE 5: Chromosome classification results of existing methods.

Method	Accuracy	#Samples	Accessible
Siamese-Networks [14]	84.6%	1,740	No
Deep CNN [5]	86.7%	9,600	No
CSVMT [15]	91.00%	4,400	No
Res-CRANN [6]	91.94%	5,256	BioImLab
Vanilla CNN [7]	92.50%	10,304	No
CNN [16]	93.79%	4,184	Github
VarifocalNet [8]	98.90%	87,831	No
ChromNet2 [17]	91.30%	21,423	No
CIR-Net [9]	95.98%	2,990	Github
MixedNet [19]	98.73%	10,856	No
Presented benchmark	**99.33%**	**126,453**	Available

its private dataset with 87,832 samples, which is currently the highest reported chromosome classification performance in the regular chromosome classification task.

The best baseline achieved an accuracy score of 99.33% on the proposed clinical dataset containing 126,453 samples, which is a state-of-the-art result among all public reported performances in terms of reported values. However, above results [3, 5–9, 14–17] are obtained on different datasets, and most of those datasets [3, 5, 7, 8, 14, 15, 17] are not available to peers, it is difficult and unrealistic to conclude which method is the most advanced.

5.6 Evaluating Existing Methods on the Proposed Dataset

To test the performance improvements of existing chromosome classification algorithms [5, 7, 9, 16, 17, 19] based on the proposed clinical dataset, this work implemented and evaluated existing algorithms on the proposed dataset and concluded their results in *Table 6*. Methods [6, 8, 14, 15] did not release their source codes or describe the reproduction details in their papers, which makes us hard to implement or evaluate these methods in this work.

TABLE 6: Experimental results of existing methods evaluated on the proposed dataset.

Method	Reported	Evaluated	Improvement
Deep CNN [5]	86.7%	92.09%	+5.39%
Vanilla CNN [7]	92.50%	95.48%	+2.98%
CNN [16]	93.79%	94.70%	+0.91%
ChromNet2 [17]	91.30%	95.23%	+3.93%
CIR-Net [9]	95.98%	98.64%	+2.66%
MixedNet [19]	98.73%	98.92%	+0.19%

According to experimental results shown in *Table 6*, the proposed clinical dataset has boosted chromosome classification performance of all existing chromosome classification methods in varying degrees. Method [5] has been boosted the most significantly with 5.39 percentage points accuracy improvement, whose chromosome classification accuracy has been improved from 86.7% to 92.09%. For method [19] with excellent chromosome classification performance, the proposed clinical dataset can still improve its classification accuracy by 0.19 percentage points. Method [16] was reported as more advanced than methods [7, 17] but less competitive on the condition of fair comparisons. *Fig. 6* visually shows the classification performance improvement of existing chromosome classification methods boosted by the proposed clinical dataset.

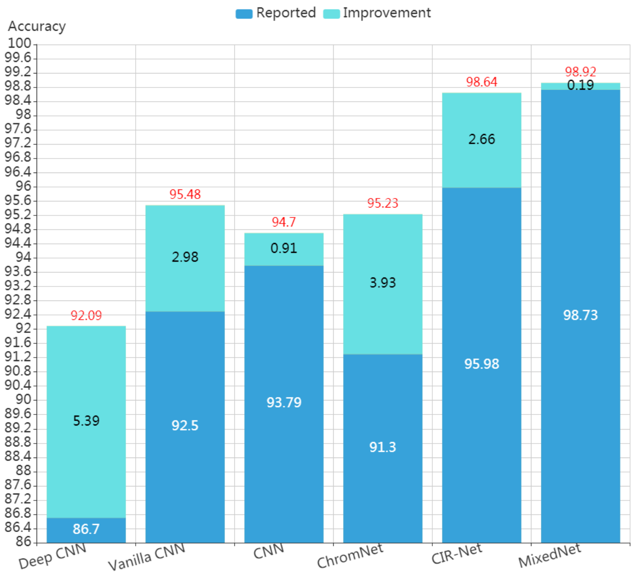


Fig. 6: The chromosome classification improvements of existing chromosome classification methods boosted by the proposed clinical dataset.

5.7 Analysis of Experimental Results

According to experimental results (*Table 2*, *Table 3*, *Table 4*, *Table 5*, and *Table 6*), at least three important facts have been observed in this work:

- A large-scale clinical chromosome dataset is vital for both evaluating existing chromosome classification methods and boosting the performance of existing methods. Based on the proposed clinical dataset, the chromosome classification performance of existing methods can be fairly compared. Meanwhile, existing methods can be boosted by a larger dataset than their original.
- A high-quality massive clinical chromosome classification dataset is more valuable than well-designed chromosome classification models. Although these chromosome classification models have been dedicated to many targeted designs for the characteristics of chromosomes and obtained quite advanced chromosome classification performance on their private small-scale datasets. However, benefit from a massive dataset, the chromosome classification performance of a classical image classification model ResNet-50 [37] is superior to the performance of most of the existing targeted chromosome classification methods [5–7, 9, 14–17]. The performance of a general image classification model ResNeSt-50 [40] exceeds to the performance of all targeted chromosome classification methods.
- Based on the proposed large-scale clinical dataset, the best baseline has achieved state-of-the-art chromosome classification performance according to existing chromosome classification methods without

modifying of a existing image classification model [40].

6 Conclusions and Limitations

6.1 Conclusions

This paper investigated deep learning chromosome classification methods in-depth and summarized two fatal challenges resulting in existing methods being difficult widely used in clinical practices. The first challenge is that the implementation details and datasets of most of the current chromosome classification methods are not available for peers, resulting in these methods being hard to be replicated, verified and applied. Meanwhile, private chromosome classification datasets are also made fair comparisons between different models impossible. The second challenge is that the current methods were trained and tested on their own small-scale datasets, making their reported performance not objective enough.

This work addressed the above two challenges of the chromosome classification issue by proposing a large-scale chromosome classification dataset and various off-the-shelf clinical chromosome classification baselines. To our best knowledge, the proposed dataset is the first publicly available dataset whose volume scale is over 120,000. Moreover, based on the proposed dataset, one of the proposed baselines achieved state-of-the-art reported classification performance with 99.33% accuracy. The most vital contributions of this paper have been concluded as follows:

- According to our best knowledge, this paper is the first work to build and release a high-quality clinical chromosome classification dataset whose data scale is more than 120,000. The proposed dataset not only inspires those with in-depth algorithm design capabilities but no corresponding clinical chromosome data teams to address chromosome classification challenges but also provide a basis for fair comparisons of different chromosome classification models and algorithms.
- This paper is the first work to reproduce, evaluate and compare existing chromosome classification models and algorithms on a more than 120,000 samples level chromosome classification dataset.
- This paper is the first work to verify that a high-quality clinical chromosome classification dataset whose data size is over 120,000 is more important than well-designed chromosome classification models.
- Based on our clinical dataset, this paper presents a state-of-the-art chromosome classification baseline whose clinical chromosome classification accuracy hit 99.33%, precision achieves 99.32%, and F_1 obtains 99.29%.

6.2 Limitations and Future Work

This work is not without limitations. The chromosome samples in the constructed clinical dataset of this study are collected from one of the largest and most authoritative prenatal diagnosis institutions in China. In addition, all the chromosome image samples were collected from CoolCube series cameras of MetaSystems company. In the future, we will collect more chromosome samples from different series of cameras from multiple different medical institutions for enriching the diversity of this dataset to promote further clinical research and applications better.

References

[1] I. Campos-Galindo, “Cytogenetics techniques,” in *Human Reproductive Genetics*. Elsevier, 2020, pp. 33–48.

[2] J. McGowan-Jordan, *ISCN 2016: An International System for Human Cytogenomic Nomenclature (2016); Recommendations of the International Standing Human Committee on Human Cytogenomic Nomenclature Including New Sequence-based Cytogenomic*. Karger, 2016.

[3] C. Lin, G. Zhao, A. Yin, B. Ding, L. Guo, and H. Chen, “A multi-stages chromosome segmentation and mixed classification method for chromosome automatic karyotyping,” in *International Conference on Web Information Systems and Applications*. Springer, 2020, pp. 365–376.

[4] C. Lin, G. Zhao, A. Yin, Z. Yang, L. Guo, H. Chen, L. Zhao, S. Li, H. Luo, and Z. Ma, “A novel chromosome cluster types identification method using resnext wsl model,” *Medical Image Analysis*, vol. 69, p. 101943, 2021.

[5] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, and S. Karande, “Crowdsourcing for chromosome segmentation and deep classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 34–41.

[6] M. Sharma, L. Vig *et al.*, “Automatic chromosome classification using deep attention based sequence learning of chromosome bands,” in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[7] W. Zhang, S. Song, T. Bai, Y. Zhao, F. Ma, J. Su, and L. Yu, “Chromosome classification with convolutional neural network based deep learning,” in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–5.

[8] Y. Qin, J. Wen, H. Zheng, X. Huang, J. Yang, N. Song, Y.-M. Zhu, L. Wu, and G.-Z. Yang, “Varifocal-net: A chromosome classification approach using deep convolutional networks,” *IEEE transactions on medical imaging*, vol. 38, no. 11, pp. 2569–2581, 2019.

[9] C. Lin, G. Zhao, Z. Yang, A. Yin, X. Wang, L. Guo, H. Chen, Z. Ma, L. Zhao, H. Luo *et al.*, “Cir-net: automatic classification of human chromosome based on inception-resnet architecture,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[10] L. Xiao and C. Luo, “Deepacc: Automate chromosome classification based on metaphase images using deep learning framework fused with priori knowledge,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 607–610.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, “A dataset for breast cancer histopathological image classification,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[12] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du *et al.*, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," *arXiv preprint arXiv:2103.04098*, 2021.

[13] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.

[14] S. Jindal, G. Gupta, M. Yadav, M. Sharma, and L. Vig, "Siamese networks for chromosome classification," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 72–81.

[15] A. O. Kusakci, B. Ayvaz, and E. Karakaya, "Towards an autonomous human chromosome classification system using competitive support vector machines teams (csvmt)," *Expert Systems with Applications*, vol. 86, pp. 224–234, 2017.

[16] X. Hu, W. Yi, L. Jiang, S. Wu, Y. Zhang, J. Du, T. Ma, T. Wang, and X. Wu, "Classification of metaphase chromosomes using deep convolutional neural network," *Journal of Computational Biology*, vol. 26, no. 5, pp. 473–484, 2019.

[17] R. Remya, S. Hariharan, V. Vinod, D. J. W. Fernandez, N. M. Ajmal, and C. Gopakumar, "A comprehensive study on convolutional neural networks for chromosome classification," in *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, 2020, pp. 287–292.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[19] C. Lin, G. Zhao, A. Yin, L. Guo, H. Chen, and L. Zhao, "Mixnet: A better promising approach for chromosome classification based on aggregated residual architecture," in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE, 2020, pp. 313–318.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[21] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 263–273.

[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[24] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[25] S.-L. Liew, J. M. Anglin, N. W. Banks, M. Sondag, K. L. Ito, H. Kim, J. Chan, J. Ito, C. Jung, S. Lefebvre *et al.*, "The anatomical tracings of lesions after stroke (atlas) dataset-release 1.1," *bioRxiv*, p. 179614, 2017.

[26] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-unet: a dimension-fusion u shape network for chronic stroke lesion segmentation," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[27] K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, and S. Wang, "X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 247–255.

[28] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[29] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[31] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[32] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[33] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.

[34] R. Huang, C. Lin, A. Yin, H. Chen, L. Guo, G. Zhao, X. Fan, S. Li, and J. Yang, "A clinical dataset and various baselines for chromosome instance segmentation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

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

[38] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[39] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[40] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," pp. 1–4, 2017.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[44] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[45] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005.

Acknowledgment

This work was supported by National Key R&D Program of China (2018YFB1404402), Guangdong Science & Technology Fund (No.2019B010137003, No.2016B030305006, No.2018A07071702, and No.201804010314), Guangzhou Science & Technology Fund (No.201804010314 and No.2012224-12), and VeChain Foundation (No.SCNU-2018-01). Z. Yang is supported by Academy of Finland (projects 314177 and 307929) and The Research Council of Norway (project 287284).