Received: 2 January 2022 | Accepted: 3 May 2022

DOI: 10.1111/srt.13166

AcneGrader: An ensemble pruning of the deep learning base models to grade acne

Shuai Liu[1] | Yusi Fan[2] | Meiyu Duan[1] | Yueying Wang[1] | Guoxiong Su[3] | Yanjiao Ren[4] | Lan Huang[1] | Fengfeng Zhou[1]

[1]College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, P.R. China

[2]College of Software, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, P.R. China

[3]Beijing Dr. of Acne Medical Research Institute, Beijing, China

[4]College of Information Technology (Smart Agriculture Research Institute), Jilin Agricultural University, Changchun, Jilin, China

Correspondence
Fengfeng Zhou, College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, P.R. China.
Email: FengfengZhou@gmail.com; ffzhou@jlu.edu.cn

Funding information
Jilin Senior and Junior Technological Innovation Team, Grant/Award Number: 20210509055RQ; National Natural Science Foundation of China, Grant/Award Numbers: 62072212, U19A2061; Jilin Provincial Key Laboratory of Big Data Intelligent Computing, Grant/Award Number: 20180622002JC; Science and Technology Project of Education Department of Jilin Province, Grant/Award Number: JJKH20200328KJ; Fundamental Research Funds for the Central Universities

Abstract

Background: Acne is one of the most common skin lesions in adolescents. Some severe or inflammatory acne leads to scars, which may have major impacts on patients' quality of life or even job prospects. Grading acne plays an important role in diagnosis, and the diagnosis is made by counting the number of acne. It is a labor-intensive job and it is easy for dermatologists to make mistakes, so it is very important to develop automatic diagnosis methods. Ensemble learning may improve the prediction results of the base models, but its time complexity is relatively high. The ensemble pruning strategy may solve this computational challenge by removing the redundant base models.

Materials and methods: This study proposed a novel ensemble pruning framework of deep learning models to accurately detect and grade acne using images. First, we train multi-base models and prune the redundancy models according to the performance and diversity of the models. Then, we construct the new features of the training data by the base models we select in the previous step. Next, we remove the redundancy models further by a feature selection algorithm. Finally, we integrate all the base models by classifiers. The ensemble pruning algorithm was proposed to prune the deep learning base models.

Results: The experimental data showed that the ensemble pruned framework achieved a prediction accuracy of 85.82% on the acne dataset, better than the existing studies. To verify our method's effectiveness, we test our method in a skin cancer dataset and greatly outperform the state-of-the-art methods.

Conclusion: The method we proposed is used to grade acne. Our method's performance outperforms state-of-the-art methods on two datasets, and it can also remove redundancy models to reduce computational complexity.

KEYWORDS
acne, acne grade, deep learning, ensemble classification, ensemble pruning

1 | INTRODUCTION

Acne is also known as acne vulgaris and is a commonly occurred skin lesion worldwide, especially in adolescents.[1,2] Acne affects 79%—95% of adolescents in western countries.[3] Additionally, 51.3% of the total adolescents in northeastern China are affected by acne.[4] Although acne vulgaris is not lethal, it may substantially influence patients' quality of life, self-esteem, mental mood, and psychological disorders.[5–7] Besides adolescents, adults may also suffer from acne.[8,9] The Hayashi criterion is usually used to grade acne severity by dermatologists,[10] and there are four severity levels of acne mainly based on the numbers of lesions, that is, mild, moderate, severe, and very severe.[11]

A graph neural network (GNN) is a type of deep learning algorithm in non-Euclidean space. They are widely used for many bioinformatics and medical image analysis studies.[12,13] A dual graph convolutional network was proposed to predict the chemical network.[14] GNNs were also utilized to predict drug–target interactions[15] and protein functions.[16] Image biomarker analysis is an important research topic. GNNs have demonstrated very powerful performance in this field.[12,17] GNNs have also been extensively used for segmentation and lesion detection tasks on biomedical images.[18,19]

Quite a few computational studies have been published to detect and classify acne by handcrafted features[20,21] or deep learning methods.[22–25] Currently, many methods detect, classify, or count acne by handcrafted features.[26–28] Deep learning algorithms are also widely used in the medical image analysis field and show impressive performance improvements. The quality of training samples is essential to generate a good acne prediction model, and ensemble algorithms tend to deliver stably accurate models by summarizing the prediction results of multiple base models.[29–33]

Ensemble algorithms are one of the widely used machine learning algorithms, such as random forest (RF).[34–36] Ensemble algorithms may achieve much improved prediction performances by integrating the results of multiple base models, which are usually weak models. Deep learning algorithms demonstrate powerful prediction performances in medical image analysis and may also serve as the base models of ensemble methods.[37] However, both ensemble algorithms and deep learning algorithms are notorious for their high computational complexities. The ensemble method may train the base models by randomly selecting training subsets. It is anticipated that there may be redundancy between the base models in the ensemble algorithms. To address this issue, ensemble pruning algorithms are used to remove the redundant models.[38–40] Recently, several works have leveraged deep learning ensemble pruning. Hu et al.[41] proposed a deep ensemble pruning algorithm to provide high-quality uncertainty prediction. Rajaraman et al.[42] proposed a deep learning ensemble pruning method for COVID-19 detection in chest X-rays. Zhang et al.[43] proposed a boosting deep learning-based ensemble model and a novel ensemble pruning method for time series forecasting. There is no work to grade acne by an ensemble pruning algorithm. In traditional ensemble pruning algorithms, the performance and diversity of the base models play important roles in pruning base models, and simple model integration strategies, such as voting are widely used, but two major issues remain: (1) quantitatively measuring the contribution of each base model to the ensemble framework, and (2) in a model-based integration strategy, whether the performance and diversity of base mode pruning strategies are better pruning strategies.

This study aims to solve these issues by proposing a novel classification framework. Our framework consists of three modules. First, we train multiple base models and use the ensemble pruning method to remove redundant base models. We prune the base model by the value of Kappa, which means the diversity of different models; in this way, we can select the model subset with the highest diversity. Then, we use the selected base models to predict the images from the training set, and the prediction results are regarded as the new features for the ensemble framework. In this way, we can represent the base models by the prediction, and we will concatenate all the results from all the base models; a vector can represent an image. We can ensemble models and pruning models at the feature level in the next step. Next, we use the feature selection algorithms to remove the features with the least impacts on the final result, that is, the feature selection methods are used to remove the redundant base models. Finally, we use a classifier to ensemble the base models we selected; in this way, we can give different weights to different base models by training a classifier.

Our contributions are threefold. (1) We propose a novel ensemble classification framework to classify acne severity. (2) An ensemble pruning strategy is designed to reduce the computational complexity of the trained ensemble classification model. The prediction results of the base models are regarded as a new set of features, and feature selection algorithms are used to find the best feature subset. (3) We use the prediction results of the base models as a new set of features and use the classifier to ensemble the results of the base models. The experimental data showed that the proposed AcneGrader outperformed the state-of-the-art methods on classifying the four acne grades.

This manuscript is organized as follows. The background and motivations are described in this section. Section 2 provides detailed descriptions of the data and algorithms evaluated in this study. The experimental data and result discussion are given in Sections 3 and 4. Concluding remarks and limitations are discussed in Section 5.

2 | MATERIALS AND METHODS

2.1 | Dataset

The ACNE04 dataset[11] was used to evaluate the proposed algorithm AcneGrader for detecting and grading the acne. This dataset annotated the local lesions and global acne severity using the Hayashi criterion by professional dermatologists.[44] There are 1457 images in this ACNE04 dataset. A fivefold stratified cross-validation strategy was used to evaluate the prediction algorithms. That is, 80% of randomly retrieved samples were used to train the model, and the remaining 20% were used to test the model. Each of the fivefolds was used as the test dataset iteratively. Dataset splitting was already carried out and released by the maintainer of the ACNE04 dataset.

To further verify the effectiveness of our model, a skin cancer dermoscopic image dataset is used to verify our model.[45] This dataset contains 3297 images; 2637 images are used to train the model, and 660 images are used to test the model. This dataset contains two categories, benign and malignant. The training set contains 1440 benign skin cancer images and 1197 malignant skin cancer images. The test set contains 360 benign skin cancer images and 300 malignant skin cancer images.

2.2 | Performance evaluation metrics

This study formulated the acne grading problem as a four-class classification problem. Acne was graded as four levels, that is, mild, moderate, severe, and very severe.[11] The prediction performances of each class were evaluated by the class-specific one-versus-others sensitivity (Sn), specificity (Sp), and precision (Pr). The two metrics Sn and Sp were the prediction accuracies of positive and negative samples, respectively. The metric Pr was the rate of correctly predicted positives among the positive predictions. The overall Sn, Sp, and Pr were defined as the average values of the four class-specific Sn, Sp, and Pr, respectively. The overall accuracy (Acc) was the percentage of correctly predicted samples among all the samples. The metric Youden index (YI) was another popular classification performance metric and defined as (Sn + Sp - 1), and a larger YI value suggested a better classification performance of the investigated model.

In the binary classification task, accuracy, precision, recall, and F1-score are used to measure our method's performance. The accuracy is the percentage of correctly predicted samples among all the samples, precision is the proportion of the number of correctly predicted samples in the total number of positive predictions, and recall is the proportion of the number of positive samples predicted as positive samples in all labeled samples. The F1-score is an evaluation index of balanced recall and precision and is defined as 2 × (Pr × recall)/(Pr + recall); a larger F1-score suggests a better classification performance of the investigated model.

2.3 | Experimental procedure

The experimental procedure of this study is illustrated in Figure 1. Firstly, we trained some base models using neural networks proposed by Wu et al.[11] In this study, we randomly select image subsets and train N base models by the image subsets. Secondly, the ensemble pruning algorithm was used to remove the redundancy models, and we remove (N - n) redundancy base models and obtain n base models. Thirdly, the base models we select are used to construct features. For instance, the training dataset consists of m images. We use n base models to predict images, and the prediction results are used as the features. In other words, each image is represented by an n-dimensional vector, and each feature corresponds to a base model. Next, feature selection algorithms are used to remove the redundancy features constructed in the previous step because each feature corresponds to a base model; in this way, we can use feature selection algorithms to remove the redundancy base models, and we can obtain better performance. Finally, some simple ensemble strategies, such as voting, are not used, and we summarize the final results by classifier; in this way, we can integrate the results nonlinearly.

2.4 | Ensemble learning

Ensemble learning is a machine learning strategy to combine the prediction results of multiple base models, and there are three main categories of supervised ensemble learning strategies, that is, bagging, boosting, and stacking. The bagging strategy is also called the bootstrap aggregating strategy and trains base models using randomly drawn sample subsets for equal-weight ensemble voting.[46–49] The boosting strategy incrementally builds the ensemble framework by training new base models with emphasis on the misclassified samples by the previously trained base models.[50–52] The boosting ensemble model may deliver better prediction performances than the bagging strategy but may also induce overfitting over the outlier samples.[53] The stacking strategy trained a meta-learner to optimally combine the prediction results of base models and relied on a sufficient number of samples to train the meta-learning step.[54]

This study used the bagging strategy to ensemble the base models. We randomly drew 30 training subsets to build 30 diversified base models. Four strategies were used to combine the prediction results of the 30 base models, that is, voting, average, sigmoid average, and weight average strategies. The voting strategy determined the final prediction by the majority of the prediction results of the base models. In the average strategy, the ensemble result was calculated by the mean of those of all the base models. In the sigmoid averaging strategy, we normalized the results of all 30 base models and output the mean value as the ensemble result. The weight average strategy assigned the validation error as the weight for each base model and calculated the weighted result as the output.

2.5 | Ensemble pruning

The ensemble pruning strategy reduces the size of the bagging ensemble model to increase the model efficiency. Two ensemble pruning strategies were used in this study. Firstly, the base models were evaluated for their validation error rates, and the top-ranked models with small error rates were selected, called error rate pruning (ErrorRate).[55] We calculated the accuracies of different models in the validation dataset and ranked the models according to the accuracies. The models with small accuracies were removed from the ensemble framework. Secondly, inter-model diversity was increased by removing the model with a correlated but better model, called diversity pruning. Three metrics were used to measure the diversity between two base models, that is, disagreement measure,[56] double-fault (DF) measure,[57] and Kappa statistic.[58]

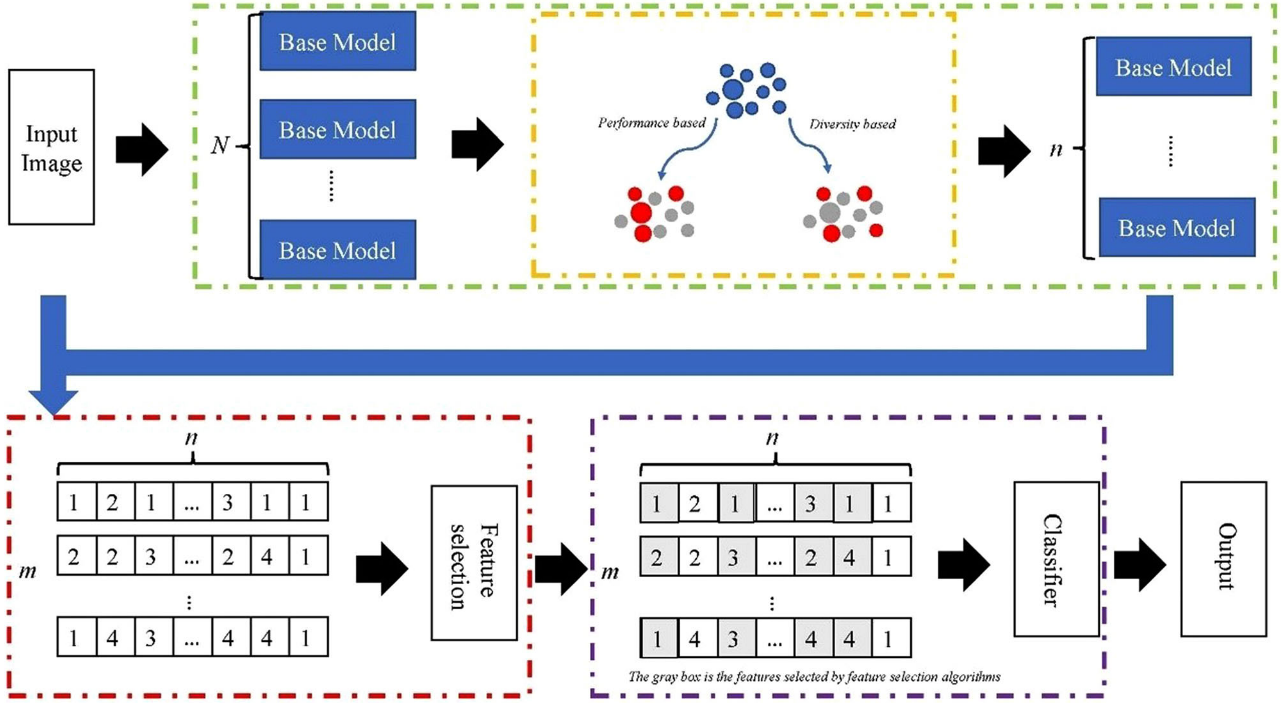

FIGURE 1 Experimental design of this study

The disagreement metric (DisM)[56] between the ith and jth base models is calculated as:

$$\mathrm{DisM}(i,j) = \frac{a}{m} \tag{1}$$

The variable a is the number of samples that the ith model and jth model assign to different classes, and the variable m is the number of all the samples. The higher DisM(i, j) is, the more diversified the two models are to each other. The proposed ensemble framework tries to select the diversified subset of models. Each model pair is evaluated for their disagreement DisM(i, j), and the model pairs with the top-ranked DisM values are chosen for further analysis.

The DF metric[57] between the ith and jth base models is calculated as:

$$\mathrm{DF}(i,j) = \frac{e}{m} \tag{2}$$

The variable m is the number of all the samples, and e is the number of the samples that neither the ith nor the jth base models assigned to the wrong classes. We rank the model pairs according to DF(i, j) and select models in the top-ranked model pairs until we find the model subset with satisfying prediction accuracies.

Kappa statistics (Kappa)[58] is used for the consistency test and may also be used to measure the classification accuracy. Kappa is calculated as:

$$\mathrm{Kappa}(i,j) = \frac{P(A) - P(E)}{1 - P(E)} \tag{3}$$

The variable $P(A)$ is the percentage agreement between the ith and jth models. $P(E)$ is the chance agreement. Kappa(i, j) = 1 indicates perfect agreement between the two models, and Kappa(i, j) = 0 indicates no agreement. We ranked the model pairs according to Kappa(i, j) to use their ascending numerical sort and select the models in the top-ranked model pairs until we find the model subset with satisfying prediction accuracies.

2.6 | Feature selection algorithms and classifiers

Ensemble pruning algorithms are used to remove redundancy in the ensemble classification framework. Most of the existing ensemble pruning strategies remove the redundancies between base models according to their performance and diversity. This study uses feature selection algorithms to remove redundant base modes. Firstly, we used the base models to predict the images in the training set and testing set, and the prediction results by the base models were used as the generated features. Secondly, we used feature selection algorithms to remove the redundant features. In this study, we used the recursive feature elimination (RFE) strategy to select the features, and the classifier support vector machine (SVM) was used to evaluate a given feature subset. We eliminated the features utilizing which we obtained the best performance. Finally, we trained an SVM model by the selected features.

2.7 | Implementation details

The base model of our framework is a convolutional neural network based on the architecture from Wu et al.[11] This study evaluates the

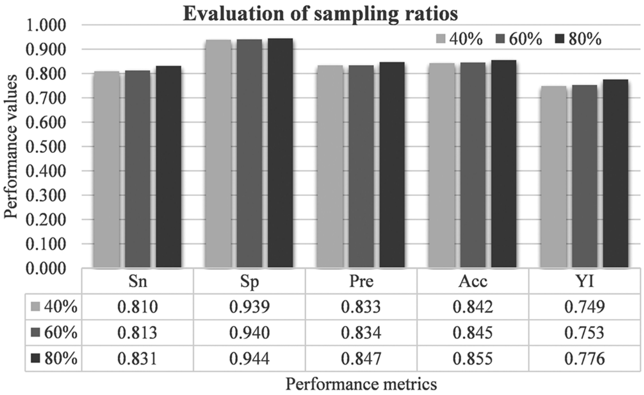

FIGURE 2 Performance comparison of different sampling ratios

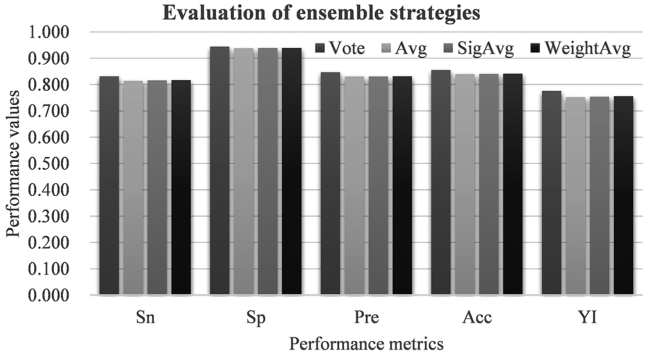

FIGURE 3 Performance comparison of different ensemble strategies

acne grading algorithm using the ACNE04 dataset, which was released by Wu et al.[11] We hypothesized that the neural network optimized by the authors releasing the ACNE04 dataset may be a good candidate as the base model for our ensemble framework. We choose the default values of the hyperparameters of the neural network from Wu et al.[11] That is, stochastic gradient descent with a mini-batch of 32 is used to optimize the base model in our framework. The experiment is trained for 120 epochs. The momentum and weight decay values are set to 0.9 and 5e-4, respectively. The learning rate starts at 0.001 and decays it by 0.5 every 30 epochs. The proposed framework AcneGrader is an ensemble learner, and the base models are trained using different sampled subsets from the training dataset. Our algorithm runs on an NVIDIA P100 GPU with 16 GB VRAM.

3 | RESULTS

3.1 | Evaluation of different sampling ratios of the bagging strategy

This study used the bagging strategy to train the ensemble classification model. The bagging stage needed to draw a sample subset from the training dataset. Therefore, this section evaluates how the sampling ratio affects the ensemble classifier, as shown in Figure 2. Three values were evaluated, that is, 40%, 60%, and 80%. The experimental data showed a positive correlation between the sampling ratio and the model performance in all five metrics. The overall accuracy reached 0.855 for the 80% sampling ratio. The other four performance metrics also reached the best values: Sn = 0.831, Sp = 0.944, Pr = 0.847, and YI = 0.776. Therefore, the following sections of this study used 80% of the sampling ratio for the bagging strategy to train the base models. As the results show, more samples randomly retrieved from the training dataset generated better prediction performances.

3.2 | Evaluation of different ensemble strategies

We also evaluated four ensemble strategies for the acne grade classification problem, as shown in Figure 3. The average strategy achieved the worst values of all five performance metrics. The sigmoid average and weight average strategies slightly improved these five performance metrics compared with the average strategy. The voting strategy achieved the best performance, with the best Acc = 0.855. Therefore, the voting strategy will be used in the following sections.

3.3 | Evaluation of different ensemble pruning algorithms

The default number of base models trained in an ensemble algorithm does not always deliver the best prediction performances, and some base models may be removed to improve the ensemble model performances. The ensemble pruning algorithm may remove the redundant base models in an ensemble learning framework.

Four ensemble strategies were evaluated for their performance effects using an incremental model selection algorithm, as shown in Figure 4. The four ensemble pruning algorithms are error rate pruning,[55] disagreement metric,[56] DF metric,[57] and Kappa statistic.[58] The best accuracies were achieved by the total number (30) of base models for the disagreement metric (Acc = 0.855) and the DF metric (Acc = 0.855) strategies. That is, the ensemble pruning algorithm based on these two strategies did not improve the ensemble learning model. It is interesting to see that the error rate pruning strategy recommended seven models to achieve the best Acc = 0.857, while the Kappa statistics recommended 22 base models to achieve Acc = 0.858. Therefore, the ensemble pruning algorithm improved the ensemble learning model, and the best model reached Acc = 0.858 using 22 base models. If the proposed AcneGrader model is deployed to a cost-sensitive platform, such as mobile phones, the ensemble model improved by the error rate pruning strategy may be chosen with a slightly reduced Acc = 0.857.

3.4 | Evaluation of pruning different numbers of base models

We further evaluated how much the ensemble pruning algorithms improved the ensemble learning algorithms initiated with different

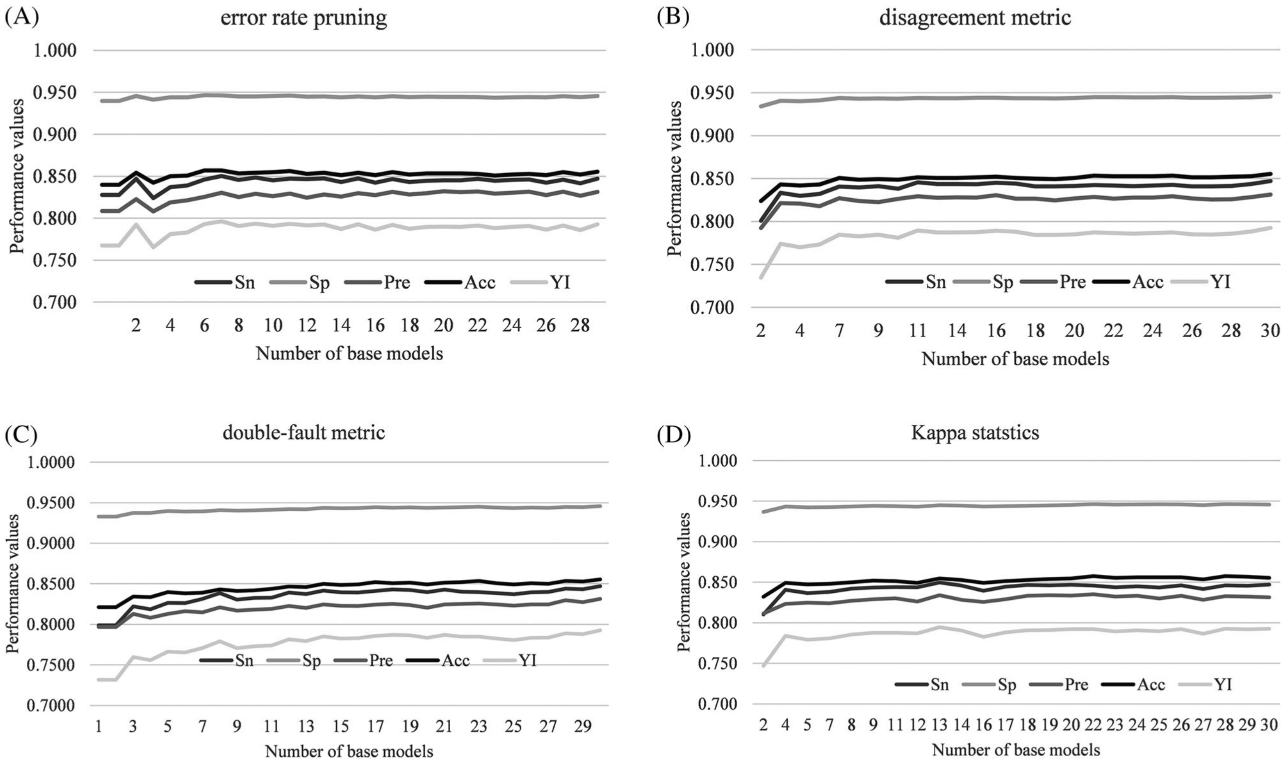

FIGURE 4 Performance comparison of the incremental model selection using different ensemble pruning algorithms. (A) The error rate pruning, (B) disagreement metric, (C) double-fault metric, and (D) Kappa statistics algorithms

numbers of base models, as shown in Figure 5. The ensemble pruning algorithms ErrorRate and DisM removed one redundant base model while achieving the same Acc = 0.857, but the overall accuracies were the same without any improvements from the four ensemble pruning algorithms, as shown in Figure 5A. When there were 20 or 30 base models, the ensemble pruning algorithm Kappa always achieved the best improvements in Acc, as shown in Figure 5B,C. The overall best model was still the ensemble learning model (Acc = 0.858) initiated with 30 base models and pruned to 22 models. The experimental data in Figure 5 further supported the necessity of pruning the ensemble learning algorithms initiated with different numbers of base models. As the results show, a larger number of base models may generate better prediction performances. The choice of the ensemble pruning method is another important factor to improve the ensemble prediction model.

3.5 | Performance comparison between different classifiers and feature selection algorithms

In this study, we used feature selection algorithms to select the base models and evaluated the prediction results of the sets of different base models. Three feature selection algorithms and five classifiers were utilized to select the best sets of base models and to predict the results. The feature selection algorithms were RFE, variance, and chi-square. The five classifiers were SVM, decision tree (DT), RF, *K*-nearest neighbor, and naïve Bayes. The RFE algorithm recursively eliminated features according to the features' weights, and only these three classifiers, SVM/DT/RF, provided such information. As shown in Table 1, the SVM–RFE achieved the best performance according to Acc in all the groups (Acc = 0.8582). Therefore, the feature selection algorithm RFE

TABLE 1 Comparison of the performance of different groups with different feature selection (FS) algorithms and classifiers

FS	Classifier	Pr	Sn	Sp	YI	Acc
RFE	SVM	**0.8556**	0.8371	**0.9455**	0.7826	**0.8582**
	DT	0.8386	0.8256	0.9419	0.7675	0.8473
	RF	0.842	0.8285	0.9432	0.7718	0.8514
Variance	RF	0.8425	**0.8404**	0.9431	**0.7835**	0.8514
	DT	0.8389	0.8319	0.9406	0.7726	0.8452
	SVM	0.8419	0.8371	0.9445	0.7816	0.8541
	KNN	0.842	0.8297	0.9446	0.7743	0.8548
	NB	0.8484	0.8271	0.9443	0.7714	0.8541
Chi2	RF	0.8436	0.835	0.944	0.779	0.8534
	NB	0.8457	0.8263	0.9443	0.7706	0.8534
	KNN	0.8447	0.8333	0.9454	0.7787	0.8568
	SVM	0.8407	0.8353	0.944	0.7793	0.8527
	DT	0.831	0.8221	0.9423	0.7645	0.8473

Abbreviations: Acc, accuracy; Chi2, chi-square; DT, decision tree; KNN, *K*-nearest neighbor; NB, naïve Bayes; Pr, precision; RF, random forest; RFE, recursive feature elimination; Sn, sensitivity; Sp, specificity; SVM, support vector machine; YI, Youden index.
The significance of bold values is the biggest values in the same column.

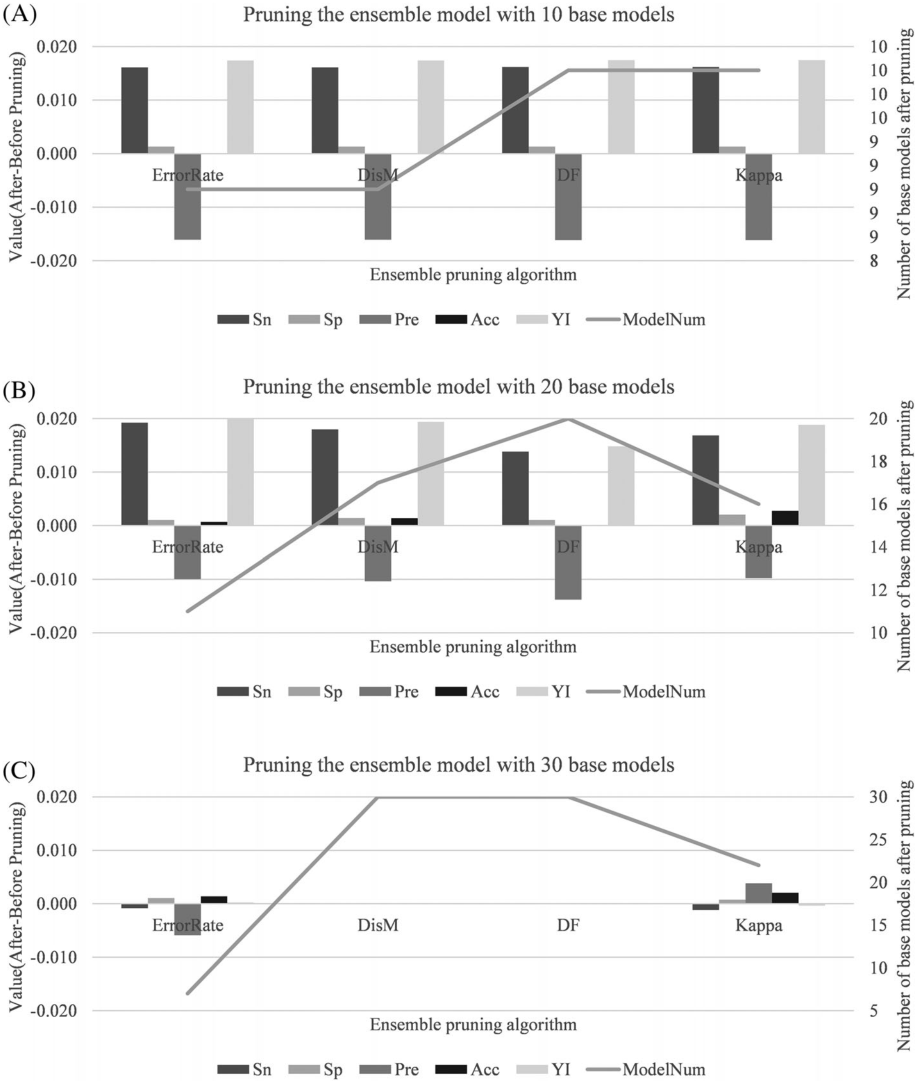

FIGURE 5 How the ensemble learning algorithms with different numbers of base models were affected by the ensemble pruning algorithms. The experiments were about the AcneGrader algorithms using (A) 10, (B) 20, and (C) 30 base models

and the classifier SVM were used to summarize the final result in the following sections of this study.

3.6 | Evaluating the influence of ensemble pruning and feature selection

The proposed AcneGrader framework consisted of the ensemble pruning module and the feature selection module. The ablation experiment was carried out to evaluate the contributions of these two modules. As shown in Table 2, when we removed the feature selection module, the accuracy was decreased by 0.07, and the number of base models was increased to 22 compared with the eight base models used by the AcneGrader framework. When we removed the ensemble pruning module, although we used eight base models to achieve the best performance, each performance metric was smaller than AcneGrader. Overall, the feature selection module and ensemble pruning module played important roles in delivering the best AcneGrader system.

TABLE 2 Verification of the importance of the feature selection (FS) module and ensemble pruning module

	Pr	Sn	Sp	YI	Acc	Num
Without FS	83.52	**94.64**	84.6	**79.23**	85.75	22
Without pruning	85.14	83.49	94.34	77.83	85.34	**8**
AcneGrader	**85.56**	83.71	**94.55**	78.26	**85.82**	**8**

Abbreviations: Acc, accuracy; Pr, precision; Sn, sensitivity; Sp, specificity; YI, Youden index.
The significance of bold values is the biggest values in the same column.

3.7 | Comparison with the state-of-the-art results

A performance comparison was carried out between our algorithm AcneGrader and the model from the same group releasing the ACNE04 dataset,[11] as shown in Table 3. The proposed algorithm AcneGrader outperformed the model from Wu et al.[11] by 1.71% in Acc, 2.69% in Sn, 0.75 in Sp, 1.19 in Pr, and 2.94% in YI. Overall, our AcneGrader

TABLE 3 Comparing our algorithm AcneGrader with the state-of-the-art result on the ANCE04 dataset

Metric	Sn	Sp	Pr	Acc	YI
VGGNet[59]	72.71 ± 2.60	90.60 ± 0.71	72.65 ± 3.42	75.17 ± 1.97	63.31 ± 3.19
Inception[60]	72.77 ± 2.61	90.95 ± 0.68	74.26 ± 3.26	76.44 ± 1.77	63.72 ± 2.92
ResNet[61]	75.36 ± 3.39	91.85 ± 0.77	75.81 ± 2.56	78.42 ± 2.11	67.21 ± 4.11
DenseNet[62]	68.64 ± 3.79	90.81 ± 0.79	75.84 ± 2.20	76.58 ± 1.77	59.45 ± 4.37
Wu et al.[11]	81.52 ± 0.02	93.80 ± 0.00	84.37 ± 0.02	84.11 ± 0.01	75.32 ± 0.02
AcneGrader (ours)	**83.71 ± 0.03**	**94.55 ± 0.05**	**85.56 ± 0.07**	**85.82 ± 0.02**	**78.26 ± 0.06**

Abbreviations: Acc, accuracy; Pr, precision; Sn, sensitivity; Sp, specificity; YI, Youden index.
The significance of bold values is the biggest values in the same column.

delivered better overall all the indices for the acne grading problem. We also compared AcneGrader with the other models, including VGGNet,[59] Inception,[60] ResNet,[61] and DenseNet,[62] and our performance outperformed theirs.

3.8 | Evaluation of AcneGrader performance on the skin cancer dataset

To further verify the effectiveness of our algorithm, we use a skin cancer dataset to evaluate the performance of AcneGrader, and we compare four state-of-the-art methods, discretized interpretable multi-layer perceptron (DIMLP)-ensemble,[63] convolutional neural network (CNN),[64] three-way decision-based Bayesian deep learning (TWDBDL),[65] and binary residual feature fusion (BARF) (cross),[45] as shown in Table 4. We can see that BARF(cross) can achieve the best performance. Our method outperforms this method by 3.94% in accuracy, 4.37% in recall, 2.43% in Pr, and 3.42% in F1-score. Overall, our method significantly outperforms the state-of-the-art methods.

Different base models may influence the performance of AcneGrader. Table 5 shows how AcneGrader performs using different neural network base models, including ResNet18 and ResNet34. Different combinations of these base models are also evaluated. We mix 10 ResNet18 and 20 ResNet34 as the denotation "ResNet18+34(10+20)," 15 ResNet18 and 25 ResNet34 as "ResNet18+34(15+15)," and 20 ResNet18 and 10 ResNet34 as "ResNet18+34(20+10)." The combinations of ResNet18 alone and "ResNet18+34(20+10)" achieved the best accuracies of 93.18%. According to the simplicity principle of Occam's Razor rule,[66] the proposed algorithm AcneGrader with ResNet18 can achieve the best F1-score, and we use ResNet18 as the base model.

TABLE 4 Comparing our method AcneGrader with other state-of-the-art methods on the skin cancer dataset

	Accuracy	Recall	Pr	F1-score
Bologna and Fossati[63]	84.90	N/A	N/A	N/A
Lee and Chin[64]	82.90	N/A	N/A	N/A
Abdar et al.[65]	88.95	N/A	N/A	89.00
Abdar et al.[45]	89.24	89.30	89.11	89.18
AcneGrader (ours)	**93.18**	**93.67**	**91.53**	**92.59**

Abbreviation: Pr, precision.
The significance of bold values is the biggest values in the same column.

TABLE 5 Evaluate how the base model influences the performance of AcneGrader

	Accuracy	Recall	Pr	F1-score
ResNet18	**93.18**	93.67	91.53	**92.59**
ResNet34	90.76	92.00	88.18	90.05
ResNet18+34 (15+15)	92.88	93.33	91.21	92.26
ResNet18+34 (10+20)	92.88	**94.33**	90.42	92.33
ResNet18+34 (20+10)	**93.18**	93.33	**91.80**	92.56

Abbreviation: Pr, precision.
The significance of bold values is the biggest values in the same column.

3.9 | Sensitivity analysis of the user-defined parameters

The setting of hyperparameters is an important step for machine learning tasks, and some comprehensive studies recommended widely used hyperparameter selection strategies.[67,68] The recommended sensitivity analysis was carried out for the user-defined parameter values of our AcneGrader framework, as shown in Table 6.

Four hyperparameters were evaluated for AcneGrader, that is, the sampling ratio of the training dataset (r), the number of base models (n), the ensemble strategy (e), and the ensemble pruning method (p). This study used Taguchi's method[69] to set the different choices of the four hyperparameters, which were $r \in \{40\%, 60\%, 80\%\}$; $n \in \{10, 20, 30\}$; $e \in$ {Vote, Avg, Sig, Wavg}; and $p \in$ {ErrorRate, DisM, DF, Kappa}. A full-factorial analysis needs to run the framework AcneGrader $3^2 \times 4^2 = 144$ times. Taguchi's method used an orthogonal array to decrease the number of experimental runs. Table 6 shows that 19 experimental runs were performed. The best Acc = 85.68

TABLE 6 Sensitivity analysis of user-defined parameters

No.	*R*	*n*	*E*	*p*	Pr	Sn	Sp	YI	Acc	*p*-Value
1	40%	10	Vote	ErrorRate	82.59	80.04	93.61	73.65	83.49	0.047
2	60%	20	Vote	DisM	83.15	81.01	94.09	75.11	84.59	0.458
3	80%	30	Vote	DF	84.71	83.14	94.44	77.58	85.55	0.299
4	60%	30	Avg	ErrorRate	82.14	80.44	93.74	74.19	83.63	0.058
5	40%	10	Avg	DisM	80.33	77.43	92.83	70.26	81.51	0.058
6	40%	20	Avg	DF	81.57	78.60	93.14	71.74	82.33	0.047
7	80%	20	Sig	ErrorRate	83.66	82.18	94.14	76.32	84.73	0.145
8	40%	30	Sig	DisM	81.44	78.02	93.22	71.24	82.47	0.058
9	60%	10	Sig	DF	81.90	79.23	93.38	72.61	82.88	0.008
10	80%	10	Wavg	Kappa	83.50	81.54	93.90	75.43	84.18	0.058
11	60%	20	Wavg	DisM	81.36	79.62	93.57	73.18	83.22	0.047
12	40%	30	Wavg	Kappa	81.17	78.55	93.39	71.94	82.88	0.006
13	60%	20	Avg	Kappa	82.07	80.07	93.62	73.69	83.42	0.006
14	**80%**	**10**	**Vote**	**DisM**	**84.92**	**83.31**	**94.49**	**77.80**	**85.68**	–
15	40%	20	Vote	Kappa	83.29	80.39	93.83	74.22	84.11	0.201
16	80%	30	Avg	ErrorRate	83.83	81.87	93.93	75.80	84.38	0.058
17	60%	10	Sig	Kappa	82.00	79.58	93.34	72.93	82.81	0.008
18	60%	10	Wavg	ErrorRate	82.16	79.81	93.44	73.25	83.01	0.087
19	80%	20	Wavg	DF	83.19	81.84	94.03	75.87	84.38	0.047

Abbreviations: Acc, accuracy; DF, double-fault; Pr, precision; Sn, sensitivity; Sp, specificity; YI, Youden index.
The significance of bold values is the biggest values in the same column.

was achieved by the Taguchi's method, which was slightly worse than the accuracy of 85.75 obtained in the above ensemble pruning experiments. The Mann–Whitney *U*-test is a nonparametric test of the null hypothesis that the probability of *X* is greater than *Y*. In this study, we compared the fold accuracies of the fivefold cross-validation by different parameters with the best model (entry 14 in Table 6) using the one-tailed Mann–Whitney *U*-test. The experimental data showed that the ensemble pruning strategy Kappa tended to be significantly smaller than the best model using the ensemble pruning strategy DisM. Most experiments with the 40% or 60% sampling rates achieved significantly smaller prediction accuracies than the best model using the 80% sampling rate. However, the number of base models and the ensemble strategy did not show consistent patterns of significantly smaller prediction performances compared with the best model.

This study recommended the ensemble model with Acc = 85.75 in the ensemble pruning module. However, the experimental data also suggested the efficiency of the Taguchi's method in tuning the hyperparameters of complicated choices.

4 | DISCUSSION

Acne is one of the most common skin lesions in adolescents, and its subtypes show substantially different patterns. For example, some acne lesions are isolated blackheads or whiteheads, while the other subtypes consist of painful pus-filled lumps under the skin. All these complications cause challenges for computational detection and grading of acne.

Ensemble learning may serve as a candidate framework to predict acne grades with multiple base models to describe acne from different aspects. However, its dependence on many base models also induces high computational requirements. Therefore, we proposed the ensemble pruning strategy to remove the redundant base models from the ensemble learning framework.

First, we tuned the parameters of the proposed framework AcneGrader. We found that more training samples did not lead to better prediction performance, and 80% of the training dataset generated the best models. The ensemble strategy is another important factor for an ensemble learning model, and the voting strategy outperformed the other ones.

Then, we tested different ensemble pruning algorithms on the trained ensemble learning model. Overall, the pruning algorithms improved the ensemble learning models. The Kappa statistics achieved the best model performance using 22 base models and was chosen as the default model. The error pruning algorithm achieved competitive performance using only seven base models. If our framework is used in an environment with limited computing resources, such as mobile devices, the ensemble model pruned by the error pruning strategy may be considered.

Next, we use different feature selection algorithms to select the important features and evaluate the feature sets with multiple classifiers. The feature selection algorithm RFE and the classifier SVM achieve the best performance. The best model uses eight base models to achieve better performance compared with the 22 base models used by the Kappa statistics.

Finally, a comparison with the previous model on the same dataset and the other algorithms showed that the proposed ensemble pruned model achieved overall the best acne grading performances.

5 | CONCLUSION

This study proposed a novel ensemble learning and pruning framework, AcneGrader. Our algorithm was applied to the four-class classification problem and achieved an improved prediction accuracy of 0.8582 compared with the state-of-the-art method. This classification model may be further improved by increasing the complexity of the ensemble learning and pruning framework, for example, more base models or better base classifiers.

ACKNOWLEDGMENTS

This work was supported by the Jilin Senior and Junior Technological Innovation Team (20210509055RQ), the National Natural Science Foundation of China (62072212 and U19A2061), the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), Science and Technology Project of Education Department of Jilin Province (JJKH20200328KJ), and the Fundamental Research Funds for the Central Universities, JLU. We truly appreciate the positive and insightful comments from the anonymous reviewers.

DATA AVAILABILITY STATEMENT

The public datasets are available at the following two websites:

1. https://github.com/xpwu95/ldl
2. https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign

ORCID

Fengfeng Zhou https://orcid.org/0000-0002-8108-6007

REFERENCES

1. Ekore RI, Ekore JO. Excoriation (skin-picking) disorder among adolescents and young adults with acne-induced postinflammatory hyperpigmentation and scars. Int J Dermatol. 2021;60(>12):1488–93.
2. Layton AM, Thiboutot D, Tan J. Reviewing the global burden of acne: how could we improve care to reduce the burden? Br J Dermatol. 2021;184(2):219–25.
3. Cordain L, Lindeberg S, Hurtado M, Hill K, Eaton SB, Brand-Miller J. Acne vulgaris: a disease of Western civilization. Arch Dermatol. 2002;138(12):1584–90.
4. Wei B, Pang Y, Zhu H, Qu L, Xiao T, Wei HC, et al. The epidemiology of adolescent acne in North East China. J Eur Acad Dermatol Venereol. 2010;24(8):953–7.
5. Aktan S, Özmen E, Sanli B. Anxiety, depression, and nature of acne vulgaris in adolescents. Int J Dermatol. 2000;39(5):354–7.
6. Dunn LK, O'Neill JL, Feldman SR. Acne in adolescents: quality of life, self-esteem, mood and psychological disorders. Dermatol Online J. 2011;17(1):1.
7. <Gupta MA, Gupta AK. Depression and suicidal ideation in dermatology patients with acne, alopecia areata, atopic dermatitis and psoriasis. Br J Dermatol. 1998;139(5):846–50.
8. Goulden V, Stables GI, Cunliffe WJ. Prevalence of facial acne in adults. J Am Acad Dermatol. 1999;41(4):577–80.
9. Khunger N, Kumar C. A clinico-epidemiological study of adult acne: is it different from adolescent acne? Indian J Dermatol Venereol Leprol. 2012;78(3):335.
10. Hayashi N, Akamatsu H, Kawashima M, Acne Study Group. Establishment of grading criteria for acne severity. J Dermatol. 2008;35(5):255–60.
11. Wu X, Wen N, Liang J, Lai Y-K, She D, Cheng M-M, et al. Joint acne image grading and counting via label distribution learning. Paper presented at the Proceedings of the IEEE International Conference on Computer Vision. 2019.
12. Li X, Zhou Y, Dvornek NC, Zhang M, Zhuang J, Ventola P, et al. Pooling regularized graph neural network for fMRI biomarker analysis. Med Image Anal. 2020;74:102233.
13. Zhang X-M, Liang L, Liu L, Tang M-J. Graph neural networks and their current applications in bioinformatics. Front Genet. 2021;12.
14. Harada S, Akita H, Tsubaki M, Baba Y, Takigawa I, Yamanishi Y, et al. Dual graph convolutional neural network for predicting chemical networks. BMC Bioinf. 2020;21(3):1–13.
15. Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. J Chem Inf Model. 2019;59(9):3981–8.
16. You R, Yao S, Mamitsuka H, Zhu S. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. Bioinformatics. 2021;37(1):i262–71.
17. Li X, Zhou Y, Dvornek N, Zhang M, Gao S, Zhuang J, et al. Braingnn: interpretable brain graph neural network for fMRI analysis. Med Image Anal. 2021;74:102233.
18. Chao C-H, Zhu Z, Guo D, Yan K, Ho T-Y, Cai J, et al. Lymph node gross tumor volume detection in oncology imaging via relationship learning using graph neural network. In: Medical Image Computing and Computer Assisted Intervention. Springer; 2020.
19. Chen X, Pan L. A survey of graph cuts/graph search based medical image segmentation. IEEE Rev Biomed Eng. 2018;11:112–24.
20. Kittigul N, Uyyanonvara B. Automatic acne detection system for medical treatment progress report. Paper presented at the 2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES). 2016.
21. Maroni G, Ermidoro M, Previdi F, Bigini G. Automated detection, extraction and counting of acne lesions for automatic evaluation and tracking of acne severity. Paper presented at the 2017 IEEE Symposium Series on Computational Intelligence (SSCI). 2017.
22. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access. 2017;6:9375–89.
23. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
24. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.
25. Xu Y, Mo T, Feng Q, Zhong P, Lai M, Eric I, et al. Deep learning of feature representation with multiple instance learning for medical image analysis. Paper presented at the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014.

26. Ko J, Cheoi KJ. Image-processing based facial imperfection region detection and segmentation. Multimedia Tools Appl. 2021;80:34283–96.
27. Park K-H, Kim Y-H. Skin condition analysis of facial image using smart device: based on acne, pigmentation, flush and blemish. J Adv Inform Technol Converg. 2018;8(2):47–58.
28. Yadav N, Alfayeed SM, Khamparia A, Pandey B, Thanh DNH, Pande S. HSV model-based segmentation driven facial acne detection using deep learning. Expert Syst. 2021;e12760.
29. Beluch WH, Genewein T, Nürnberger A, Köhler JM. The power of ensembles for active learning in image classification. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
30. Chen W, Hong H, Li S, Shahabi H, Wang Y, Wang X, et al. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. J Hydrol. 2019;575:864–73.
31. Liu S-J, Luo H, Shi Q. Active ensemble deep learning for polarimetric synthetic aperture radar image classification. IEEE Geosci Remote Sens Lett. 2020;18(9):1580–4.
32. Liu W, Zhang M, Luo Z, Cai Y. An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors. IEEE Access. 2017;5:24417–25.
33. Zheng X, Chen W, You Y, Jiang Y, Li M, Zhang T. Ensemble deep learning for automated visual classification using EEG signals. Pattern Recognit. 2020;102:107147.
34. Kuncheva LI, Rodríguez JJ, Plumpton CO, Linden DEJ, Johnston SJ. Random subspace ensembles for fMRI classification. IEEE Trans Med Imaging. 2010;29(2):531–42.
35. Takemura A, Shimizu A, Hamamoto K. Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the AdaBoost algorithm with feature selection. IEEE Trans Med Imaging. 2009;29(3):598–609.
36. Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A. Melanoma classification on dermoscopy images using a neural network ensemble model. IEEE Trans Med Imaging. 2017;36(3):849–58.
37. Cao Y, Geddes TA, Yang JYH, Yang P. Ensemble deep learning in bioinformatics. Nat Mach Intell. 2020;2(9):500–8.
38. Hernández-Lobato D, Martinez-Munoz G, Suárez A. Statistical instance-based pruning in ensembles of independent classifiers. IEEE Trans Pattern Anal Mach Intell. 2008;31(2):364–9.
39. Hernández-Lobato D, Martínez-Muñoz G, Suárez A. Statistical instance-based pruning in ensembles of independent classifiers. IEEE Trans Pattern Anal Mach Intell. 2009;31(2):364–9.
40. Martinez-Munoz G, Hernández-Lobato D, Suárez A. An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Trans Pattern Anal Mach Intell. 2008;31(2):245–59.
41. Hu R, Huang Q, Chang S, Wang H, He J. The MBPEP: a deep ensemble pruning algorithm providing high quality uncertainty prediction. Appl Intell. 2019;49(8):2942–55.
42. Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK. Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. IEEE Access. 2021;8:115041–50.
43. Zhang S, Chen Y, Zhang W, Feng R. A novel ensemble deep learning model with dynamic error correction and multi-objective ensemble pruning for time series forecasting. Inform Sci. 2021;544:427–45.
44. Hayashi N, Akamatsu H, Kawashima M, Group AS. Establishment of grading criteria for acne severity. J Dermatol. 2008;35(5):255–60.
45. Abdar M, Fahami MA, Chakrabarti S, Khosravi A, Pławiak P, Acharya UR, et al. BARF: a new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification. Inform Sci. 2021;577:353–78.
46. Chen W, Pradhan B, Li S, Shahabi H, Rizeei HM, Hou E, et al. Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis. Nat Resour Res. 2019;28(4):1239–58.
47. Chen Y, Wang Y, Gu Y, He X, Ghamisi P, Jia X. Deep learning ensemble for hyperspectral image classification. IEEE J Selected Topics Appl Earth Observ Rem Sens. 2019;12(6):1882–97.
48. Moral-García S, Mantas CJ, Castellano JG, Benítez MD, Abellan J. Bagging of credal decision trees for imprecise classification. Expert Syst Appl. 2020;141:112944.
49. Yariyan P, Janizadeh S, Van Phong T, Nguyen HD, Costache R, Van Le H, et al. Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. Water Resour Manage. 2020;34(9):3037–53.
50. Jauhari F, Supianto AA. Building student's performance decision tree classifier using boosting algorithm. Indones J Electr Eng Comput Sci. 2019;14(3):1298–304.
51. Qian N, Wang X, Fu Y, Zhao Z, Xu J, Chen J. Predicting heat transfer of oscillating heat pipes for machining processes based on extreme gradient boosting algorithm. Appl Therm Eng. 2020;164:114521.
52. Tanha J. A multiclass boosting algorithm to labeled and unlabeled data. Int J Mach Learn Cybernet. 2019;10(12):3647–65.
53. Vezhnevets A, Barinova O. Avoiding boosting overfitting by removing confusing samples. Paper presented at the European Conference on Machine Learning. 2007.
54. Tsakiridis NL, Tziolas NV, Theocharis JB, Zalidis GC. A genetic algorithm-based stacking algorithm for predicting soil organic matter from vis–NIR spectral data. Eur J Soil Sci. 2019;70(3):578–90.
55. Margineantu DD, Dietterich TG. Pruning adaptive boosting. Paper presented at the ICML. 1997.
56. Skalak D. The sources of increased accuracy for two proposed boosting algorithms. In: Proceedings of the American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop. 1999.
57. Giacinto G, Roli F. Design of effective neural network ensembles for image classification purposes. Image Vision Comput. 2001;19(9–10):699–707.
58. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
59. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
60. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
61. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
62. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
63. Bologna G, Fossati S. A two-step rule-extraction technique for a CNN. Electronics. 2020;9(6):990.
64. Lee KW, Chin RKY. The effectiveness of data augmentation for melanoma skin cancer prediction using convolutional neural networks. Paper presented at the 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). 2020.
65. Abdar M, Samami M, Dehghani Mahmoodabad S, Doan T, Mazoure B, Hashemifesharaki R, et al. Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput Biol Med. 2021;135:104418.
66. Orozco-Sevilla V, Coselli JS. Commentary: Occam's razor: the simplest solution is always the best. J Thorac Cardiovasc Surg. 2020;S0022-5223(20):32915–9.

67. Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, et al. A novel method for improved accuracy of transcription factor-binding site prediction. Nucleic Acids Res. 2018;46(12).
68. Wang J, Kumbasar T. Parameter optimization of interval Type-2 fuzzy neural networks based on PSO and BBBC methods. IEEE/CAA J Autom Sin. 2019;6(1):247–57.
69. Jugulum R, Taguchi S. Computer-based robust engineering: essentials for DFSS. Quality Press; 2004.

How to cite this article: Liu S, Fan Y, Duan M, Wang Y, Su G, Ren Y, et al. AcneGrader: An ensemble pruning of the deep learning base models to grade acne. Skin Res Technol. 2022;28:677–688. https://doi.org/10.1111/srt.13166