IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 28, 2021, accepted March 1, 2021, date of publication March 24, 2021, date of current version April 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3064040

Research on Application of Classification Model Based on Stack Generalization in Staging of Cervical Tissue Pathological Images

SHUAILEI ZHANG[1], CHEN CHEN[1], CHENG CHEN[1], FANGFANG CHEN[1], MIN LI[2], BO YANG[1], ZIWEI YAN[1], AND XIAOYI LV[1,2,3]

[1]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
[2]Key Laboratory of Software Engineering Technology, Xinjiang University, Urumqi 830046, China
[3]Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi 830046, China

Corresponding authors: Chen Chen (1343432873@qq.com) and Xiaoyi Lv (xjuwawj01@163.com)

This work was supported in part by the State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia under Grant SKL-HIDCA -2019-5, and in part by the Special Scientific Research Project for young Medical Science under Grant 2019Q003.

ABSTRACT Cervical cancer is a malignant tumor that threatens women's health and life. Cervical pathology examination, as the gold standard for cervical cancer diagnosis, provides an important basis for the surgical plan and postoperative follow-up strategy for cervical cancer. Cervical biopsy diagnosis includes normal, low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL) and squamous cell carcinoma (SCC). At present, cervical pathology examination still relies on the doctor's personal clinical experience and subjective judgment, which is time-consuming and may cause misdiagnosis or missed diagnosis. In addition, the current intelligent classification of cervical pathological images still has disadvantages such as imperfect classification system and low classification accuracy. Therefore, this experiment uses the ResNet50 model of the convolutional neural network as the feature extractor, and selects the K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM) classifiers in Machine Learning to perform cervical tissue pathological images Discrimination, the accuracy of the classification results were 85.83%, 80.33%, and 86.67%. In order to further improve the accuracy of the model and enhance the applicability and stability of the model, this experiment proposes the Stacked Generalization (SK) classification model. The first-layer base learner of the SK classification model selects CNN-KNN, CNN-RF, CNN-SVM, and the second-layer classifier selects Multilayer Perceptron (MLP). Among them, MLP makes the final result by learning the classification performance of the base learner for label discrimination, the accuracy of the classification model after ensemble learning is 90.00%. In addition, this experiment uses the Synthetic Minority Oversampling Technique (SMOTE) algorithm to amplify the training samples, and the amplified data set has a classification accuracy of 89.17% under the training of the SK classification model. The results show that the SK classification model in this experiment has a high classification ability for cervical histopathological images, and has good generalization ability and robustness.

INDEX TERMS Cervical pathology image, ResNet50, machine learning, multilayer perceptron, stacked generalization, SMOTE.

I. INTRODUCTION

Cervical cancer is one of the fastest growing health problems in the world and the main cause of death among women in developing countries [1]. In developed countries, rapid advances in screening technology and prevention methods have effectively controlled the incidence of cervical cancer [2].In developing countries, the incidence of cervical cancer is relatively high because of underdeveloped medical services and unbalanced medical resources. For example, the uneven distribution of medical resources in China is particularly obvious in the development of the eastern and western regions and has had a tendency to intensify. More regrettable, Xinjiang has a vast terrain, and differences in the medical environment between the north and the south are even worse [3]. Cervical cancer, thyroid cancer and hydatid

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

disease are the three high-incidence diseases in Xinjiang. Therefore, when cervical lesions can be detected early and treated as an intervention, the economic loss of and health threats to the cervical cancer patient can be reduced.

There are three steps [4] in the diagnosis of cervical cancer. The first step involves TCT liquid-based cytology and HPV testing. When abnormal TCT and HPV results are detected [5], electronic colposcopy plays a key role as a ''bridge''. Colposcopy can be used to identify suspicious lesions and take spot biopsy samples. The obtained pathological tissue is sliced and imaged under the microscope. The pathologist makes a diagnosis based on cervical pathological images. The diagnosis result is the ''gold standard'' for the patient's entire treatment process. In the classification of cervical pathological images, as the severity of the disease worsens, the number of atypical immature cells in the epithelium increases from the bottom to the top of the sample, the nucleoli become larger, and the cytoplasm becomes thick and deepens [6]–[8].

Cervical histopathological images play a decisive role in the diagnosis of cervical cancer. However, the current classification and discrimination of cervical tissue pathological images still rely on the pathologist's personal clinical experience and subjective judgment, and the entire discrimination process has problems such as excessive time consumption and large manual investment. At the same time, fatigue caused by pathologists' work experience and working hours is likely to cause misdiagnosis or missed diagnosis. Therefore, how to effectively use computer-aided diagnosis technology to shorten the pathological diagnosis time and improve the classification accuracy of cervical tissue pathological images has important research value. One study [4] proposed a computer-aided diagnosis with a cervical pathology imaging system, which quantifies and classifies the nuclear structure of ultralarge-scale Cervical Intraepithelial Neoplasia(CIN) images with an accuracy rate of 94.25%. The experiment focused on exploring the pathological images of preSCCous cervical lesions. The degree of discrimination has certain practical value for the training and practice of pathologists, but it lacks a more complete classification system for cervical pathology images. A study [9] proposed extracting the texture feature information of pathological cervix images through the grey level co-occurrence matrix (GLCM). The image is segmented using K-means clustering and a label-controlled watershed algorithm, and a support vector machine (SVM) is used for classification. This method has a classification accuracy of 90% for Normal and squamous cell carcinoma (SCC), but the classification accuracy of CIN images is only 70%. Therefore, the classification model of this study is not suitable for sample sets with similar feature vectors, and the classification effect seldom meet the satisfactory clinical requirements. Another study [10] developed a machine vision system using the KS400 macro editing language for 230 images at all levels. The classification scores of the Normal and CIN3 samples in the scoring system reached 98.7%, and the classification scores of the microcytosis and CIN1 samples reached 76.5%. This experimental study lacks a connection between and an evaluation of various Normal- and CIN-level images and microcytosis tissue characteristics. The study [11] proposes an automated and localized method based on fusion to evaluate abnormalities in cervical cancer tissues. After using support vector machines and linear discriminant analysis methods to vote on the vertical stage of 61 cases, CIN has the highest classification accuracy rate of 88.5%.

The overall classification system of most studies is not perfect, and the discrimination effect for highly similar LSILs and HSILs is poor. Aiming at the imaging characteristics of the abovementioned pathological cervical tissues, a single classification model has low applicability. Therefore, the experiment in this article constructed the Stacked Generalization (SK) classification model, In the experiment, feature extraction is performed through the ResNet50 model of convolutional neural network (CNN), The base learner chooses K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), The second layer classifier selects Multilayer Perceptron (MLP).

II. MATERIALS

Cervical Cancer screening for women of childbearing age is mainly divided into three steps. The first step is cervical cytology, called thin-layer liquid-based cytology (TCT), which is currently the main method of screening for early cervical cancer. By observing an exfoliated vaginal cell smear under the microscope, an abnormal cervix is found very early. If the patient's TCT result is diagnosed as suspicious high-grade cancer, the second step of electronic colposcopy is required. Colposcopy involves the use of a microscope to magnify the epidermal tissue of the cervix or genitals with a light source and filter to observe the subtle changes in the cervical epithelium and blood vessels, locate suspicious lesions and determine their severity. If abnormalities are found during colposcopy, then multiple biopsies of suspicious lesions should be performed under the guidance of acetic acid or iodine solution staining and sent for histopathological examination. Histopathological examination as the third step of cervical cancer screening is the ''gold standard'' for the diagnosis of cervical disease in patients.

In this experiment, in the preparation of histopathological cervical slices, the first step is to place the removed tissue block into a pre-prepared 10% formalin solution to denature and coagulate the protein to prevent autolysis or bacterial degradation after cell death and to maintain the original morphological structure of the cell. The second step is to use low-concentration to high-concentration alcohol as a dehydrator to gradually remove the water in the tissue block. The third step is to immerse the transparent tissue block in melted paraffin to cool and solidify into the block. The fourth step is to fix the embedded wax block on a microtome and cut into thin slices. Finally, the tissue slices attached to the glass slide are stained with HE and labeled.

The tissue slice imaging instrument in this experiment was provided by Ningbo Kangfeng Bioinformatics Co., Ltd.

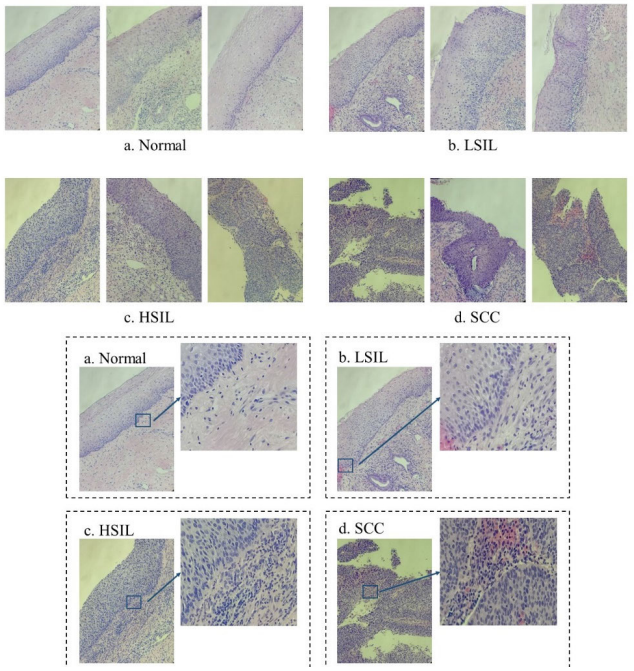

FIGURE 1. Cervical pathology image.

(KFBIO), including processing equipment, digital pathology scanning system, and pathology information system. Figure 1 shows four types of cervical pathology images.

In Figure 1, a, b, c, d are Normal, low-grade squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL) and squamous cell carcinoma (SCC). Respectively, from these four cervical pathological images, it can be seen that as the degree of a cervical lesion increases, the epithelial cell arrangement becomes increasingly disordered until the polarity disappears, the cells are significantly heterogeneous, the nucleus is enlarged, the chromatin distribution is uneven, and there is a mitotic phase. When SCC cells penetrate the epithelial basement membrane, dendritic, cord-like, diffuse or mass SCC nests may appear in the interstitium. The magnification of the images collected in this experiment is x200. Due to the presence of cell superposition, abnormal cell nuclei, and various cancer nest shapes in the image, the effect of using conventional cell segmentation techniques is not good. This article will use Resne50 model to extract image features.

III. METHODS

A. EXPERIMENT PROCEDURE

The experimental data was provided by the First Affiliated Hospital of Xinjiang Medical University. A total of 480 cervical tissue biopsy imaging data of 84 patients who were in the hospital from September 2018 to October 2019 were collected. Among them, 144 were normal, 108 were low-grade squamous intraepithelial lesions (LSIL), 100 were high-grade squamous intraepithelial lesions (HSIL), and 128 were squamous cell carcinoma (SCC). In order to prevent over-fitting, this experiment divides the data according to all the images of

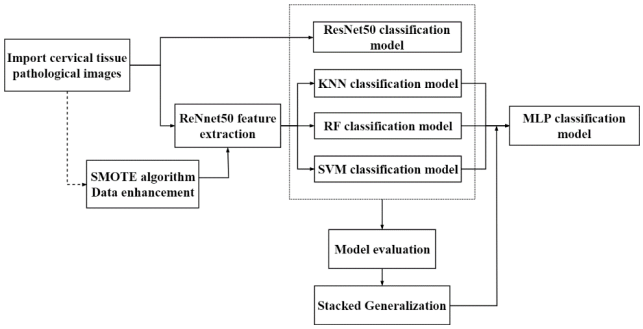

FIGURE 2. General flow chart.

each patient as the smallest set. 75% of the training validation set and 25% of the test set are divided randomly according to each category. Therefore, in normal cervical pathology images, there are 108 training validation sets and 36 testing sets; in LSIL cervical pathology images, there are 81 training validation sets and 27 testing sets; in HSIL cervical pathology images, training There are 75 validation sets and 25 test sets; in the SCC cervical pathology images, there are 96 training validation sets and 32 test sets. Therefore, there are a total of 360 sheets in the training verification set and 120 sheets in the test set.

Figure 2 is the general flow chart of this research experiment. The ResNet50 classification model in the convolutional neural network trains and discriminates the input cervical tissue pathological images. In addition, the experiment retained the 2048-dimensional feature vector extracted from the ResNet50 model, on this basis, three classifiers of KNN, RF, and SVM were selected for training and discrimination. In the evaluation of the experimental results of the CNN classification model, CNN-KNN classification model, CNN-RF classification model, and CNN-SVM classification model, this experiment selects CNN-KNN, CNN-RF, CNN-SVM classification model for ensemble learning, the purpose of ensemble learning is to further improve the classification performance and generalization ability of the model.The ensemble learning algorithm selects the stack generalization classification model. Among them, the second-level classifier of the stack generalization classification model makes the final label discrimination by learning the classification performance of the first-level base learner, the second layer classifier of the stack generalization model in this experiment selects MLP. In addition, this experiment did a comparative experiment before and after data enhancement, and the enhancement algorithm selected SMOTE.

B. CNN CLASSIFICATION MODEL

In the digital image classification problem, the complexity of the CNN model directly affects the recognition effect. The more layers that the convolutional neural network has, the richer the features that are extracted from different labels [12]. However, the continuous deepening of the network will cause the problems of gradient explosion and gradient disappearance, and the optimization effect will

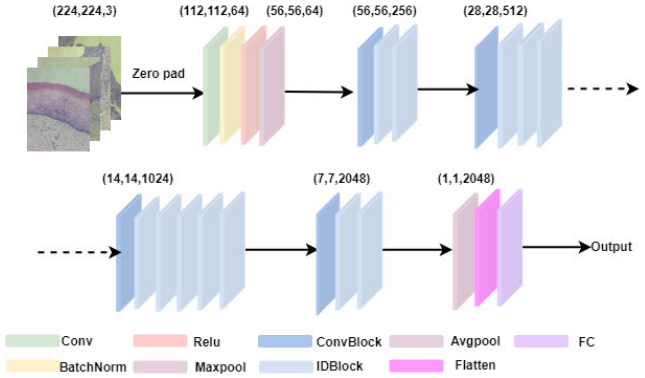

FIGURE 3. CNN classification model.

be worse. To solve this problem, K He et al proposed a network structure based on VLAD(Vector of Local Aggregated Descriptors) and Highway Network—ResNet [13]. This paper uses the ResNet50 model of convolutional neural network. In order to reduce the time consumption of the model, the high-dimensional pathological tissue slice image needs to be cropped, and the cropped size is (224, 224, 3). In addition, in order to improve the performance of the model, this experiment enhances the data by rotating, horizontal and vertical flipping, brightness and contrast. Figure 3 is a diagram of the ResNet50 network structure. In order to prevent the lack of pixel information, the input cervical tissue pathological image (224, 224, 3) needs to be zero-filled (3*3), and the input image after zero-filling passes through the convolutional layer (kernel_size=7*7, filters = 64, stride=2*2, padding=3) extract the feature vector (112, 112, 64), the channel axis is standardized by BatchNorm and Relu nonlinear activation function, and then through a MaxPooling layer (windows=3*3, Stride=2*2) Perform maximum pooling and obtain the converged feature map. The network layer structure of steps 2 to 4 are four large blocks respectively. Among them, ConvBlock is used to change the dimension of the network, and IDBlock is used to deepen the network. As the number of network layers increases, the local and global feature vectors (7, 7, 2048) of the cervical pathology image are extracted and sent to the Average Pooling layer (windows=7*7), and the 2048-dimensional feature vector is obtained after Flatten flattening. On this basis, the 2048-dimensional feature vector connect the fully connected layer, and finally output the image prediction label. Table 1 is the experimental environment of this Resnet50. Since the number of images is too small, we will use transfer learning for ResNet50 model. We will use frozen the previous layers until flatten layers.

C. CNN-KNN CLASSIFICATION MODEL

It is worth noting that traditional digital image classification algorithms require manual selection of feature extraction algorithms. This step not only requires investigating the color, texture, and shape feature details of the image itself, but also constantly trying various feature extraction algorithms and adjusting parameters. The whole process takes a long time. The neural network algorithm avoids the complexity of feature extraction and has an excellent performance in the field of digital images [14]. Therefore, the feature vectors of KNN, RF, and SVM classifiers in this experiment use the 2048-dimensional feature vectors extracted by CNN in Section 3.1.

TABLE 1. Lab environment.

Batch_size	Learning_rate	Momentum	Optimizer	Epochs
16	0.001	0.9	Adam	30

The K-nearest neighbour(KNN) algorithm, as a well-known statistical method for pattern recognition, occupies an important position in machine learning classification algorithms [15], [16]. The basic idea of the KNN algorithm is that if most of the k most similar samples in the feature space of a sample belong to a certain category, the sample also belongs to this category. In the KNN algorithm, the distance between objects is calculated as an index of dissimilarity between objects. This experiment uses Euclidean distance. As shown in formula (1), Euclidean distance calculates the difference between x individual and y individual, and both x individual and y individual contain n-dimensional features. The greater the distance, the greater the difference between individuals.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

KNN only determines the category of the sample to be classified based on the nearest one or a few samples in the decision-making process. Therefore, a too large k value may easily cause under-fitting, and too small a value may easily cause over-fitting. In this experiment, a 10-fold cross-validation was performed on the divided training validation set to determine the best k value. In the 10-fold cross-validation process, the training validation set (360 cases) needs to be evenly divided into 10 groups, and each subset of data (36 cases) is used as a validation set, and the remaining 9 groups (324 cases) are used as training Set, this will get 10 models, and use the average of the classification accuracy of the final validation set of these 10 models as the performance index of the classifier under this 10-fold cross-validation. The optimization result is shown in Figure 4. In Figure 4, the abscissa represents the value of k, and the ordinate represents the average value of accuracy after 10 cross-validation corresponding to each determination of a k value. The optimal range of k value in this experiment is 1-30. Experimental results show that when k = 10, the average accuracy of 10-fold cross-validation tends to converge, which is 86.11%. Therefore, in this experiment, the k value is set to 10.

D. CNN-RF CLASSIFICATION MODEL

In 2001, Breiman made a modification on the basis of bagging and proposed a classifier that uses multiple trees to train and predict samples – the random forest algorithm. The algorithm is based on a random method to build decision trees, and

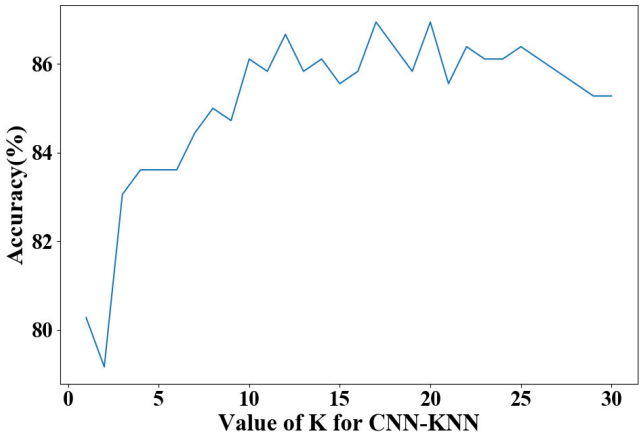

FIGURE 4. k-value parameter optimization.

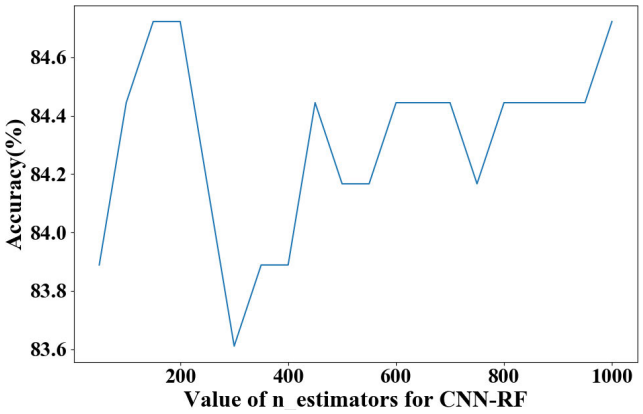

FIGURE 5. n_estimators-value parameter optimization.

then, these decision trees form a forest. There is no correlation between decision trees. When there is a new sample input, each tree processes it independently. A judgement is made and the classification result of the sample is determined according to the principle of majority [17]–[19]. To construct a random forest classification model, this experiment needs to determine the number of trees in the forest n_estimators. Although the larger the n_estimator, the higher the discrimination accuracy of the classifier, but the memory occupied and the training and prediction time will increase accordingly, and the marginal benefit is decreasing. Therefore, this experiment needs to select the largest n_estimator from the acceptable memory/time. Similarly, this experiment optimizes the best n_estimators value through 10-fold cross-validation on the training and validation set. The optimization result is shown in Figure 5. In Figure 5, the abscissa represents the value of n_estimators, and the ordinate represents the average value of the accuracy after 10-fold cross-validation. The optimization range of n_estimators value in this experiment is 50-1000. It can be found that when n_estimators=150, the 10-fold cross-validation average accuracy rate is the highest, which is 84.75%. Therefore, the value of n_estimators in this experiment is set to 150.

E. CNN-SVM CLASSIFICATION MODEL

Support vector machines have the characteristics of a low generalization error rate and fast classification speed, and they are easy to explain and suitable for small -sample classification problems [20], [21]. The main goal of an SVM is to classify samples mapped to the hyperplane, as shown in equation (2).

$$min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\iota_{0/1}\left(y\left(w^T x_i + b\right) - 1\right).\iota_{0/1} = \begin{cases}1, & \left(y\left(w^T x_i + b\right) - 1\right) < 0\\ 0, & \left(y\left(w^T x_i + b\right) - 1\right) \geq 0\end{cases} \quad (2)$$

where $\iota_{0/1}$ is the loss function, which has the characteristics of non-convexity and discontinuity. In future solutions, functions with good mathematical properties are often used to replace it in a process called a substitution loss. In addition, C is the penalty coefficient. Facing the constrained quadratic linear programming problem in equation (2), this paper uses the Lagrangian multiplier method to solve the problem. The equivalent conversion of the dual problem is given in equation (3).

$$max_{\alpha} = \sum_{i}^{m}\alpha_i - \frac{1}{2}\sum_{i}^{m}\sum_{j}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j.$$
$$s.t.\sum_{i}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \quad (3)$$

Equation (3) shows that selecting an appropriate kernel function is very important because it not only can reduce the calculation of the inner product of $x_i^T x_j$ but can also improve the accuracy of classification. In addition, the distribution of the penalty coefficient C and the γ value also affects the number of support vectors and the spatial distribution of the sample map. In this experiment, the optimal C and γ are determined through grid optimization and 10-fold cross-validation. The optimization results are shown in Figure 6. In Figure 6, the x-axis represents the C value, the optimization range is 0.1-1, the y-axis represents the γ value, and the optimization range is 0.001-0.01, and the z-axis represents the average accuracy after 10-fold cross-validation. It can be found that when $C = 0.9$ and $\gamma = 0.001$, the optimization result is the best, and the average accuracy rate is 88.09%.

F. STACKED GENERALIZATION

Ensemble Learning is a type of machine learning framework that completes learning tasks by constructing and combining multiple learners. Its main frameworks include bagging, boosting and stacking [22]–[24].Among them, in order to achieve better performance than a single classifier, the Stacked Generalization algorithm uses the category probability output of several single classifiers as the feature input of the training model of the next layer. The classification model of the next layer makes the final classification judgment by learning new features. It is worth noting

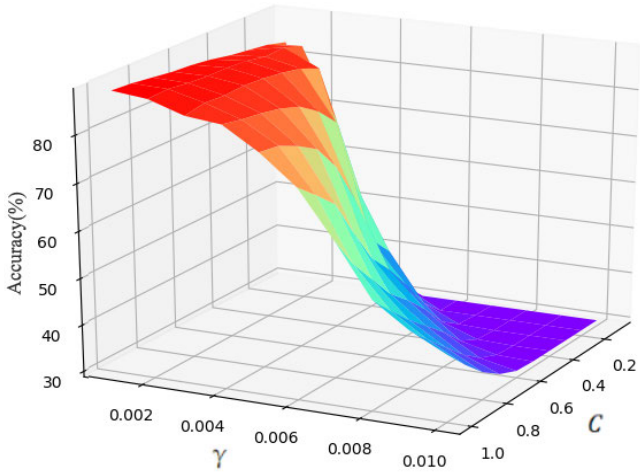

FIGURE 6. *C*, γ-value parameter optimization.

that the choice of the base learner of the SK classification model needs to meet the differences. In addition, classification models with poor classification effects need to be eliminated, because this type of classification model will have a negative effect on improving the classification ability of the fusion model. In this experiment, the SK classification model is selected. The first-layer base learner selects CNN-KNN, CNN-RF, CNN-SVM, and the second-layer learner selects MLP [25], [26]. As shown in Figure 7, the data labels of this experiment are Normal, LSIL, HSIL, SCC. 360 cases in the training validation set and 120 cases in the test set. In the construction of the stack generalization classification model, the 360 training validation sets are divided into 10 groups in turn. When each subset of data (36 cases) is used as a validation set, the remaining 324 data will be used as the training set, so a total of 10 training verification models were obtained. In the first cross-validation model, 324 training sets are trained, and the four category judgment probabilities of 36 verification sets are output, and the dimensions are (36, 12). Therefore, when these 10 models are trained and verified sequentially, a (360, 4) dimensional vector will be obtained. Different from the 10-fold cross-validation of 3.2, 3.3, and 3.4 to determine the best parameters, the 10-fold cross-validation of the stack generalization classification model aims to preserve the discriminative characteristics of the data by the base classifier. This discriminative feature represents the probability of the base classifier discriminating the data set as a certain category. Since there are 3 base classifiers in this experiment, the feature vector of the new training validation set in the stack generalization model is (360, 12) dimensions in total, which is the feature vector of the new training set. Similarly, in the process of each cross-validation, the test set will be discriminated 10 times, so that 10 sets of (120,4) test set feature vectors will be obtained. By averaging the 10 sets of feature vectors, the new test set feature vector (120,12) will be obtained. In this experiment, the second-level classifier of the stack generalization model selects MLP. The new training verification set feature vector is 360 examples,

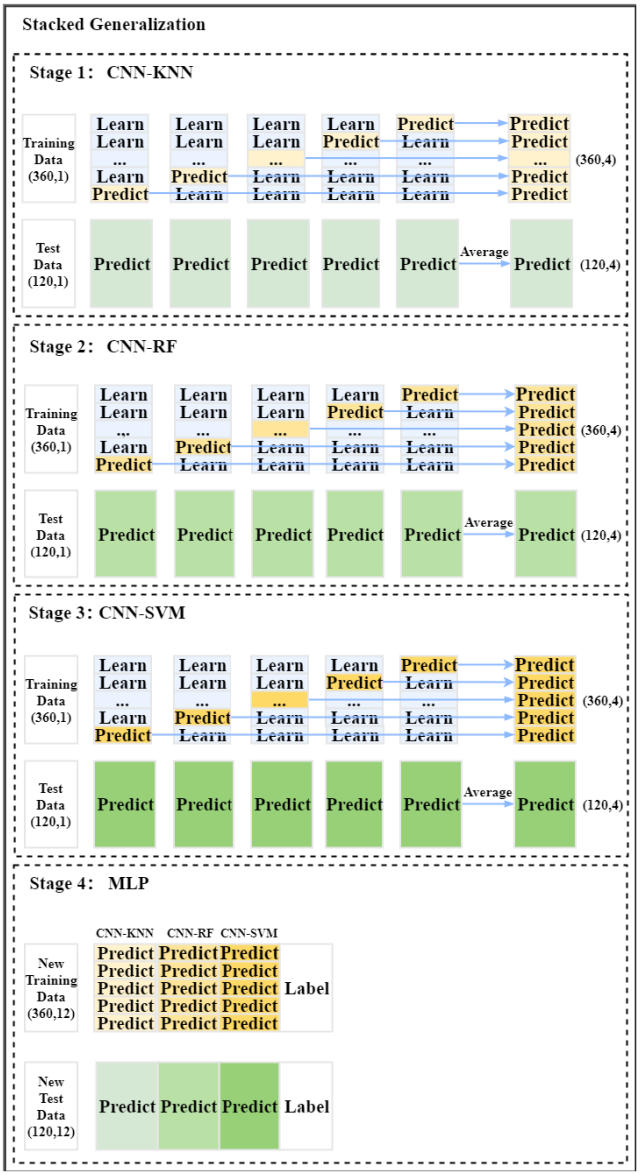

FIGURE 7. Stacked Generalization Classification model.

the dimension is (360, 12), and the new test set feature vector is 120 examples, and the dimension is (120, 12). Finally, this experiment obtains the final discrimination result of the stack generalization classification model by optimizing the parameters of the MLP classifier.

The layers of MLP are fully connected. The bottom layer is the input layer, which is the new feature vector (360, 12), the middle is the hidden layer, and the final is the output layer. Assuming that the input layer is represented by a vector x, the output of the hidden layer is $f(w_1\mathrm{x}+b_1)$,Among them, w_1 is the weight (also called the connection coefficient), b_1 is the bias, the function f in this experiment is the sigmoid function, as shown in formula 3. Its purpose is to directly use numerical optimization methods to learn network parameters. In addition, the hidden layer to the output layer can be regarded as softmax regression, so the output of the output layer is $softmax(w_1x_1+b_1)$. x_1 represents the output $f(w_1\mathrm{x}+b_1)$ of the

hidden layer. The determination of the connection weight and bias between the layers of MLP is the core of the optimization problem. This experiment uses gradient descent (SGD) to randomly initialize all parameters and iteratively train.

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{4}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EVALUATION PARAMETERS

Each column of the confusion matrix represents the predicted category, and the total number of each column represents the number of data predicted to be that category; each row represents the true attribution category of the data, and the total number of data in each row represents the number of data instances of that category. Taking the two classification as an example, this experiment evaluates the accuracy, precision, specificity, and sensitivity of the parameters. The F1-score calculation formula is (5)-(8). It is worth noting that the output result of this experiment is the prediction judgment of the four categories of Normal, LSIL, HSIL, and SCC. Taking Normal as an example, when calculating the evaluation parameters, it is necessary to classify LSIL, HSIL, and SCC into one category and transform it into the label is Normal and the label is not Normal in the binary classification calculation problem. Similarly, the evaluation parameters labeled LSIL, HSIL, and SCC are consistent with the calculation method labeled Normal.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{7}$$

$$\text{F1} = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \tag{8}$$

B. CLASSIFICATION RESULTS OF CNN MODEL

Figure 8 and Table 2 are the confusion matrix of the CNN classification model for each single classification result of this cervical tissue pathology image ('0' means Normal, '1' means LSIL, '2' means HSIL, '3' means Cancer) and Accuracy, specificity, sensitivity, F1-score calculation results. The results show that the CNN classification effect is not good. The reason is that the Resnet50 classification model is suitable for larger training samples. The training sample data in this experiment is less, and the training parameters are not very detailed, so the CNN classification model cannot be obtained on the new instance samples to better classification ability. It is worth noting that the accuracy of the CNN classification model in this experiment is 70.83%, and the accuracy of the CNN-KNN, CNN-RF, CNN-SVM classification model is 85.83%, 80.83% and 86.67%. Since the accuracy of the CNN classification model is relatively low, in the later ensemble learning, the CNN classification model is not used as one of the base learners. The reason is that the selection of the base learner by the SK classification

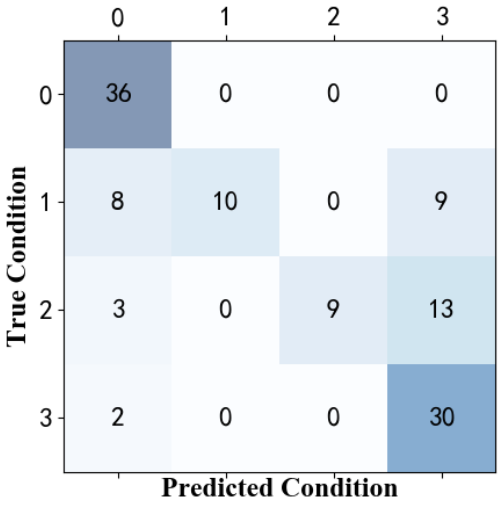

FIGURE 8. Confusion matrix of CNN classification results.

TABLE 2. Experimental results of CNN classification model.

		Precision	Sensitivity	Specificity	F1
CNN	Normal	0.73	1.00	0.85	0.85
	LSIL	1.00	0.37	1.00	0.54
	HSIL	1.00	0.36	1.00	0.53
	SCC	0.58	0.94	0.73	0.71

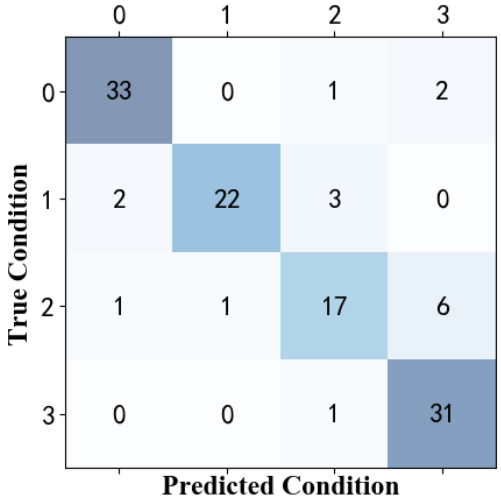

FIGURE 9. Confusion matrix of CNN-KNN classification results.

model not only requires greater differences but also A close and high accuracy rate is required. The 70.83% accuracy rate of the CNN classification model in this experiment will have a negative effect on the final decision-making judgment in the later integrated learning, so it will not be considered.

C. CLASSIFICATION RESULTS OF CNN-KNN MODEL

Figure 9 is the confusion matrix of the CNN-KNN classification model for each single classification result of this cervical tissue pathology image ('0' means Normal, '1' means LSIL, '2' means HSIL, and '3' means Cancer), Table 3 is

TABLE 3. Experimental results of CNN-KNN classification model.

		Precision	Sensitivity	Specificity	F1
CNN-KNN	Normal	0.92	0.92	0.95	0.92
	LSIL	0.95	0.81	0.99	0.88
	HSIL	0.77	0.68	0.95	0.72
	SCC	0.79	0.97	0.91	0.87

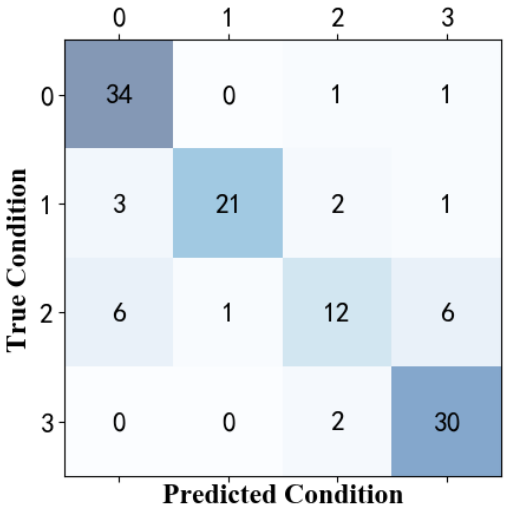

FIGURE 10. Confusion matrix of CNN-RF classification results.

the calculation result of accuracy, specificity, sensitivity and F1-score. It can be found that the classification effect of the CNN-KNN classification model in HSIL is poor. The reason is that the amount of HSIL and SCC data is small, and the sample size of SCC is greater than that of HSIL. After the value of k is determined to be 10, unbalanced samples can easily lead to misjudgement of HSIL as SCC.

D. CLASSIFICATION RESULTS OF CNN-RF MODEL

Figure 10 is the confusion matrix of the CNN-KNN classification model for each single classification result of this cervical tissue pathology image ('0' means Normal, '1' means LSIL, '2' means HSIL, and '3' means Cancer), Table 4 is the calculation result of accuracy, specificity, sensitivity and F1-score. Similarly, in the case of unbalanced samples, the discriminant results of CNN-RF will tend to have more classification labels in the sample set. Therefore, in this classification experiment, when the CNN-RF classification model is applied to cervical histopathological images with greater feature similarity, the classification results of LSIL and HSIL will be more inclined to the Normal and Cancer categories with more samples in the training validation set.

E. CLASSIFICATION RESULTS OF CNN-SVM MODEL

Figure 11 is the confusion matrix of the CNN-KNN classification model for each single classification result of this cervical tissue pathology image ('0' means Normal, '1' means LSIL, '2' means HSIL, and '3' means Cancer), Table 5 is the calculation result of accuracy, specificity, sensitivity

TABLE 4. Experimental results of CNN-RF classification model.

		Precision	Sensitivity	Specificity	F1
CNN-RF	Normal	0.85	0.94	0.90	0.89
	LSIL	0.95	0.78	0.99	0.86
	HSIL	0.75	0.60	0.95	0.67
	SCC	0.79	0.94	0.91	0.86

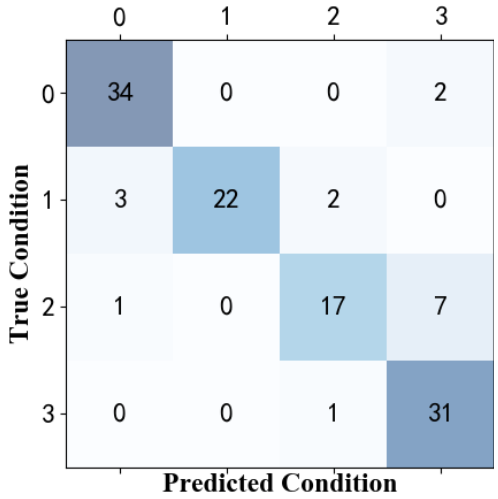

FIGURE 11. Confusion matrix of CNN-SVM classification results.

TABLE 5. Experimental results of CNN-SVM classification model.

		Precision	Sensitivity	Specificity	F1
CNN-SVM	Normal	0.89	0.94	0.95	0.92
	LSIL	1.00	0.81	1.00	0.90
	HSIL	0.85	0.68	0.97	0.76
	SCC	0.78	0.97	0.90	0.86

and F1-score. Similar to CNN-KNN and CNN-RF, the classification ability of CNN-SVM makes up for the lack of classification ability of small sample data sets in the deep learning field to a certain extent, but it is difficult to distinguish LSIL and HSIL with small training sample size, its classification and discrimination ability has not achieved satisfactory improvement effects.

F. CLASSIFICATION RESULTS OF STACK GENERALIZATION MODEL

Figure 12 is the confusion matrix of the CNN-KNN classification model for each single classification result of this cervical tissue pathology image ('0' means Normal, '1' means LSIL, '2' means HSIL, and '3' means Cancer), Table 6 is the calculation result of accuracy, specificity, sensitivity and F1-score. It can be found that the MLP classification model has learned the discriminative ability and the difference of the three base classifiers of CNN-KNN, CNN-RF, and CNN-SVM, and finally has a significant improvement to classify imbalanced data sets.

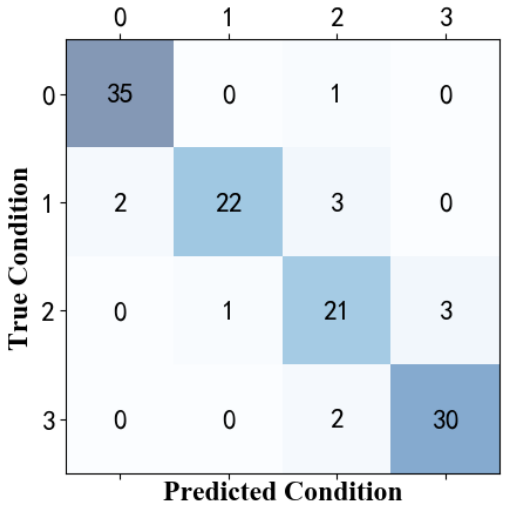

FIGURE 12. Confusion matrix of SK classification results.

TABLE 6. Experimental results of SK classification model.

		Precision	Sensitivity	Specificity	F1
SK	Normal	0.95	0.97	0.98	0.96
	LSIL	0.96	0.81	0.99	0.88
	HSIL	0.78	0.84	0.94	0.81
	SCC	0.91	0.94	0.97	0.92

G. ACCURACY AND TIME CONSUMPTION

The accuracy and time consumption of the base learner and integrated learning built in this experiment are shown in Table 5. And the training time includes the process of optimizing the parameters of each classifier. It can be found that the SK classification n model has a good advantage in the accuracy of the identification of cervical pathological tissues. Compared with the base learner, the time consumption of the model is longer, but considering the importance of clinical diagnosis and treatment plan, the accuracy of computer-aided diagnosis is an indicator that we need to consider more.

V. MODEL EVALUATION

A. ROC CURVE AND AUC VALUE

This experimental model evaluation selects the ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) value. The ROC curve is called the receiver operating characteristic curve. Each point on the ROC curve reflects the sensitivity to the same signal stimulus. The horizontal axis of the ROC curve is the specificity of the false positive rate (FPR), formula 9, and the vertical axis is the sensitivity of the true positive rate (TPR), formula 10. AUC is defined as the area under the ROC curve, and its value is between 0-1, the closer to 1 the better. Figures 13-17 are the ROC curves and AUC values of CNN, CNN-KNN, CNN-RF, CNN-SVM, and stack generalization classification models, where '0' means Normal, '1' means LSIL, and '2' means

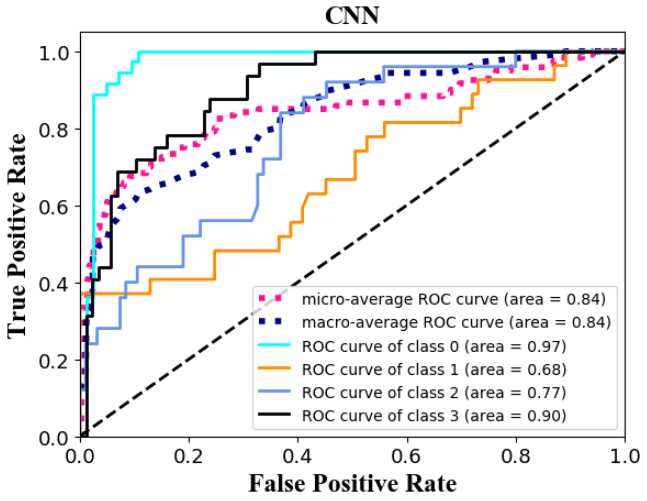

FIGURE 13. ROC curve and AUC value of CNN classification model.

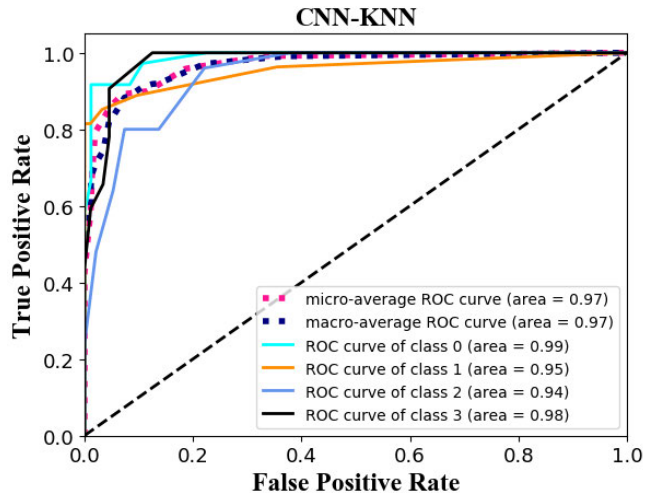

FIGURE 14. ROC curve and AUC value of CNN-KNN classification model.

HSIL, '3' means Cancer. In addition, in order to evaluate the performance of the algorithm on the entire data set, this experiment uses two evaluation models, macro-average and micro-average. The macro average is the arithmetic average of the performance indicators of each class, and the micro average It is the arithmetic average of the performance indicators of each instance (document).

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

It can be seen from Figure 13-17 that the classification effect of the CNN classification model is poor, especially in the classification capabilities of LSIL and HSIL; in addition, compared to the classification of the base learner CNN-KNN, CNN-RF, and CNN-SVM As a result, the stack generalization classification model has further improved the ability to discriminate the same batch of sample instances.

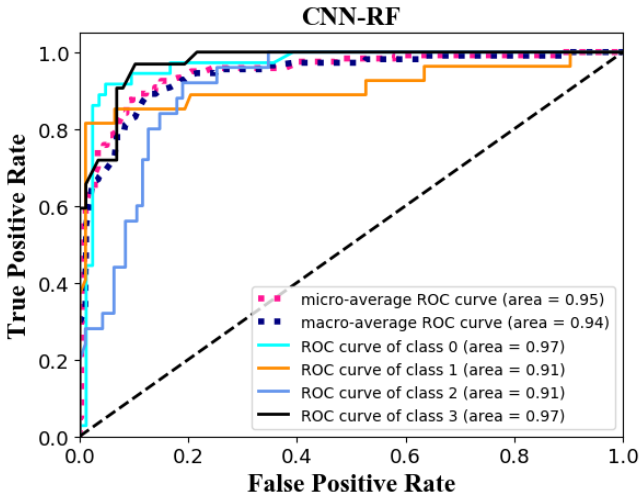

FIGURE 15. ROC curve and AUC value of CNN-RF classification model.

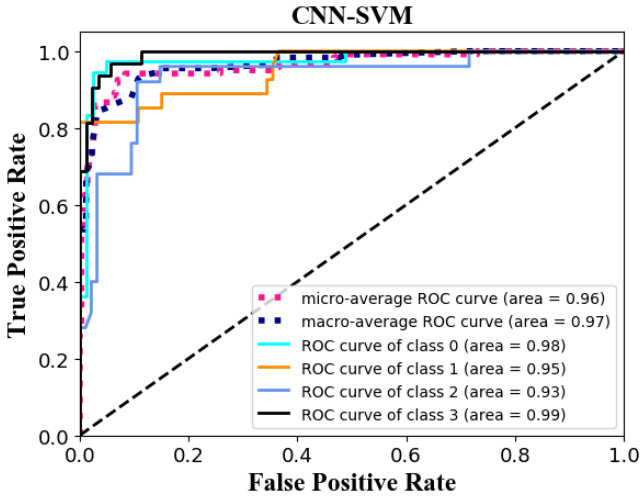

FIGURE 16. ROC curve and AUC value of CNN-SVM classification model.

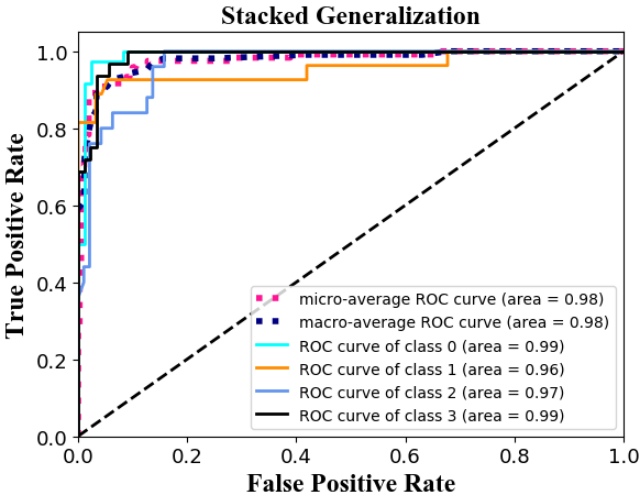

FIGURE 17. ROC curve and AUC value of SK classification model.

VI. DATA ENHANCEMENT

At the end of the experiment, we used the SMOTE algorithm to enhance the 360 training and verification sample sets. Among them, the synthesis strategy of the SMOTE algorithm is to randomly select a sample b from its nearest neighbours for class sample a, and randomly select a point on the line between a and b as the newly synthesized sample [27]–[29]. The enhanced data set is 3000 cases. The data volume of normal, LSIL, HSIL, and SCC are all 750. This experiment uses the previously built SK classification model to train and test the data enhanced by the SMOTE algorithm. The experimental results show that the accuracy of the three base learners of SMOTE-CNN-KNN, SMOTE-CNN-RF, and SMOTE-CNN-SVM are 86.67%, 79.16%, and 85.83% respectively, and the classification accuracy rate based on SMOTE-SK is 89.17%. The single classification accuracy, specificity, sensitivity and F1 value of each classifier are shown in Table 5-8.

From Table 8 to Table 11, it can be seen that compared with the three classification models of SMOTE-CNN-KNN,

TABLE 7. Accuracy and Time consumption.

	Accuracy	training time	testing time
CNN	70.83%	47min	60.06
CNN-KNN	85.83%	42s	0.24s
CNN-RF	83.33%	345s	2.03s
CNN-SVM	86.67%	373s	1.23s
SK	90.00%	791s	0.06s

TABLE 8. Experimental results of SMOTE-CNN-KNN classification model.

		Precision	Sensitivity	Specificity	F1
SMOTE-CNN-KNN	Normal	0.94	0.94	0.98	0.94
	LSIL	0.89	0.89	0.97	0.89
	HSIL	0.75	0.72	0.94	0.73
	SCC	0.85	0.88	0.94	0.86

TABLE 9. Experimental results of SMOTE-CNN-RF classification model.

		Precision	Sensitivity	Specificity	F1
SMOTE-CNN-RF	Normal	0.85	0.94	0.91	0.89
	LSIL	0.95	0.78	0.99	0.86
	HSIL	0.75	0.48	0.96	0.59
	SCC	0.71	0.94	0.93	0.81

TABLE 10. Experimental results of SMOTE-CNN-SVM classification model.

		Precision	Sensitivity	Specificity	F1
SMOTE-CNN-SVM	Normal	0.92	0.94	0.96	0.93
	LSIL	0.92	0.85	0.98	0.88
	HSIL	0.78	0.72	0.95	0.75
	SCC	0.80	0.88	0.92	0.84

TABLE 11. Experimental results of SMOTE-SK classification model.

		Precision	Sensitivity	Specificity	F1
SMOTE-SK	Normal	0.90	0.97	0.95	0.93
	LSIL	0.96	0.81	0.99	0.88
	HSIL	0.86	0.76	0.97	0.81
	SCC	0.86	0.97	0.94	0.91

SMOTE-CNN-RF, and SMOTE-CNN-SVM, SMOTE-SK has further improved performance. This is consistent with the experimental conclusions before data enhancement. Through the comparative experiments before and after data enhancement, it can be known that the SK classification model built in this experiment has good robust stability and generalization ability.

VII. CONCLUSION

This experimental research adopts the computer-aided diagnosis mode and establishes the SK classification model based on MLP. Among them, the base learner selects CNN-KNN, CNN-RF, and CNN-SVM classifiers. During the experiment, traditional deep learning methods require high sample data volume, so the training parameters are not very detailed, and the classification effect of new instance samples is poor. Considering the applicability of machine learning to the classification ability of small sample data sets. Retained the 2048-dimensional feature vector extracted by CNN and built KNN, RF, SVM classification models. In order to further improve the classification performance, the SK algorithm in ensemble learning was selected in the experiment. The SK classification algorithm retained the classification of the first-level base learner. The discriminant output probability is imported into the MLP as a new feature, and the MLP classifier makes a final judgment on the new instance sample through the learning ability and difference of the learning base learner. In addition, this experiment uses the SMOTE algorithm to enhance the data set, and the classification results are good when the original classification model parameters are unchanged. The data enhancement comparison experiment shows that the MLP-based SK classification model built in this experiment has good generalization ability and robustness. This study explored the relationship between Normal, LSIL, HSIL, and SCC four categories of cervical pathological tissue images. The proposed SK ensemble classification model algorithm based on MLP has good classification capabilities. This experiment is an applied research, using traditional classification algorithms to apply to the classification of cervical pathological tissue images in this experiment, and achieved good classification and discrimination capabilities, aiming to provide a reference for the intelligentization of future medical diagnosis vision Classification model.

ACKNOWLEDGMENT

(Shuailei Zhang and Cheng Chen are co-first authors.)

REFERENCES

[1] A. Khamparia, D. Gupta, V. H. C. de Albuquerque, A. K. Sangaiah, and R. Jhaveri, "Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning," *J. Supercomput.*, vol. 76, pp. 1–19, Jan. 2020.

[2] D. Saslow, D. Solomon, H. W. Lawson, M. Killackey, S. L. Kulasingam, J. Cain, F. A. R. Garcia, A. T. Moriarty, A. G. Waxman, D. C. Wilbur, N. Wentzensen, L. S. Downs, M. Spitzer, A.-B. Moscicki, E. L. Franco, M. H. Stoler, M. Schiffman, P. E. Castle, and E. R. Myers, "American cancer society, American society for colposcopy and cervical pathology, and American society for clinical pathology screening guidelines for the prevention and early detection of cervical cancer," *Amer. J. Clin. Pathol.*, vol. 137, no. 4, pp. 516–542, Apr. 2012.

[3] C.-Y. Feng, R.-H. Liang, and X.-M. Jiang, "Analysis of the government health expenditure in the first decade of Chinese new medical reform (2009–2018): Xinjiang Uygur autonomous region as an example," *Risk Manage. Healthcare Policy*, vol. 13, p. 387, May 2020.

[4] Y. Wang, D. Crookes, O. S. Eldin, S. Wang, P. Hamilton, and J. Diamond, "Assisted diagnosis of cervical intraepithelial neoplasia (CIN)," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 112–121, Feb. 2009.

[5] H. Liang, M. Fu, J. Zhou, and L. Song, "Evaluation of 3D-CPA, HR-HPV, and TCT joint detection on cervical disease screening," *Oncol. Lett.*, vol. 12, no. 2, pp. 887–892, Aug. 2016.

[6] A. Robertson, J. Anderson, J. S. Beck, R. Burnett, S. Howatson, F. Lee, A. Lessells, K. McLaren, S. Moss, and J. G. Simpson, "Observer variability in histopathological reporting of cervical biopsy specimens," *J. Clin. Pathol.*, vol. 42, no. 3, pp. 231–238, 1989.

[7] L. He, L. R. Long, S. Antani, and G. Thoma, "Computer assisted diagnosis in histopathology," *Sequence Genome Anal., Methods Appl.*, vol. 3, pp. 271–287, Nov. 2010.

[8] S. De, R. J. Stanley, C. Lu, R. Long, S. Antani, G. Thoma, and R. Zuna, "A fusion-based approach for uterine cervical cancer histology image classification," *Comput. Med. Imag. Graph.*, vol. 37, nos. 7–8, pp. 475–487, Oct. 2013.

[9] L. Wei, Q. Gan, and T. Ji, "Cervical cancer histology image identification method based on texture and lesion area features," *Comput. Assist. Surg.*, vol. 22, no. 1, pp. 186–199, Oct. 2017.

[10] S. J. Keenan, J. Diamond, W. G. McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton, "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)," *J. Pathol.*, vol. 192, no. 3, pp. 351–362, 2000.

[11] P. Guo, K. Banerjee, R. J. Stanley, R. Long, S. Antani, G. Thoma, R. Zuna, S. R. Frazier, R. H. Moss, and W. V. Stoecker, "Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 6, pp. 1595–1607, Nov. 2016.

[12] J. Yuan, Y. Fan, X. Lv, C. Chen, D. Li, Y. Hong, and Y. Wang, "Research on the practical classification and privacy protection of CT images of parotid tumors based on ResNet50 model," *J. Phys., Conf. Ser.*, vol. 1576, Jun. 2020, Art. no. 012040.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[15] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2126–2136.

[16] B. V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Tutorial, 1991.

[17] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[18] J. Corcoran, J. Knight, K. Pelletier, L. Rampi, and Y. Wang, "The effects of point or polygon based training data on RandomForest classification accuracy of wetlands," *Remote Sens.*, vol. 7, no. 4, pp. 4002–4025, Apr. 2015.

[19] T. G. Pavey, N. D. Gilson, S. R. Gomersall, B. Clark, and S. G. Trost, "Field evaluation of a random forest activity classifier for wrist-worn accelerometer data," *J. Sci. Med. Sport*, vol. 20, no. 1, pp. 75–80, Jan. 2017.

[20] P. Kaur, H. S. Pannu, and A. K. Malhi, "Plant disease recognition using fractional-order zernike moments and SVM classifier," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8749–8768, Dec. 2019.

[21] Y. Liu, F. Zhang, C. Wang, S. Wu, J. Liu, A. Xu, K. Pan, and X. Pan, "Estimating the soil salinity over partially vegetated surfaces from multispectral remote sensing image using non-negative matrix factorization," *Geoderma*, vol. 354, Nov. 2019, Art. no. 113887.
[22] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[23] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.
[24] A. Ledezma, R. Aler, A. Sanchis, and D. Borrajo, "GA-stacking: Evolutionary stacked generalization," *Intell. Data Anal.*, vol. 14, no. 1, pp. 89–119, Jan. 2010.
[25] F. Sharifzadeh, G. Akbarizadeh, and Y. Seifi Kavian, "Ship classification in SAR images using a new hybrid CNN–MLP classifier," *J. Indian Soc. Remote Sens.*, vol. 47, no. 4, pp. 551–562, Apr. 2019.
[26] T. K. Bhowmik, U. Bhattacharya, and S. K. Parui, "Recognition of Bangla handwritten characters using an MLP classifier based on stroke features," in *Proc. Int. Conf. Neural Inf. Process.*, 2004, pp. 814–819.
[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[28] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, 2005, pp. 878–887.
[29] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

SHUAILEI ZHANG received the bachelor's degree in communications engineering from Xinjiang University, China, in June 2019, where she is currently pursuing the master's degree in information and communications engineering. She was recommended to be admitted to Xinjiang University, in September 2019. Her main research interests include image processing and bioinformatics.

CHEN CHEN received the bachelor's degree from Xinjiang University, China, in June 2018, where he is currently pursuing the Ph.D. degree. His main research interest includes medical signal processing.

CHENG CHEN received the bachelor's degree from Xinjiang University, China, in June 2018, where he is currently pursuing the master's degree. His main research interest includes bioinformatics.

FANGFANG CHEN received the bachelor's degree from Xinjiang University, China, in June 2019, where she is currently pursuing the master's degree. Her main research interest includes bioinformatics.

MIN LI received the bachelor's degree in computer science and technology from Shandong Women's University, China, in 2019. She is currently pursuing the master's degree in software engineering with Xinjiang University. Her current research interest includes medical image processing.

BO YANG received the B.Sc. degree from Sichuan Technology and Business University, China, in June 2019. He is currently pursuing the M.Sc. degree with Xinjiang University. His main research interests include medical image processing and analysis.

ZIWEI YAN received the bachelor's degree from Xinjiang University, China, in June 2018, where she is currently pursuing the master's degree. Her main research interest includes medical signal processing.

XIAOYI LV received the M.S. degree in information and communication engineering from Xinjiang University, China, in 2006, and the Ph.D. degree in electronic and information engineering from Xi'an Jiaotong University, China, in 2010. He is currently a Professor with the School of Software, Xinjiang University. His current work is focused upon bioinformatics and artificial intelligence in medical diagnosis.