



# Prediction of ROP Zones Using Deep Learning Algorithms and Voting Classifier Technique

Nazar Salih<sup>1,2</sup> · Mohamed Ksantini<sup>2</sup> · Nebras Hussein<sup>3</sup> · Donia Ben Halima<sup>2</sup> · Ali Abdul Razzaq<sup>4</sup> · Sohaib Ahmed<sup>4</sup>

Received: 30 January 2023 / Accepted: 11 May 2023  
© The Author(s) 2023

## Abstract

Retinopathy of prematurity (ROP) can cause blindness in premature neonates. It is diagnosed when new blood vessels form abnormally in the retina. However, people at high risk of ROP might benefit significantly from early detection and treatment. Therefore, early diagnosis of ROP is vital in averting visual impairment. However, due to a lack of medical experience in detecting this condition, many people refuse treatment; this is especially troublesome given the rising cases of ROP. To deal with this problem, we trained three transfer learning models (VGG-19, ResNet-50, and EfficientNetB5) and a convolutional neural network (CNN) to identify the zones of ROP in preterm newborns. The dataset to train the model contains 1365 fundus images from the ROP screening. This dataset was gathered from the Private Clinic Al-Amal Eye center in Baghdad, Iraq. The models above are ensemble through voting classifier techniques to increase the performance. The proposed method had an overall accuracy of 88.82 percent when employing the voting classifier. On the other hand, EfficientNetB5 has outperformed other models in terms of accuracy with 87.27%.

**Keywords** Artificial intelligence · Machine learning · Deep learning · Deep convolutional neural networks · Voting classifier · Retinopathy of prematurity · Fundus images

## Abbreviations

ROP	Retinopathy of prematurity
AI	Artificial intelligence
ICROP	International classification of retinopathy of prematurity
RETCAM	Retinal camera
DL	Deep learning
ML	Machine learning
CNN	Convolutional neural network
DCNN	Deep convolutional neural network
DNN	Deep neural network
AUC	Area under the curve
ROC	Receiver operating characteristics

KIDROP	Karnataka internet-assisted diagnosis of retinopathy of prematurity
VGG	Visual geometry group
GA	Genetic algorithm
GPUs	Graphics processing units
SVM	Support vector machine
Tele-ROP	Telemedical retinopathy of prematurity

## 1 Introduction

In 1940, Terry was the primary researcher to define and characterize retinopathy of prematurity (ROP) as retrolental fibroplasia caused by a retinal detachment behind the lens [1]. Since then, it has been generally accepted that this is the primary cause of childhood blindness worldwide [2–4]. As a result, with the advent of neonatal intensive care units, there has been an increase in the percentage of newborns that survive birth at a gestational age of fewer than 37 weeks [2]. As many as 15 million infants are born each year prematurely worldwide [3]. These days, ROP is a major public health issue [4]. Some significant causes of ROP-related blindness are a paucity of ophthalmologists trained to diagnose and treat the disease and a lack of delayed screening [4]. If

✉ Nazar Salih  
nazar.s2009@yahoo.com

<sup>1</sup> National School of Electronic and Telecommunications, University of Sfax, Sfax, Tunisia

<sup>2</sup> Control and Energies Management Laboratory (CEM-Lab), National Engineering School of Sfax, University of Sfax, Sfax, Tunisia

<sup>3</sup> Biomedical Engineering Department, Al-Khwarizmi College of Engineering University of Baghdad, Baghdad, Iraq

<sup>4</sup> Ibn AL Haitham Teaching Eye Hospital, Baghdad, Iraq

high-risk individuals can have their ROP treated early on, they will be able to keep most or all of their eyesight. Thus, screening for ROP at an early stage is essential to prevent permanent visual impairment [5].

ROP is categorized into stages (1–5) depending on the severity of the illness, zones (1–3), and the existence of the plus disease, as per the International Classification of Retinopathy of Prematurity (ICROP) principles issued in 1984 [6], 1987 [7], and 2005 [8]. Zone 1 is a circular region with a radius equal to twice the distance between the center of the optic disc and the fovea of the macula. Except for zone 1, zone 2 is an annular gap with a radius equal to the distance from the optical disc to the serrated nasal border. Zone 3 consists of the remaining crescent-shaped territories outside zones 1 and 2. Figure 1 depicts how the severity of the condition defines ROP as per the International Classification of Retinopathy of Prematurity (ICROP).

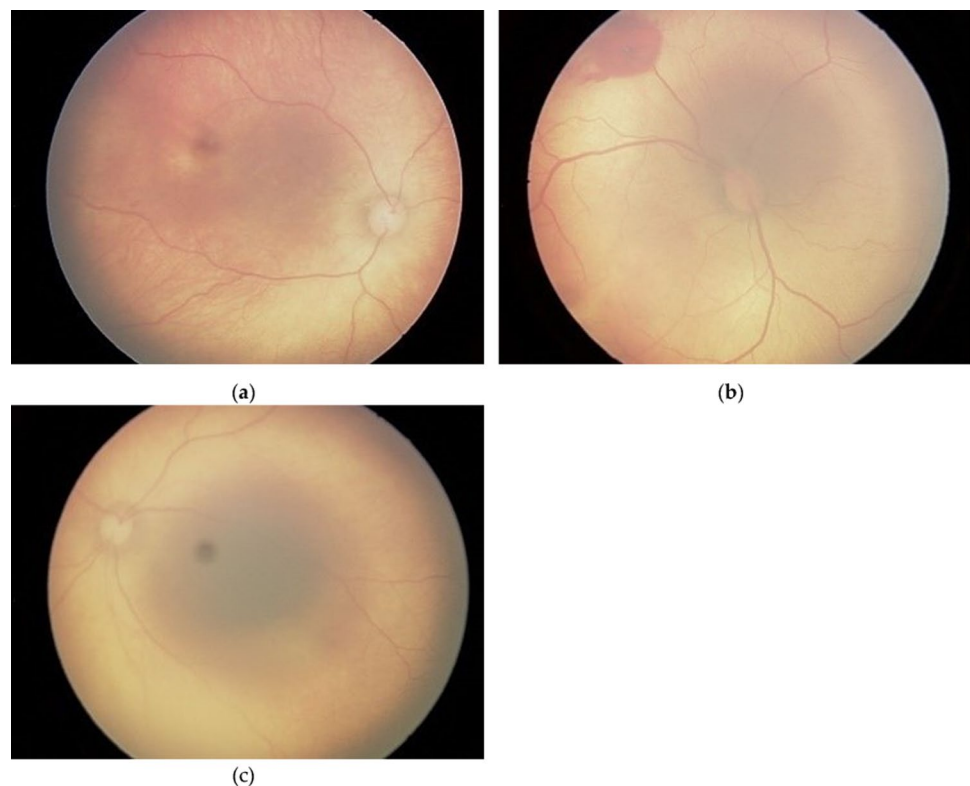
The diagnosis of ROP mainly involves retinal image acquisition. The wide-angle optical retinal imaging system: retinal camera (Retcam), is rapidly being used for ROP fundus inspection. It can photograph, store, output, and transmit multidirectional fundus images. In addition, its framework is better for follow-up, consultation, clinical research, and education. Numerous obtained morphological datasets have become accessible [12–14].

Improvements in AI technology have led to a rise in the application of AI in the medical profession. The area of medical diagnosis has significantly benefited from the use

of deep learning models, which have been used in various computer vision tasks such as picture classification, object recognition, image segmentation, and illness diagnosis. It has been proposed to employ AI to alleviate some pressure placed on medical professionals due to the availability of large data sets and the development of deep network designs. Numerous conventional machine learning techniques have been applied to the problem of picture identification and categorization. However, other methods, like extraction of features and dimensionality reduction, are needed to complete the job, adding extra time to the process. However, the models' effectiveness may suffer if the image matrix is converted to a one-dimensional vector and essential details are lost. The convolutional layers of a convolutional neural network (CNN) are responsible for feature extraction from the input pictures, while the pooling layer is responsible for dimensionality reduction.

Many methods, such as ophthalmoscope, can do diagnostic analysis on ROP using these large databases. But using artificial intelligence (AI), we can obtain efficient results. Notably, deep learning (DL) algorithms are capable of recognizing pathogenic features in retinal disorders. Similarly, transfer learning is a valuable paradigm that applies prior learned information and abilities to a new but related task. Pretrained models such as VGG-19, ResNet-50, and EfficientNetB5 were trained using significant data sources like ImageNet, which has over a thousand categories and 1.2 million natural pictures [9]. These models are constructed from

**Fig. 1** Illustration of retina images of the zones 1–3 of ROP: **a** zone 1, **b** zone 2, **c** zone 3



the ground up utilizing considerable computer resources. These models have acquired knowledge of characteristics such as edges, forms, illumination, rotation, and geographical data. These data might be used to extract features from photos in various fields. As a result, the amount of available training datasets is critical for a model to attain excellent efficiency; training the model with small data may cause low performance or overfitting, which may be corrected by transfer learning. As a result, transfer learning is beneficial for classification issues; it boosts a model's generalization ability when the training dataset is restricted (less than a thousand examples) [10]. This method is excellent for detecting images and predicting disease in small datasets like the one used in this study.

Majority and weighted voting for categorization are two of the most used combination techniques. In a simple majority vote, the option with the most votes among several possible choices is chosen. After all the classifiers have been trained separately, majority voting does not need additional parameter adjustment [11, 12]. With weighted voting, the relative importance of each classifier's output may be adjusted independently of the others. The weight should be relatively high for the output class where the classifier shows the most promise. Therefore, choosing the correct weights of votes for all the classes of each classifier [13] is a significant challenge. Consider the weighting problem as an optimization problem. As a result, artificial intelligence methods like genetic algorithms (GA) and particle swarm optimization can be used to find a solution (PSO).

A comprehensive and dependable automatic ROP detection system is needed to identify ROP at its earliest phases of development. Using fundus pictures obtained from premature newborns, the present study sought a high degree of accuracy in diagnosing ROP. The model was trained using a dataset, and eye-based disorders were evaluated to predict classification performance. CNN is also used to evaluate the effectiveness of transfer learning models. This project aimed to develop a CAD system for diagnosing ROP in a clinical context. We acquired a high degree of precision in predicting eye-based situations.

The structure of the rest of the paper is as follows: the introduction is in Sect. 1. Section 2 discusses the related work. The proposed methodology, dataset collection, and characteristics are described in Sect. 3. Section 4 discusses the classification model performance results in the diagnosis of ROP. While Sect. 5 highlights the conclusion.

## 2 Related Work

Scientists worldwide have been interested in ROP research and the subsequent creation of computer-aided diagnostic tools for ROP screening. Algorithms for ROP screening

using standard image processing and ML were proposed using small datasets of images. They were inaccurate when applied to complex problems and time consuming to complete because they relied on manually crafted features. Deep learning has become increasingly popular in the medical imaging industry due to an increase in both the size of the dataset and the accessibility of graphics processing units (GPUs). A deep learning network can learn high-level data features without any human intervention.

Recent studies in detecting ROP have shown significant advancements, but still, a few gaps make this study different from previous studies. Such as Agrawal et al. [14] proposed a novel method for detecting zones 1, 2, and 3 in retinal images where the macula is not present using an ensemble of U-Net and Circle Hough Transform. The proposed model is generalizable because it was trained using a variety of image sizes. One image was tested in 30 ms, and the entire training process only took 14 min. Also, Mulay et al. [15] proposed a CNN-based system focused on ridge detection, an important landmark in ROP diagnosis. Their contribution is to use a CNN-based model called Mask R-CNN for demarcation line/ridge detection, which will help clinicians detect ROP stage 2 more accurately. To overcome poor image quality, the proposed system employs a pre-processing step of image enhancement. They use labeled neonatal images in this study and investigate the use of CNN to localize ridges in these images. They used the KIDROP project's dataset of 220 images of 45 babies. The system was trained on 175 retinal images using ground truth ridge region segmentation.

Moreover, Wang et al. [16] proposed using deep neural networks to create the automated ROP detection method known as DeepROP (DNNs). ROP identification and grading activities made up ROP detection. Id-Net and Gr-Net, two distinct DNN models, were created for identification and grading tasks. Large-scale datasets of retinal fundus pictures were created by clinical ophthalmologists tagging the photos from ROP exams to develop the DNNs. Vijayalakshmi et al. [17] proposed an automated detection and classification method for determining the extent of ROP in wide-field telemedical photos. Sixty-five normal, eighty-nine stage 2, and forty-five stage 3 telemedical ROP (tele-ROP) pictures were obtained. Tele-ROP images have been analyzed using a Hessian classifier and a support vector machine (SVM) classifier for ROP detection and severity classification.

On the other hand, only some authors have employed pre-trained deep learning models to increase the prediction power of the model. Such as, Lei et al. [18] developed a unique deep convolutional neural network (DCNN) for wide-angle retinal pictures to diagnose ROP automatically. To be more precise, they start with ResNet-50 as their foundation architecture and then enhance it by including a channel and a spatial attention module. The trained

models are then shown to realize pathological structural localization using a class-discriminative approach called gradient-weighted class activation mapping (Grad-CAM). Furthermore, Worrall et al. [19] examine the following two methods for employing convolutional neural networks (CNN) to assist physicians with ROP detection: (1) They refine a pre-trained GoogLeNet to serve as an ROP detector and, with minor adjustments, also produce an approximation of a Bayesian probability about the existence of illness. (2) To help with grading even further, they train a second CNN to provide original feature maps of diseases learned from the data. These feature maps draw attention to discriminative data that, in their opinion, might be utilized by physicians in conjunction with their classifier to help with screening. Moreover, Huang et al. [5] used transfer learning to five different architectures of deep neural networks to diagnose ROP in preterm newborns. According to their research findings, the VGG-19 model performed significantly better than the other models when detecting whether or not a preterm newborn had ROP. They observed that the VGG-19 model displayed exemplary performance in predicting ROP after doing fivefold cross-validation on the datasets to authenticate the dependability of the VGG-19 model. This was done to ensure that the VGG-19 model was accurate. Furthermore, Tong et al. [20] suggested a powerful deep learning-based prediction model to identify the nature of ROP from fundus images,

as well as to identify the phase of ROP and the occurrence of another disease, allowing for automated diagnosis and therapy. Thirteen certified retinal specialists annotated a total of 36,231 fundus images. A 101 layer ResNet and a Faster-RCNN were trained for picture classification and identification. To train and optimize our algorithms, we employed a tenfold cross-validation approach. The effectiveness of the intelligent system was evaluated using a four-degree classification test, which assessed its accuracy, sensitivity, and specificity (Table 1).

Subsequently, a few existing studies review the current technologies to detect ROP at early stages. Authors have compared different machine learning-based methodologies and proposed the research gaps. Such as Gensure et al. [21] analyze the present state of AI applications in ROP and shed light on the obstacles and solutions that need to be implemented to bring these algorithms to the bedside. They concluded that the generalizability of the algorithms to sustain performance on unseen data and the integration of artificial intelligence technologies into new or current clinical processes are stumbling blocks to the widespread adoption of AI for ROP detection and treatment. Furthermore, Scruggs et al. [22] focus on creating artificial intelligence for the automated detection of severe illness in ROP and draw attention to the clinical and technological issues associated with the design of artificial intelligence and its implementation in the real world.

**Table 1** A comparison of the proposed study with recent studies

Ref	Year	Dataset	VGG	ResNet	EfficientNet	CNN	Voting classifier	Others
[14]	2020	HVDROPDB Retcam HVDROPDB Neo HVDROPDB OD HVDROPDB BV	–	–	–	–	–	U-net
[15]	2019	KIDROP	–	–	–	✓	–	–
[16]	2018	Identification Dataset	–	–	–	–	–	Id-Net, Gr-Net
[18]	2021	<i>Shenzhen Screening for Retinopathy of Prematurity Cooperative Group</i>	–	✓	–	–	–	–
[19]	2016	Diabetic-retinopathy-detection (Kaggle)	–	–	–	–	–	GoogleNet
[5]	2020	Chang Gung Memorial Hospital, Linkou, Taiwan	–	–	–	–	–	Inception, DenseNet, MobileNet
[17]	2019	KIDROP	–	–	–	–	–	Hessian analysis, Support vector machine
[21]	2020	Diabetic-retinopathy-detection (Kaggle)	–	–	–	✓	–	–
[20]	2020	Wuhan University Eye Center	–	✓	–	✓	–	–
[22]	2020	KIDROP	✓	–	–	✓	–	–
This study	2022	Al-Amal eye center	✓	✓	✓	✓	✓	–

From the above discussion, it is clear that researchers have been using transfer learning models and CNN-based models to classify ROP. It is also clear that these studies have been using small-scale datasets to train the models, which could become a significant factor in the proposed model's performance. Lastly, to the best of our knowledge, no study has used a voting classifier to combine the prediction ability of all the models to increase the overall performance.

### 3 Materials and Methods

#### 3.1 Dataset

All photos were taken in the Private Clinic Al-Amal Eye Center in Baghdad, Iraq. Expert specialists captured all images using a RetCam3 imaging device. This specialized facility has been providing ROP screening services for many years. From 2015 to 2019, 1365 fundus pictures were acquired using ROP screening.

#### 3.2 Labeling

The study includes two senior ophthalmologists who have worked with ROP patients for more than 15 years. These specialists assigned three zones of classification to each of the fundus photos. The three ophthalmologists independently categorized the images before comparing them to discover any inconsistencies in the labeling procedure (i.e., to find whether the specialists allocated a specific image different labels). After a debate among the specialists, the labels were subsequently organized collectively, and a label was assigned to the photographs.

#### 3.3 Pre-processing

The fundus photos were 640 480 pixels in quality. However, they were shrunk to 224 224 when fed into our deep learning models. For training, we used data from 1029 patients. The study did not include unclear, blurry, or dark photos. We analyzed fundus images presenting various phases of ROP in a similar infant, ensuring no overlays between patients in the training and test datasets. The dataset utilized for training, evaluation, and testing of the model is split at random, as summarized in Table 2.

##### 3.3.1 Data Augmentation

Overfitting can occur during training if the model is trained with little data. To get around this problem, we used data augmentation to generate fresh retinal fundus pictures from the current training dataset. To create new datasets, data augmentation was employed. We employed augmentation

**Table 2** ROP zone dataset

	Zone 1	Zone 2	Zone 3
Train set (80%)	349	327	416
Validation set	44	41	52
Test set (20%)	87	82	104
Total	436	409	520

techniques such as rotation range [3, 3], width shift range [0.1, 0.1], height shift range [0.1, 0.1], zoom range [0.85, 1.15], and horizontal flip in this investigation. The training dataset was expanded 7 times, yielding 18,808 pictures for training.

#### 3.3.2 Deep Learning Models

In this work, we choose one model among many backbone models, including VGG-19 from the VGG group, ResNet-50 from the ResNet group, and EfficientNetB5 from the Efficient groups. CNN with full connectivity is used for comparison and categorization.

#### 3.4 VGG-19

The VGG-19 model is a type of VGG model that has 19 layers. VGG-19 is a progressive CNN with layers that have already been trained and a significant consideration of what shapes, colors, and structures make up an image. VGG-19 is an intense neural network trained on millions of images with complex classification tasks.

##### 3.4.1 ResNet-50

The ResNet-50 model, in particular, is a 50 layer convolutional neural network (CNN). In a ResNet, a sort of artificial neural network, residual blocks are piled on top of one another to form a network (ANN). Deep residual nets employ residual blocks to enhance the precision of the models. These skip connections can be used in either direction. First, they fix the vanishing gradient problem by creating a new path the gradient can take. The model can also learn an identity function with their help. This ensures that the model's upper layers maintain performance relative to their lower counterparts.

##### 3.4.2 EfficientNetB5

One of the EfficientNet models created for image classification is the Efficientnet-b5 model. TensorFlow was used to train this model. The ImageNet picture database was used to train all of the EfficientNet models. It is a simple and successful approach that allows more principled scaling up of



a baseline ConvNet to any goal resource limits while maintaining model efficiency.

### 3.4.3 CNN

A CNN or ConvNet is a subset of machine learning. It is one of the numerous artificial neural network types used for various tasks and data sources. A CNN is a deep learning network design used primarily for image processing and pixel computational applications. CNN has three layers: a convolutional layer, a pooling layer, and a fully connected (FC) layer. The first layer is the convolutional layer, while the last is the FC layer.

**3.4.3.1 Proposed Methodology** These DNN models were chosen to accomplish our primary objective of recognizing ROP. After obtaining the findings, a voting classifier is constructed to improve performance and compare with the previously described DNN models. The performance of the models was then reviewed based on the findings, and the model with the best performance was selected for detecting the disease's severity. Our solution comprised the weight loading of the Keras-provided pre-trained model. We merged classifiers by substituting the model's FC layers with four hidden layers.

The size of the first and second FC layers in all DNN models was 250, and the dropout layers had a drop rate of 20%. The third FC layer was 128 bytes, and the third dropout layer had a 20% drop rate. The last layer was a softmax layer for categorizing the fundus images, preceded by the FC layer for image categorization. The ReLU activation function is used in all dense layers. The loss function is categorical cross-entropy, and the "Adam" optimizer was chosen above alternative optimizers.

Rather than choosing only one model, a voting classifier may anticipate the average value of all the findings. This voting classifier receives the findings or results of all classifiers and produces a single resulting value [25, 26]. The hard voting classifier used in this work needs each model (from the four classification models that came before it) to cast a vote for each instance in the dataset. More than half of the votes have been disclosed for the final output prediction. The model would be classified using the majority class label.

Furthermore, a machine learning model's efficiency can be drastically altered by the hyperparameters used. Because of its reputation as a powerful optimization method, the Adam optimizer was selected for this task. Since 0.0001 is a widely used default number that has proven effective for many different kinds of deep learning models, it was selected as the learning rate. By gradually slowing down the training rate over time, the decay rate of 1e-6 helps prevent overfitting. Important hyperparameters included the number of epochs as well as the batch size. How many times the full training dataset is used

to update the model weights is set by the number of epochs. Preliminary trials suggested that the model maintained its gains in performance up to 50 epochs, but exhibited no further meaningful gains after that point, which led to the decision to limit the training to 50 epochs. The number of samples utilized to compute each gradient update during training is controlled by the batch size. Training times often decrease with increasing batch size, but more storage is needed. Although it may take longer to converge, a lower batch size can result in more precise updates. Given the amount of RAM available and the results of some exploratory trials, a batch size of 32 was settled on for this particular application.

The overall categorization process workflow is depicted in Figure 2. After the data had been pre-processed, it was split into two sets: the training set and the test set. Next, the enhanced training data were used to train the model. They deployed the model to the test dataset for classification, followed by model testing and hyperparameter adjustment to achieve the best results on the dataset. After that, the model's efficacy was assessed according to how well it made predictions and distinguished between different data classes. In addition, we displayed the difficulties posed by several models in categorizing the phases of ROP and computed the area under the curve (AUC) to compare their performances. Lastly, the outputs of the four DNN models are combined to construct the voting classifier.

**3.4.3.2 Evaluation Metrics** In this study, precision, recall, F-measure, and area under the curve were some of the metrics used to examine and compare the efficacy of each trained model. After this study, the data from the different classifiers were incorporated into the voting classifier to forecast accuracy.

This work assessed the trained model's precision and recall to categorize the ROP zones. Whether these values are actual or not, precision is measured as a propensity toward correctness and discusses how closely two or more quantities are related.

Accuracy (ACC): The proportion of properly identified samples to total samples [23]:

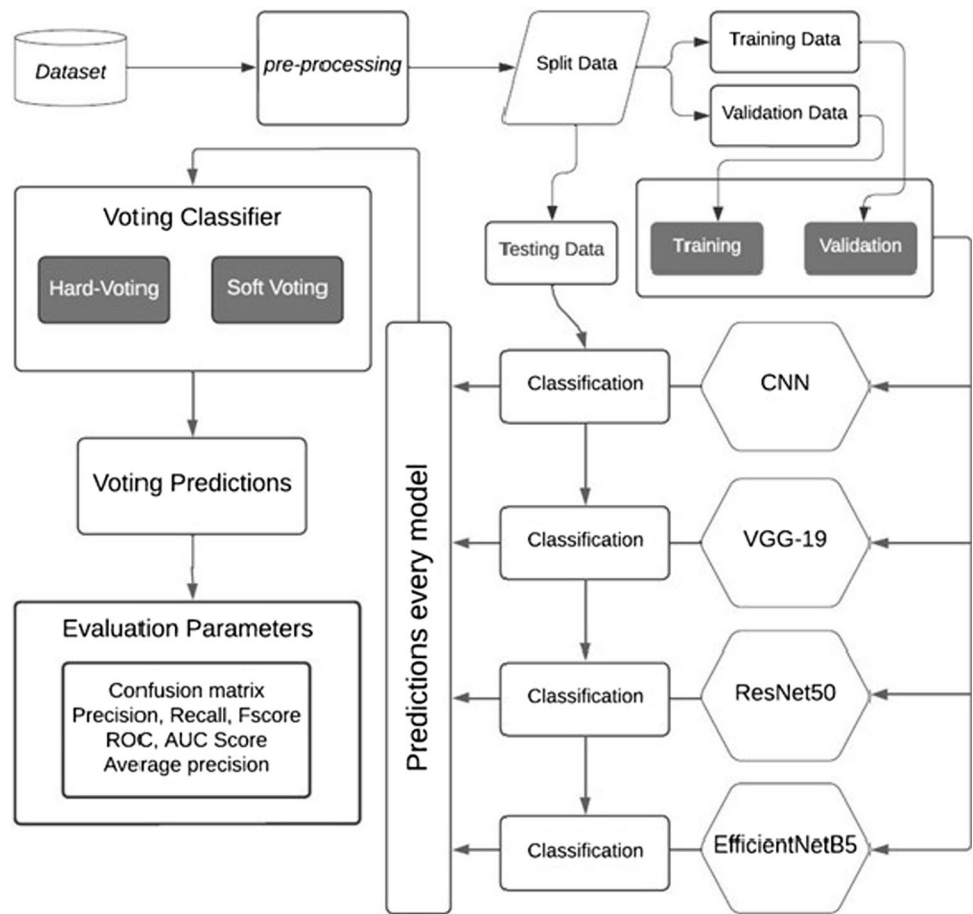
$$\text{Accuracy (Acc)} : \frac{((TP + TN))}{((TP + TN + FP + FN))} \quad (1)$$

Precision (Prec): precision is defined as the division of truly positive cases among all examples that we projected to be positive [24]:

$$\text{Precision (Prec)} : \frac{((TP))}{((TP + FP))} \quad (2)$$

Recall: the proportion of positive samples accurately identified as positive to the total number of positive models [25]:

Fig. 2 Workflow of research



$$\text{Recall} : \frac{((TP))}{((TP + FN))} \tag{3}$$

F1 score: the F1 score is the harmonic mean of precision and sensitivity [25]:

$$\text{F1 Score} : 2 \times \frac{(\text{Precision} \times \text{Recall})}{((\text{Precision} + \text{recall}))} \tag{4}$$

Area under the curve (AUC): The ROC curve is also known as (receiver operating characteristics). The area under the curve is an important performance metric that shows how effectively the model can differentiate between numerous classes. Remember that the higher this area, the more accurate the model for identifying [26]. The ROC curve is calculated using the true positive rate (TPR) and false positive rate (FPR).

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \tag{5}$$

where TP true positives, TN true negatives, FP false positives and FN false negatives.

## 4 Results and Discussion

The goal of this study is to identify ROP zones in preterm newborns. The efficacy of our models in distinguishing the three zones' categorization of ROP from fundus pictures was tested. The datasets were trained on four different classifiers to anticipate the trends in the data using precision, recall, the F1 measure, and the area under the curve. Finally, the voting classifier was utilized to integrate each classifier's output and offer an overall accuracy result.

### 4.1 Experimental Setup

We used an Intel Core i7 PC running at 2.7 GHz with 8 GB RAM for processing. Scikit-learn, an open-source ML package written in Python, is employed. Google Colab is an open-source web tool that allows us to create and share reports incorporating live code, graphics, equations, and narrated text.

### 4.2 Diagnosis of ROP by DL Models

Early and accurate diagnosis of ROP is essential for effective therapy. As a result, we utilized categorical categorization to ascertain whether the condition was moderate or severe. A timely diagnosis like that would enable optimal care to be provided. To do this, we trained models that showed promise in predicting ROP severity, including the VGG-19, ResNet-50, EfficientNetB5, and CNN. We fed the model the photos of the test cases and evaluated its performance based on its prediction.

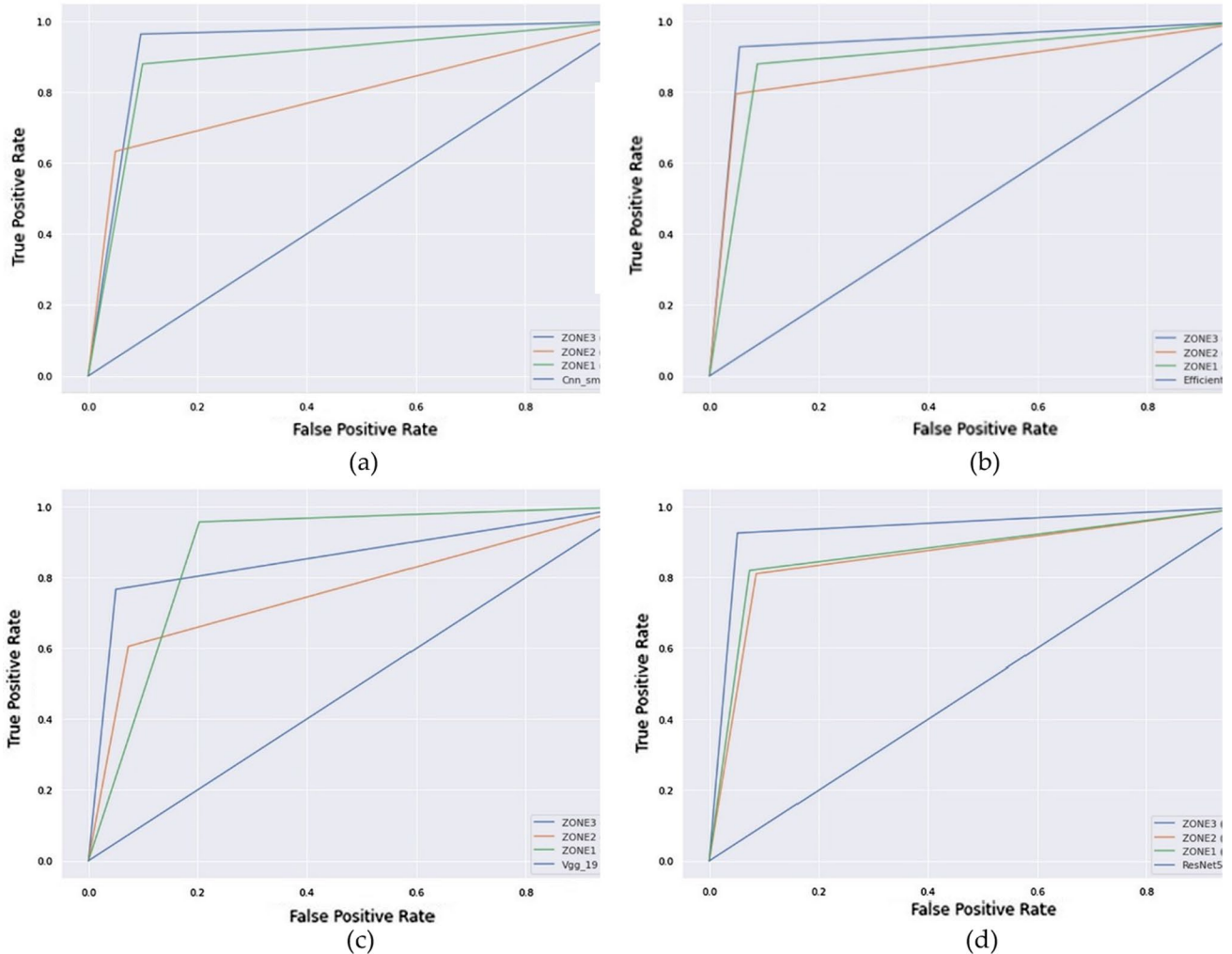
Our results show that EfficientNetB5 has outperformed other models in terms of accuracy with 87.27% and VGG-19 with the lowest score of 77.93%. In terms of precision, EfficientNetB5 has surpassed other models with 87.11% and VGG-19 with the lowest score of 79.06, as shown in Figure 3. Also, in terms of recall, EfficientNetB5 has again outperformed other models with 86.75% and VGG-19 with

lowest score of 77.65; furthermore, EfficientNetB5 shows better performance as compared to other models with 86.75% and VGG-19 with lowest score of 77.65 in terms of F1-score as shown in Table 3.

In the second step, the ensemble method proposed the voting classifier. We produced different deep learning classifiers and combined their results into a single verdict. In this study, a voting classifier was used, one type of ensemble

**Table 3** Performance evaluation of the deep neural networks (DNN) models

Method	Accuracy	Precision	Recall	F1-score
ResNet-50	85.71	85.2	85.18	85.18
VGG-19	77.93	79.06	77.65	77.07
EfficientNetB5	87.27	87.11	86.75	86.84
CNN	83.79	83.71	82.55	82.47



**Fig. 3** ROC-AUC curves of the deep learning classification models: **a** CNN, **b** EfficientNetB5, **c** VGG-19, **d** ResNet-50



method. This is achieved by employing several relatively DL methods on the same dataset [25, 26]. We used a voting classifier in this study, which means that each classifier (from the previous classification models) vote for each occurrence.

Moreover, half of all votes are cast for the most recent output projection. The case would be tagged with the more common class label; compare the models using the results tables. The performance evaluation using ROC curves, which demonstrates the consistency of the deep learning models, is shown in Fig 3.

The results show that the voting classifier enhanced the results and gave better prediction, i.e., precision: 88.6, Recall:88.03, F1 Score:88.17, and AUC: 91.20, as shown in Table 4. The reason for the better performance of the voting classifier technique is that it combines two or more DL techniques, and each classifier votes for a class and the class with the most votes win (Fig 4). A detailed comparison of the proposed method with other methods is shown in Table 4.

### 4.3 Limitations and Future Directions

Research studies must take into account their limitations and potential future paths. Several caveats in this study need discussion and potential attention in follow-up

**Table 4** Comparison of the proposed method (voting classifier) with other methods

Ref	Accuracy	Precision	Recall	F1-score
Pour et al. [27]	72.36	–	–	–
Coyner et al. [28]	87.5	86	76	–
A grawal et al. [14]	84.4	69.6	–	–
Proposed method (voting classifier)	91.2	88.6	88.03	88.17

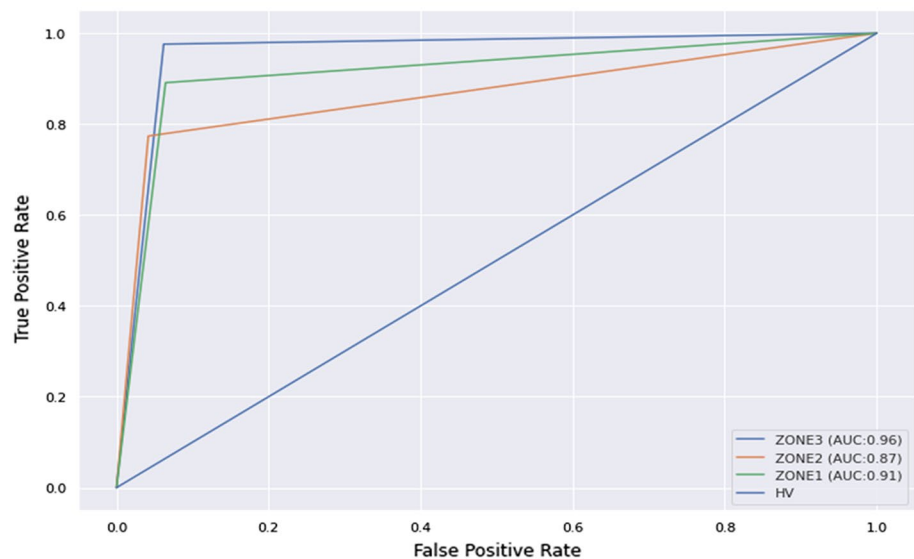
studies. One potential limitation of the study is that its sample size was restricted to a single geographical area. A larger representative sample may be used in subsequent studies to broaden the applicability of the findings. Second, self-reported information was used, which could introduce error and bias into the study. To corroborate the reliability of self-reported data, future studies may employ objective metrics, such as physiological or behavioral assessments.

## 5 Conclusion

This study presents the first use of transfer learning to a CNN to identify ROP illness in newborns automatically. In light of the scarcity of ophthalmologists and the prevalence of premature babies in rural regions, developing a system with high prediction accuracy is crucial. In this paper, we applied different classifiers (VGG-19, ResNet-50, EfficientNetB5, and CNN) to train a dataset for detecting the zones of ROP in preterm infants. The classification accuracies for these four different classifiers were 77.93, 83.79, 85.71, and 87.27%, and when using the voting classifier, the proposed model provided an overall accuracy of 88.82%. Moreover, the accuracy, recall, F1 measure, and area under the curve of the trained module were determined by testing. Then a voting classifier was applied to the total score derived from the evaluation matrices.

In the future, we will concentrate on the continued development of algorithms, more methodologies, and the production of a larger training dataset, all of which will aid in advancing medical reform in the current circumstances.

**Fig. 4** ROC-AUC curves of voting classifier



**Acknowledgements** The authors are grateful to the Private Clinic of Al-Amal Eye center in Baghdad, Iraq, for providing permission to use the data source.

**Authors contribution** Conceptualization: NS, MK, and NH; methodology: NS, MK, and NH.; software: NS; validation: NS, MK, and NH; formal analysis: NS, MK, and NH; investigation: NS, MK, and NH; resources: NH, AAR and SA; data curation: AAR, and SA; writing—original draft preparation: NS; writing—review and editing: NS; visualization: MK, NH and DBH; supervision: MK, NH and DBH; project administration: MK. All authors have read and agreed to the published version of the manuscript.

**Funding** This research received no external funding.

**Data availability** Datasets utilized with the author's permission.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. 'A Weighted Voting Framework for Classifiers Ensembles | SpringerLink'. Accessed 7 December 2022 <https://link.springer.com/article/https://doi.org/10.1007/s10115-012-0586-6>.
2. Ranjana, A., Kulkarni, S., Walambe, R., Kotecha, K.: Assistive framework for automatic detection of all the zones in retinopathy of prematurity using deep learning. *J. Dig. Imag.* **34**(4), 932–947 (2021). <https://doi.org/10.1007/s10278-021-00477-8>
3. Garner, A., Ben-Sira, I., Konen, W., Majima, A., McCormick, A., Mushin, A.: An International classification of retinopathy of prematurity. *Pediatrics* **74**(1), 127–133 (1984)
4. Aaberg, T.: An International classification of retinopathy of prematurity. II. The classification of retinal detachment. The International Committee for the classification of the late stages of Retinopathy of Prematurity. *Archiv Ophthalmol.* **105**(7), 906–912 (1987)
5. 'Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance | IEEE Journals & Magazine | IEEE Xplore'. Accessed 7 December 2022 <https://ieeexplore.ieee.org/document/618255>.
6. Lei, B., Zeng, X., Huang, S., Zhang, R., Chen, G., Zhao, J., Wang, T., Wang, J., Zhang, G.: Automated detection of retinopathy of prematurity by deep attention network multimedia tools and applications. *Multimed. Tools. Appl.* (2022). <https://doi.org/10.1007/s11042-021-11208-0>
7. Bao, Y., Ming, W.-K., Mou, Z.-W., Kong, Q.-H., Li, A., Yuan, T.-F., Mi, X.-S.: Current application of digital diagnosing systems for retinopathy of prematurity. *Comput. Methods. Prog. Biomed.* (2020). <https://doi.org/10.1016/j.cmpb.2020.105871>
8. Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michałowski, Ł., Paluszkiwicz, R., Ziarkiewicz-Wróblewska, B., Zieniewicz, K., Sobieraj, P., Nowicki, A.: Transfer Learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *Int J Comput Assist Radiol Surg* **13**(12), 1895–1903 (2018). <https://doi.org/10.1007/s11548-018-1843-2>
9. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**(1), 6 (2020). <https://doi.org/10.1186/s12864-019-6413-7>
10. Coyner, A.S., Peter Campbell, J., Ostmo, S., Kim, S.J., Jonas, K.E., Paul Chan, R.V., Chiang, M.F.: Machine learning for prediction of retinopathy of prematurity fundus image quality from clinical data. *Investigat. Ophthalmol. Vis. Sci.* **60**(9), 1525 (2019)
11. 'Detecting Fake News in Social Media Using Voting Classifier | IEEE Journals & Magazine | IEEE Xplore'. Accessed 19 November 2022 <https://ieeexplore.ieee.org/document/9632574>.
12. Early Treatment for Retinopathy of Prematurity Cooperative Group: The incidence and course of retinopathy of prematurity: findings from the early treatment for retinopathy of prematurity study. *Pediatrics* **116**(1), 15–23 (2005). <https://doi.org/10.1542/peds.2004-1413>
13. Gensure, R.H., Chiang, M.F., Campbell, J.P.: Artificial intelligence for retinopathy of prematurity. *Curr. Opin. Ophthalmol.* **31**(5), 312–317 (2020). <https://doi.org/10.1097/ICU.00000000000000680>
14. Goldenberg, R.L., Culhane, J.F., Iams, J.D., Romero, R.: Epidemiology and causes of preterm birth. *Lancet* **371**(9606), 75–84 (2008). [https://doi.org/10.1016/S0140-6736\(08\)60074-4](https://doi.org/10.1016/S0140-6736(08)60074-4)
15. Huang, Y.-P., Vadloori, S., Chu, H.-C., Kang, E.-C., Wei-Chi, Wu., Kusaka, S., Fukushima, Y.: Deep learning models for automated diagnosis of retinopathy of prematurity in preterm infants. *Electronics* **9**(9), 1444 (2020). <https://doi.org/10.3390/electronic9091444>
16. International Committee for the Classification of Retinopathy of Prematurity: The International classification of retinopathy of prematurity revisited. *Archiv. Ophthalmol.* **123**(7), 991–999 (2005). <https://doi.org/10.1001/archophth.123.7.991>
17. Martin, J.A., Kochanek, K.D., Strobino, D.M., Guyer, B., MacDorman, M.F.: Annual summary of vital statistics--2003. *Pediatrics* **115**(3), 619–634 (2005). <https://doi.org/10.1542/peds.2004-2695>
18. Supriti, M., Ram, K., Sivaprakasam, M., Vinekar, A.: Early detection of retinopathy of prematurity stage using deep learning approach. *Med. Imag. Comput. Aid. Diagn.* **10950**, 758–764 (2021). <https://doi.org/10.1117/12.2512719>
19. Pour, E.K., Pourreza, H., Zamani, K.A., Mahmoudi, A., Sadeghi, A.M.M., Shadravan, M., Karkhaneh, R., Pour, R.R., Esfahani, M.R.: Retinopathy of prematurity-assist: novel software for detecting plus disease. *Korean J. Ophthalmol.* **31**(6), 524 (2017). <https://doi.org/10.3341/kjo.2015.0143>
20. Rashaed, S.A.: Retinopathy of prematurity—a brief review. *Dr Sulaiman Al Habib Med. J.* **1**(3–4), 58–64 (2019). <https://doi.org/10.2991/dsahmj.k.191214.001>
21. Rolls, E.T.: The storage and recall of memories in the Hippocampo-cortical system. *Cell Tissue Res.* **373**(3), 577–604 (2018). <https://doi.org/10.1007/s00441-017-2744-3>

22. Scruggs, B.A., Paul Chan, R.V., Kalpathy-Cramer, J., Chiang, M.F., Peter Campbell, J.: Artificial intelligence in retinopathy of prematurity diagnosis. *Trans. Vis. Sci. Technol.* **9**(2), 5 (2020). <https://doi.org/10.1167/tvst.9.2.5>
23. Terry, T.L.: Extreme prematurity and fibroblastic overgrowth of persistent vascular sheath behind each crystalline lens\* \*from the massachusetts eye and ear infirmary. This investigation is made possible through the special fund for research for pathology laboratory. *Am. J. Ophthalmol.* **25**(2), 203–204 (1942). [https://doi.org/10.1016/S0002-9394\(42\)92088-9](https://doi.org/10.1016/S0002-9394(42)92088-9)
24. Tong, Y., Wei, Lu., Deng, Q.-Q., Chen, C., Shen, Y.: Automated identification of retinopathy of prematurity by image-based deep learning. *Eye Vis.* **7**(1), 40 (2020). <https://doi.org/10.1186/s40662-020-00206-2>
25. Vijayalakshmi, C., Sakthivel, P., Vinekar, A.: Automated detection and classification of telemedical retinopathy of prematurity images. *Telemed. J. E-Health Off. J. Am. Telemed. Assoc.* **26**(3), 354–358 (2020). <https://doi.org/10.1089/tmj.2019.0004>
26. Wang, J., Rong, Ju., Chen, Y., Zhang, L., Junjie, Hu., Yu, Wu., Dong, W., Zhong, J., Yi, Z.: Automated retinopathy of prematurity screening using deep neural networks. *EBioMedicine* **35**, 361–368 (2018). <https://doi.org/10.1016/j.ebiom.2018.08.033>
27. Worrall, Daniel E., Clare M. Wilson, and Gabriel J. Brostow. 'Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks'. In: *Deep Learning and Data Labeling for Medical Applications*, edited by Gustavo Carneiro, Diana Mateus, Loïc Peter, Andrew Bradley, João Manuel R. S. Tavares, Vasileios Belagiannis, João Paulo Papa, et al., 10008:68–76. *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2016. Accessed 12 February 2021 doi:[https://doi.org/10.1007/978-3-319-46976-8\\_8](https://doi.org/10.1007/978-3-319-46976-8_8).
28. Jafar, Z.S.A., Buriro, A., Riaz, M., Mahboob, A., Riaz, M.N.: Implementation and comparison of text-based image retrieval schemes. *Int J Advan Comput Sci Appl (IJACSA)* **10**(1), 611–618 (2022). <https://doi.org/10.14569/IJACSA.2019.0100177>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.