



Research article

An ensemble-acute lymphoblastic leukemia model for acute lymphoblastic leukemia image classification

Mei-Ling Huang* and Zong-Bin Huang

Department of Industrial Engineering & Management, National Chin-Yi University of Technology, Taichung, Taiwan

* **Correspondence:** Email: huangml@ncut.edu.tw; Tel: +886423924505; Fax: +886423934620.

Abstract: The timely diagnosis of acute lymphoblastic leukemia (ALL) is of paramount importance for enhancing the treatment efficacy and the survival rates of patients. In this study, we seek to introduce an ensemble-ALL model for the image classification of ALL, with the goal of enhancing early diagnostic capabilities and streamlining the diagnostic and treatment processes for medical practitioners. In this study, a publicly available dataset is partitioned into training, validation, and test sets. A diverse set of convolutional neural networks, including InceptionV3, EfficientNetB4, ResNet50, CONV_POOL-CNN, ALL-CNN, Network in Network, and AlexNet, are employed for training. The top-performing four individual models are meticulously chosen and integrated with the squeeze-and-excitation (SE) module. Furthermore, the two most effective SE-embedded models are harmoniously combined to create the proposed ensemble-ALL model. This model leverages the Bayesian optimization algorithm to enhance its performance. The proposed ensemble-ALL model attains remarkable accuracy, precision, recall, F1-score, and kappa scores, registering at 96.26, 96.26, 96.26, 96.25, and 91.36%, respectively. These results surpass the benchmarks set by state-of-the-art studies in the realm of ALL image classification. This model represents a valuable contribution to the field of medical image recognition, particularly in the diagnosis of acute lymphoblastic leukemia, and it offers the potential to enhance the efficiency and accuracy of medical professionals in the diagnostic and treatment processes.

Keywords: medical image classification; acute lymphoblastic leukemia; deep learning; convolutional neural networks

1. Introduction

The recognition of medical diseases through image analysis relies heavily on the rapid advancement of artificial intelligence and machine learning technologies. Artificial intelligence, designed to replicate human cognitive functions, is ushering in a profound transformation in healthcare. This transformation is propelled by the growing abundance of healthcare data and the swift evolution of analytical capabilities, effectively reshaping conventional healthcare into an era of smart healthcare [1]. In recent years, the convergence of big data and deep learning technology has propelled medical disease image recognition to the forefront of research in the medical field. Historically, physicians heavily leaned on their experience and professional expertise for disease diagnosis and treatment. Nevertheless, given the immense diversity of diseases and their intricate presentations, relying solely on human judgment, doctors are susceptible to subjective biases and the risk of misdiagnosis. Furthermore, healthcare professionals must contend with a substantial volume of cases and a vast array of imaging data, resulting in an overwhelming workload. Consequently, there is a growing demand for swift and precise automated systems to support physicians in their diagnostic endeavors. Leveraging machine learning and deep learning techniques, these systems can autonomously scrutinize and detect anomalies in medical images, providing valuable aid to doctors in the process of diagnosing and treating diseases. Currently, medical disease image recognition technology has found extensive applications, with research conducted in various domains, such as brain tumors [2], diabetic retinopathy [3], lung diseases [4], and more. A meticulously crafted convolutional neural network (CNN) model not only offers valuable reference for medical professionals but also has the capability to automatically outline or pinpoint the specific areas of interest for observation through computer-assisted detection. This capability aids physicians in promptly interpreting medical images and taking the appropriate actions based on the findings.

One of the most prevalent form of cancer in children is acute lymphoblastic leukemia (ALL) [5–7], a condition that impacts white blood cells (WBCs). In ALL, patients typically exhibit an overabundance of immature WBCs within the bone marrow. These immature WBCs can potentially disseminate to other organs, including the spleen, liver, lymph nodes, central nervous system, and testicles. It is noteworthy that a substantial proportion of ALL cases, approximately 55% of the global total, are concentrated in the Asia-Pacific region [7].

The CNN stands as one of the prominent techniques in the realm of deep learning. It excels in training and categorizing extensive volumes of intricate data through its model. CNNs find application in a diverse range of domains, encompassing image recognition [8–10], and voice recognition [11–14]. The architecture of a CNN typically comprises one or more convolutional layers, a pooling layer, a rectified linear units layer (ReLU layer), a fully connected layer, and various hyperparameters, among other components.

Jawahar et al. [15] introduced a cutting-edge deep neural network model called ALNett, which leverages deep convolutions featuring varying dilation rates for the classification of microscopic leukocyte images. The model's performance was assessed against several pre-trained models, including VGG16, ResNet-50, GoogleNet, and AlexNet. The experimental findings revealed that the proposed ALNett model achieved an impressive classification accuracy of 91.13%, all while maintaining low computational complexity. Mondal et al. [7] also contributed to the field by deploying a deep CNN to identify ALL. Their study delved into the development of weighted ensembles of deep CNN with the aim of recommending improved classifiers for ALL diagnosis.

Das and Meher [16] introduced an effective deep CNN framework designed to enhance the accuracy of ALL detection. This framework incorporates innovative components, including depth-wise separable convolution, a linear bottleneck architecture, reverse residual structures, and skip connections. As a result, it not only achieves greater speed but also gains popularity in the field. A key feature of this method is the introduction of a novel probability-based weighting factor, skillfully combining the strengths of MobilenetV2 and ResNet18 while preserving the advantages of both approaches. Anilkumar et al. [17] have effectively classified ALL using deep CNNs, achieving an impressive accuracy rate of 94.12%. In a related study, Duggal et al. [18] introduced an innovative approach by incorporating a stain de-convolutional layer at the front of a CNN. They utilized this model to discriminate between malignant immature white blood cells and benign immature white blood cells, thereby enabling the detection of ALL. All three of these studies have contributed significantly to the field by proposing deep neural network models that leverage deep convolution and feature extraction techniques to successfully classify and identify ALL images.

In 2022, Ghaderzadeh et al. [19] assembled a dataset for ALL and conducted a comprehensive study. They employed ten different CNN models, including EfficientNet, MobileNetV3, VGG-19, Xception, InceptionV3, ResNet50V2, VGG-16, NASNetLarge, InceptionResNetV2, and DenseNet201, for various feature extraction tasks. Among these ten models, DenseNet201 emerged as the top-performing model, achieving exceptional classification results with an accuracy of 99.85%, sensitivity of 99.52%, and specificity of 99.89%. Panthakkan et al. [20] devised an innovative approach by integrating the Xception and ResNet50V2 models, ultimately creating a $4 \times 4 \times 2048$ feature map in the final layer of feature extraction. They then concatenated and integrated the features extracted from Xception and ResNet50V2 into a 1×1 convolutional layer. This concatenated feature representation was subsequently combined with a classifier to form a unified deep network. The purpose of this proposed model is to provide valuable assistance to dermatologists and clinicians in the early detection of skin cancer within the clinical process.

Elashiri et al. [21] conducted a study in which they integrated features extracted by three different models: ResNet50, VGG16, and DeepLabv3. These integrated features were then passed to the feature transformation stage, employing weighted feature extraction facilitated by a hybrid squirrel butterfly search optimization technique. The transformed features were subsequently input to a modified long short-term memory (LSTM) for the final classification process. Remarkably, this approach yielded the highest accuracy rate, reaching an impressive 93%. Silva et al. [22] introduced an EnsembleDVX model, which combines the capabilities of DenseNet169, VGG16, and Xception in the context of COVID-19 detection. The ensemble EnsembleDVX model demonstrates superior performance when compared to individual single models. Additionally, Aurna et al. [23] introduced a two-phase ensemble model designed for the accurate and automated detection of brain tumors. This ensemble model effectively combines features extracted from five individual models and, through the use of principal component analysis, selects the most significant combinations of these features. The results obtained from the ensemble model surpass the performance of the single models, demonstrating its efficacy in precise brain tumor detection. Su et al. [24] utilized a combination of deep learning and ensemble learning techniques to tackle the challenging task of glioma grading, specifically focusing on the classification between glioma grades II and III. They constructed an ensemble model called LR-14, which is based on logistic regression and integrates the predictions from 14 different deep learning classifiers. This ensemble LR-14 model emerged as the top-performer, achieving the highest level of performance for this critical classification task.

As indicated by the examples mentioned, ensemble models have proven their effectiveness in the realm of medical classification. The outcomes from these integrated models have generally outperformed those from single models, underlining the value of combining different approaches for improved results. However, it is worth noting that many ensemble models tend to be complex with a significant number of parameters, which can lead to longer training times and increased computational demands. Hence, our primary objective of this study is to introduce a streamlined ensemble model with fewer parameters. The aim is to reduce training time while achieving enhanced stability and accuracy compared to single models. This approach seeks to strike a balance between computational efficiency and performance in the context of medical classification, potentially offering a more practical and accessible solution.

2. Materials and methods

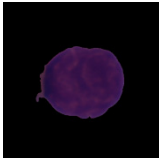

2.1. ALL challenge dataset of ISBI 2019 (C-NMC 2019)

ALL accounts for around 25% of all pediatric cancers [25]. Identifying immature leukemic blasts from normal cells under a microscope is a challenging task because, morphologically, the images of these two cell types can be quite similar. The dataset used in this study was from the ALL Challenge dataset of ISBI 2019 (C-NMC 2019) [26]. The whole training set comprises a total of 73 subjects (10661 images), including 46 with ALL (7272 images) and 26 with Normal conditions (3,389 images). The training set is split into three folds including fold0, fold1, and fold2 (Table 1). Among the three folds, fold0, containing 2397 ALL images and 1,130 normal images, is split into training, validation, and test sets in this study. Table 2 provides an overview of example images and the number of training, test, and validation sets used in this investigation.

Table 1. The number of images for C-NMC 2019 dataset.

	ALL	Normal	Total
train_fold0	2397	1130	3527
train_fold1	2418	1163	3581
train_fold2	2457	1096	3553
Total	7272	3389	10661

Table 2. Examples and number of images used in this study.

Class	Example	Training	Test	Validation	Total
ALL		1535	479	383	2397
Normal		724	226	180	1130
Total		2259	705	563	3527

2.2. Methodological framework

The proposed ensemble-ALL model is outlined and introduced as follows:

- 1) Data Splitting (Step 1): The initial dataset is divided into distinct subsets, including training, validation, and test sets.
- 2) Individual Model Training (Step 2): In addition to the well-known models, including InceptionV3, EfficientNetB4, ResNet50, and AlexNet, we have selected three models, ALL-CNN [27], CONV_POOL-CNN [27], and Network in Network (NIN-CNN) [28], with simple architecture in this study. The above seven models are independently applied to the C-NMC 2019 dataset.
- 3) Model Selection and Enhancement (Step 3): The top-performing four individual models are meticulously selected and integrated with the squeeze-and-excitation (SE) module to evaluate whether this fusion enhances performance.
- 4) Ensemble Model Formation (Step 4): The two most promising SE-embedded models are combined to create the proposed ensemble-ALL model.
- 5) Bayesian Optimization (Step 5): Bayesian optimization techniques are harnessed within the ensemble-ALL model to enhance its performance.
- 6) Cross-Validation (Step 6): A five-fold cross-validation strategy is employed to rigorously assess the model's performance.
- 7) Performance Evaluation (Step 7): Performance metrics including accuracy, precision, recall, F1-score, and kappa are recorded and evaluated for all models.

This structured approach shown in Figure 1 ensures the creation of a robust and optimized ensemble-ALL model for the task of medical image classification.

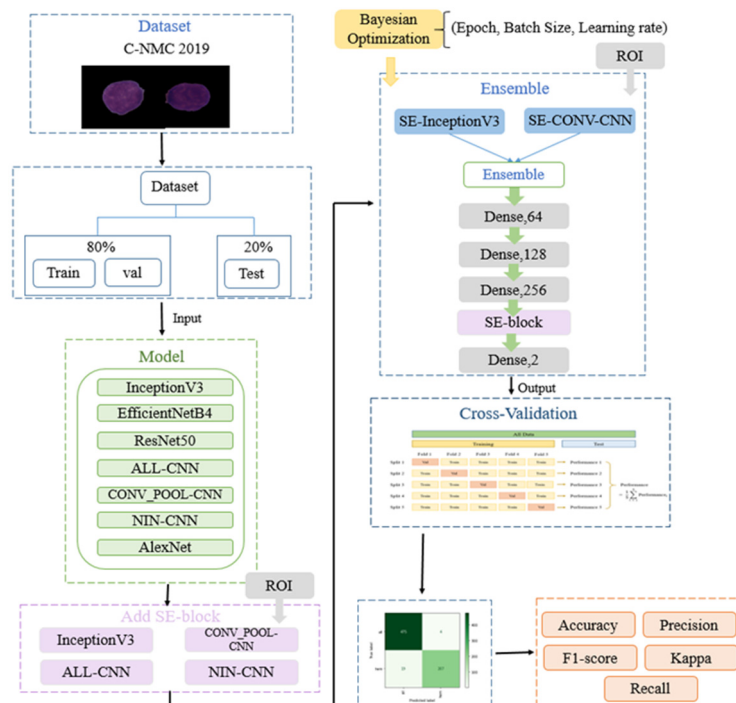


Figure 1. Flowchart of the study.

2.3. Squeeze-and-excitation

Squeeze-and-Excitation Networks, known as SE Networks, secured the championship in the ImageNet 2017 classification competition. The SE module’s concept is straightforward, easy to implement, and can be seamlessly integrated into existing network model frameworks [29]. Consequently, SE modules have been incorporated into models in numerous academic works [29–34]. It’s evident that the addition of SE attention modules consistently leads to improved accuracy results in the studies, outperforming the original models. For instance, you can find examples such as SE-MobileNet [30], SE-ResNext [31], SE-ResNet-18 [32], and SE-DenseNet [33], all of which have demonstrated enhanced performance thanks to the inclusion of SE attention modules.

The decision to employ the SE module to enhance InceptionV3 in this study is rooted in the observation that many existing studies primarily focus on the SE-ResNet model, with only a limited number exploring SE-InceptionV3. Additionally, deeper network structures, such as the one in InceptionV3, are susceptible to feature distortion. To mitigate this, many studies have introduced attention modules into deeper network models [30,33,34]. Consequently, in this study, the SE module is introduced after the global average pooling layer of InceptionV3 to address these considerations. Figures 2–5. provide a visual representation of the architectures of the SE-InceptionV3, SE-CONV_POOL-CNN, SE-ALL-CNN, and SE-NIN-CNN models employed in the study.

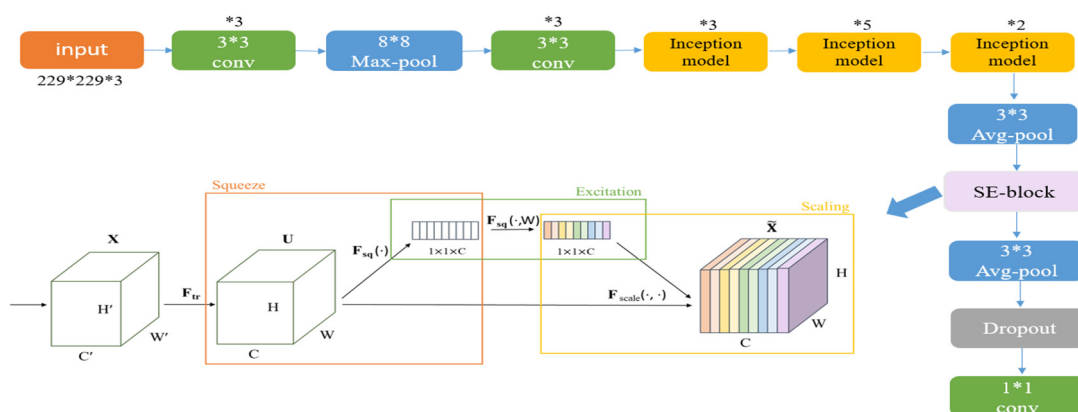


Figure 2. Architecture of SE-InceptionV3.

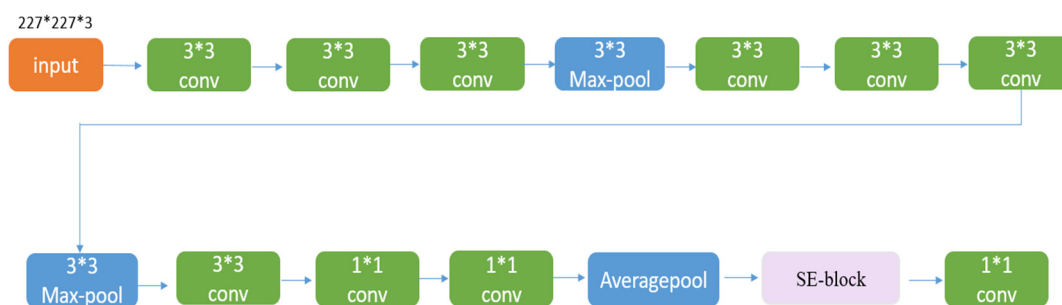


Figure 3. Architecture of SE-CONV_POOL-CNN.

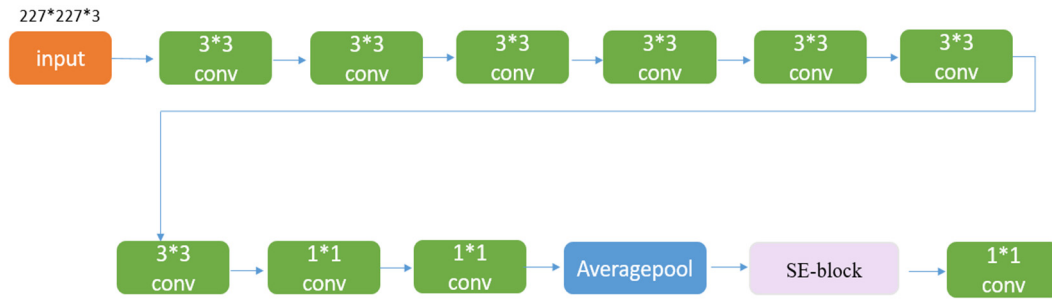


Figure 4. Architecture of SE-ALL-CNN.

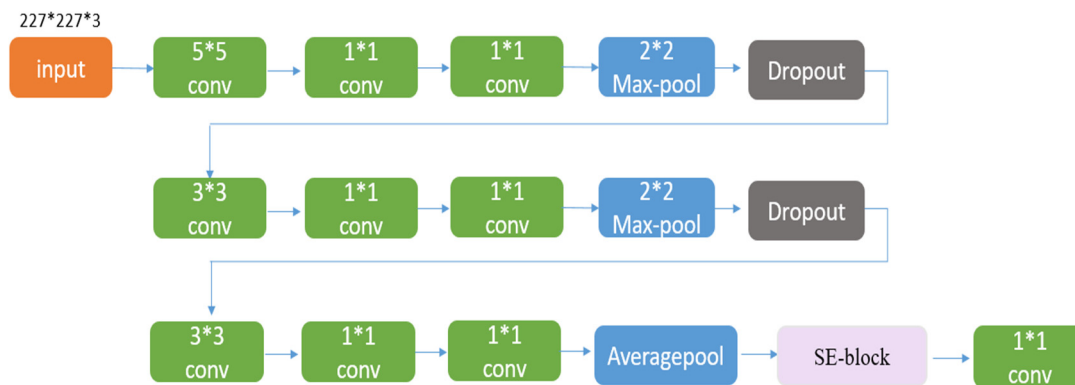


Figure 5. Architecture of SE-NIN-CNN.

2.4. Region of interest

To streamline the feature extraction process and remove extraneous elements, the study employs a region of interest (ROI) feature extraction method within image processing. Specifically, the black background has a different proportion in each image, we can directly outline the ROI area through the pixel matrix [100:350, 200:400] from the images. Figure 6(a). is an example of original image. The proportion of black background is greatly reduced in Figure 6(b). It's important to note that, given the variability in feature sizes in each picture, the elimination of the black background may not be entirely comprehensive in all cases.



Figure 6. (a) Original image, (b) ROI.

2.5. Bayesian optimization

Bayesian optimization is a frequently employed technique for optimizing highly complex or non-differentiable functions, often referred to as black-box functions [35]. Bayesian optimization algorithms have extensive application in areas such as automated hyperparameter tuning, machine learning, and deep learning. In the context of this study, Bayesian optimization is utilized to determine appropriate hyperparameters, including the number of epochs (ranging from 10 to 100), batch size (ranging from 8 to 32), and learning rate (ranging from 1×10^{-5} to 1×10^{-2}). This approach enables the automatic selection of optimal hyperparameters for the task at hand. Due to the limitation of computer equipment, the execution of hyperparameter optimization is time consuming. It is not applied foron the training for seven individual CNN models in Section 2.2 Step 2, and is solely used for training the ensemble model in Section 2.2 Step 4 in this study.

2.6. Five-fold cross-validation

The initial step involves partitioning all the available data into two distinct sets: a training set and a test set. Subsequently, the training set is further divided into five approximately equal and non-overlapping validation folds. During each iteration of the process, one of these folds is designated as the validation set, while the remaining folds are combined to form the training set. To illustrate, in Fold1, the training set is composed of the validation sets from Folds 2–5. The model’s performance is assessed using the validation set (Val) in each iteration, and this process is repeated until every fold has been used as the validation set. This method, known as five-fold cross-validation, is visually depicted in Figure 7. to provide a clear overview of the process.

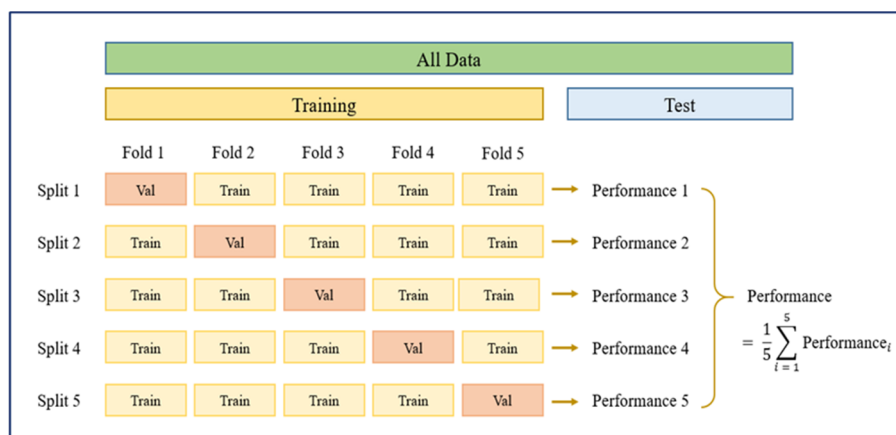


Figure 7. The process of five-fold cross-validation.

2.7. Computational environment

The equipment used in the experiment is an Intel(R) Core(TM) i7-10700 3.80GHz 32GB CPU, NVIDIA GeForce RTX 3060 GPU. The whole experiment process is performed using Python 3.8 [Python Software Foundation, Fredericksburg, Virginia, USA], which contains Keras 2.6 and Tensorflow GPU 2.6.0.

3. Results

3.1. Performance of individual models

The results of individual models are recorded in Table 3. The top four highest accuracies are 90.78, 90.35, 89.21, and 84.82%, which are from CONV_POOL-CNN, ALL-CNN, NIN-CNN, and InceptionV3 models. The four best-performing models are chosen for further enhancement. These models are combined with the SE module to create four modified models, denoted as SE-InceptionV3, SE-CONV_POOL-CNN, SE-ALL-CNN, and SE-NIN-CNN. This integration with the SE module is intended to improve their performance and effectiveness.

Table 3. Results of individual models.

Model	Accuracy	Precision	Recall	F1-score	kappa
ResNet50	82.83%	84.96%	82.84%	83.27%	63.22%
InceptionV3	84.82%	85.63%	84.82%	85.06%	66.37%
EfficientnetB4	83.26%	83.57%	83.26%	82.22%	57.93%
CONV_POOL-CNN	90.78%	90.70%	90.78%	90.69%	78.45%
ALL-CNN	90.35%	90.33%	90.35%	90.34%	77.80%
NIN-CNN	89.21%	89.97%	89.22%	88.69%	73.31%
AlexNet	84.11%	83.83%	84.11%	83.86%	62.47%

3.2. Performance of selected SE models

Performance of SE-InceptionV3, SE-CONV_POOL-CNN, SE-ALL-CNN, and SE-NIN-CNN models are listed in Table 4. Compared with Table 3, The SE module has a significant improvement on InceptionV3 model, but has less improvement on the other three models. The top two accuracies, precisions, recalls, F1-scores and kappas are from SE-InceptionV3 and SE-CONV_POOL-CNN models. The accuracy 84.82% and kappa 66.37% from InceptionV3 model were greatly increased to 95.03 and 88.69% for SE-InceptionV3 model. The accuracy 90.78% from CONV_POOL-CNN model was slightly increased to 91.34% for SE- CONV_POOL-CNN model.

Table 4. Results of individual SE models.

Model	Accuracy	Precision	Recall	F1-score	Kappa
SE- InceptionV3	95.03%	95.10%	95.04%	95.06%	88.69%
SE-CONV_POOL-CNN	91.34%	91.41%	91.35%	91.16%	79.39%
SE- ALL-CNN	89.07%	89.34%	89.08%	89.17%	75.35%
SE- NIN-CNN	87.51%	87.36%	87.52%	87.34%	70.58%

3.3. Performance of the proposed ensemble-ALL model

The two best-performing SE models, SE-InceptionV3 and SE-CONV_POOL-CNN, are carefully chosen and integrated into the proposed ensemble-ALL model, as depicted in Figure 8. In this architecture, the convolutional layer replaces the traditional fully connected layer to serve

as the output of a single model. Subsequently, these outputs are merged and fed into the fully connected layer, effectively increasing the number of features. Following this, three fully connected layers are employed, and an SE module is introduced to preserve critical features. Finally, the output of the fully connected layer serves as the output of the ensemble model. This approach optimizes the model's architecture for the task of medical image classification.

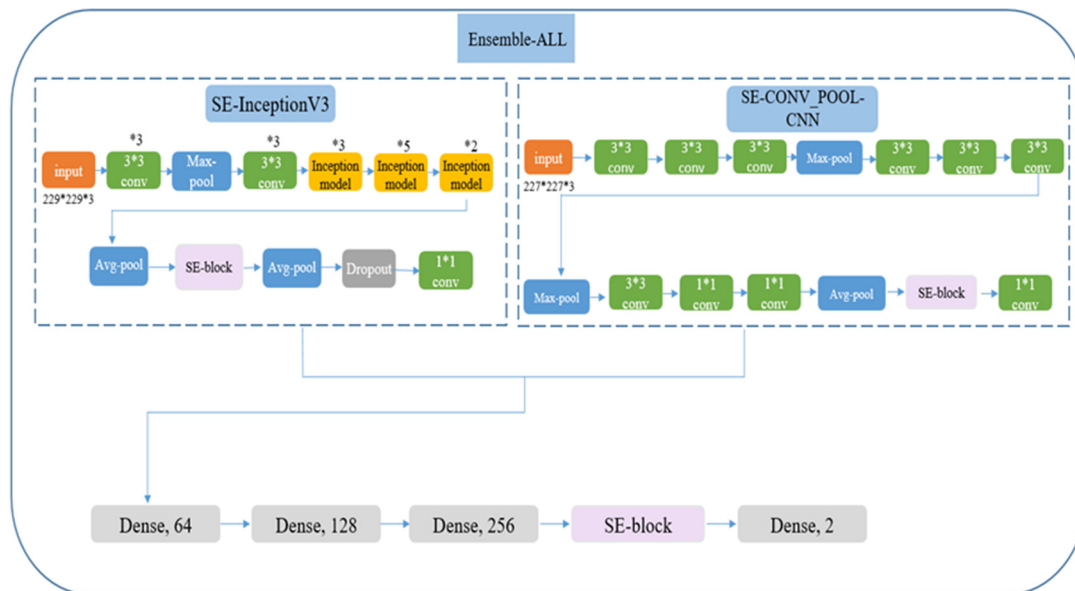


Figure 8. The proposed ensemble-ALL model.

In the study, the Bayesian optimization method is employed to identify the optimal hyperparameter combinations. Specifically, the chosen parameter values are as follows: Epoch = 50, Batch size = 15, and learning rate at 1.11×10^{-05} . The results of the five-fold cross-validation for the SE-InceptionV3, SE-CONV_POOL-CNN, and the proposed ensemble-ALL models are presented in Tables 5–7, respectively. These tables likely provide valuable insights into the performance and efficacy of these models in the context of medical image classification. The average of accuracy, precision, recall, F1-score and kappa from SE-InceptionV3 were 94.893, 95.054, 94.894, 94.877, and 88.229%, respectively. The accuracy, precision, recall, F1-score and kappa from SE-CONV_POOL-CNN were 89.985, 90.178, 89.986, 89.987, and 77.066%, respectively. The best accuracy, precision, recall, F1-score and kappa come from the proposed ensemble-ALL at 96.255, 96.259, 96.255, 96.246, and 91.362%, respectively. Figure 9 illustrates the performance of the three models, and it's clear that the proposed ensemble-ALL model outperforms the other two SE models across all major metrics. This highlights the effectiveness of the ensemble-ALL model in achieving superior results in the context of medical image classification.

The training curves for accuracy, loss, and the confusion matrix for SE-InceptionV3, SE-CONV_POOL-CNN, and the proposed ensemble-ALL models are presented in Figures 10–12. A comparison of these curves reveals that the proposed ensemble-ALL model exhibits higher accuracy, more stable and converged loss, and smoother training patterns in comparison to the two SE models. These observations further reinforce the superior performance of the ensemble-ALL model in the context of medical image classification.

Table 5. Results of SE-InceptionV3 model.

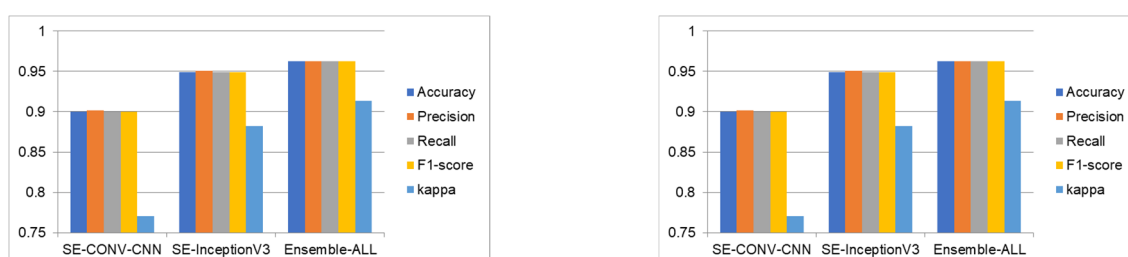
SE-InceptionV3	Accuracy	Precision	Recall	F1-score	Kappa
fold1	95.319%	95.306%	95.319%	95.311%	89.216%
fold2	95.319%	95.357%	95.319%	95.332%	89.316%
fold3	94.893%	95.126%	94.894%	94.943%	88.519%
fold4	94.042%	94.460%	94.043%	93.873%	85.653%
fold5	94.893%	95.022%	94.894%	94.927%	88.439%
AVG	94.893%	95.054%	94.894%	94.877%	88.229%
SD	0.00521	0.00359	0.00521	0.00594	0.01493
Confidence interval	94.25–95.54%	94.61–95.50%	94.25–95.54%	94.14–95.61%	86.37–90.08%

Table 6. Results of SE-CONV_POOL-CNN model.

SE-CONV-CNN	Accuracy	Precision	Recall	F1-score	Kappa
fold1	89.219%	89.307%	89.220%	89.256%	75.424%
fold2	90.780%	90.978%	90.780%	90.846%	79.151%
fold3	90.212%	90.308%	90.213%	89.958%	76.516%
fold4	90.496%	90.533%	90.496%	90.513%	78.259%
fold5	89.219%	89.763%	89.220%	89.363%	75.982%
AVG	89.985%	90.178%	89.986%	89.987%	77.066%
SD	0.00728	0.00655	0.00727	0.00696	0.01577
Confidence interval	89.08–90.89%	89.36–90.99%	89.08–90.89%	89.12–90.85%	75.11–79.02%

Table 7. Results of ensemble-ALL model.

Ensemble-ALL	Accuracy	Precision	Recall	F1-score	Kappa
fold1	95.744%	95.732%	95.745%	95.724%	90.139%
fold2	96.453%	96.473%	96.454%	96.425%	91.733%
fold3	96.595%	96.589%	96.596%	96.592%	92.166%
fold4	96.312%	96.336%	96.312%	96.321%	91.573%
fold5	96.170%	96.166%	96.170%	96.168%	91.197%
AVG	96.255%	96.259%	96.255%	96.246%	91.362%
SD	0.00326	0.00334	0.00327	0.00330	0.00767
Confidence interval	95.85–96.66%	95.84–96.67%	95.85–96.66%	95.84–96.66%	90.41–92.31%

**Figure 9.** Performance of three models.

SE-InceptionV3

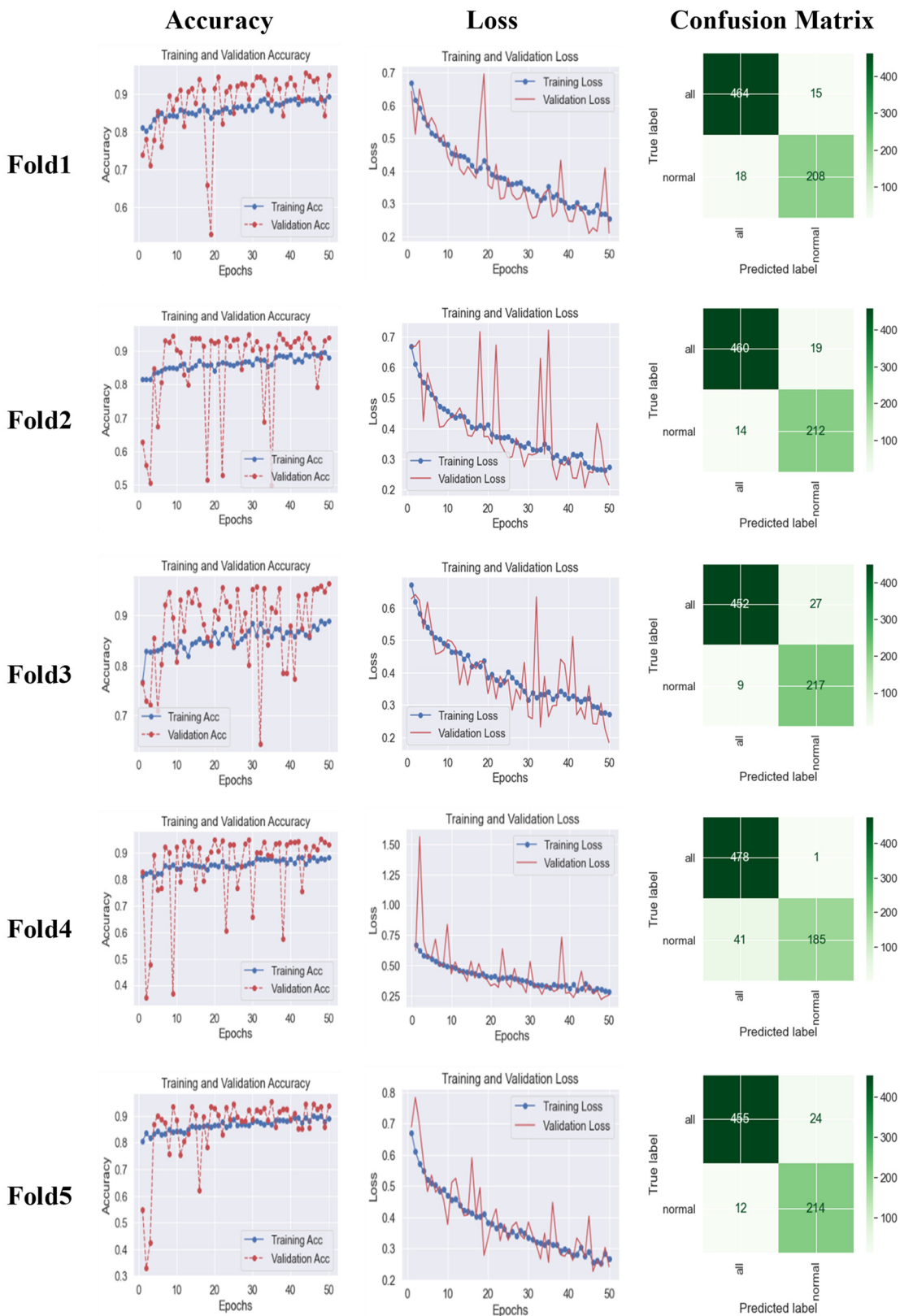


Figure 10. The accuracy, loss and confusion matrix for SE-InceptionV3 model.

SE-CONV_POOL-CNN

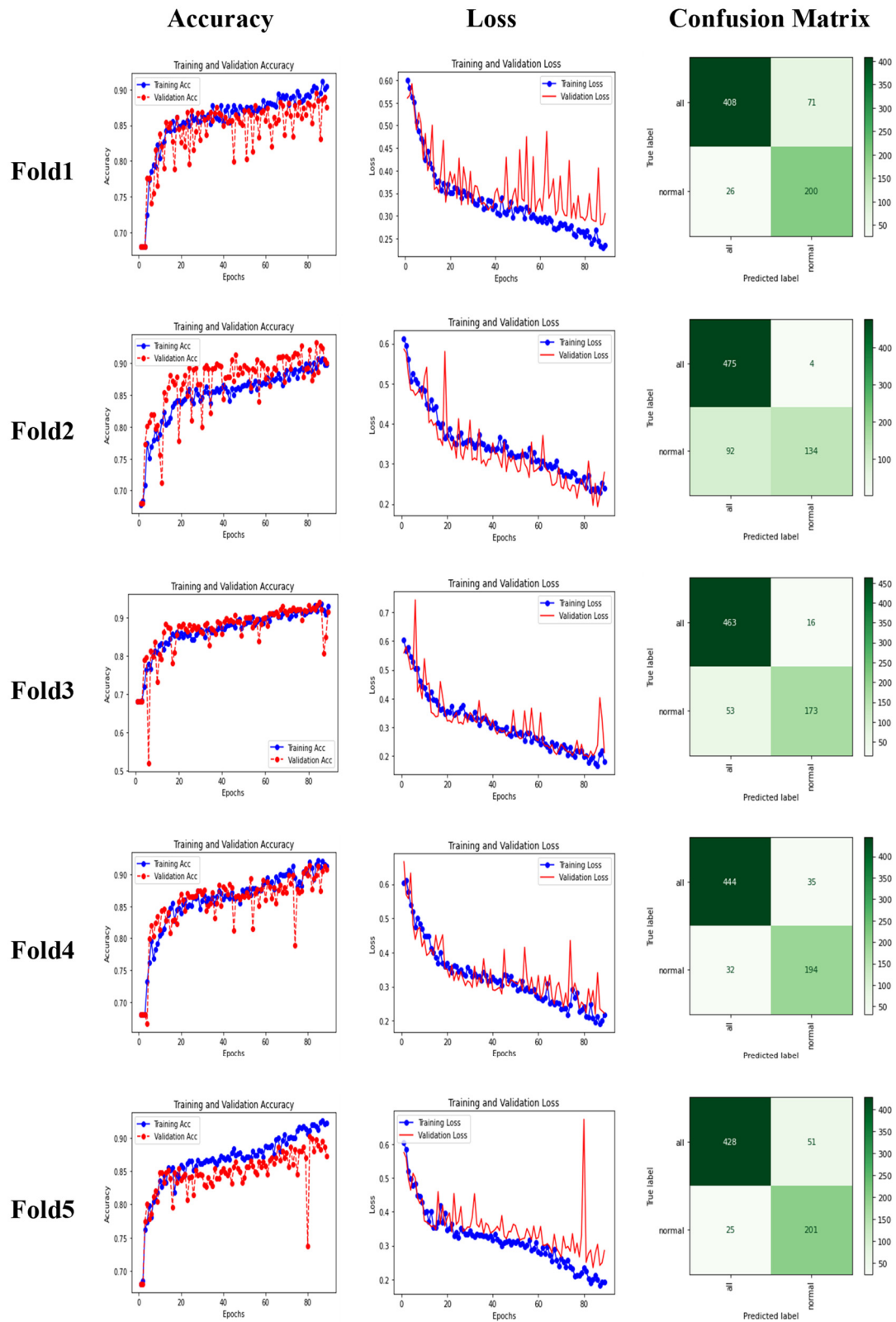


Figure 11. The accuracy, loss and confusion matrix for SE-CONV_POOL-CNN model.

Ensemble-ALL (SE-InceptionV3+SE-CONV)

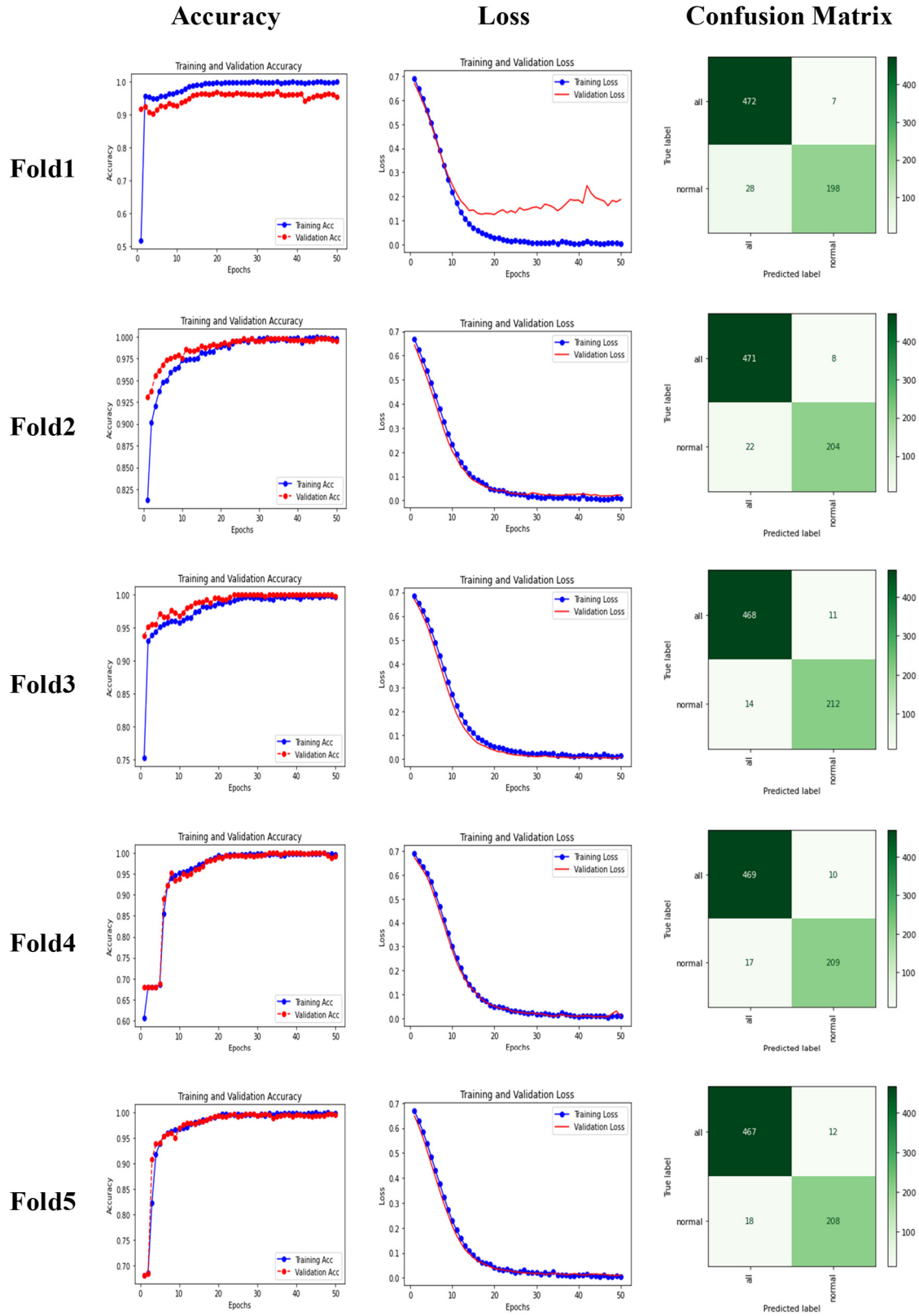


Figure 12. The accuracy, loss and confusion matrix for ensemble-ALL model.

4. Discussion

4.1. Selection process of the proposed ensemble-ALL model

The top-performing four out of seven individual models are selected and integrated with the SE module. Then, the two most promising SE-embedded models are combined to create the proposed ensemble-ALL model in this study. Definitely, the proposed model is not the only solution. We have completed many different selections among models in the experimental processes, and finally the proposed model stood out all combinations. Some of the results are presented in Table 8 as follows:

Table 8. Results of Different Ensemble models.

Model	Accuracy	Precision	Recall	F1-score	kappa
InceptionV3 + NIN + ALL	94.89%	96.33%	92.15%	93.88%	87.79%
InceptionV3 + ALL + CONV	93.19%	93.07%	91.13%	92.01%	84.03%
InceptionV3 + NIN + ALL + CONV	94.18%	92.59%	94.67%	93.49%	86.99%
NIN + ALL + CONV	92.19%	91.52%	90.40%	90.93%	81.85%
NIN + ALL + AlexNet	90.78%	90.98%	90.78%	90.85%	79.15%
ALL + CONV + AlexNet	91.48%	91.70%	91.49%	91.55%	80.77%
NIN + CONV + AlexNet	90.35%	90.54%	90.36%	90.42%	78.16%

In addition, due to limitations of computer equipment, the author also needs to consider the amount of computer calculations. This is part of the reason that only two models are selected in the ensemble stage.

4.2. Comparison with SOTA

Table 9 offers a comparative analysis of the results achieved by the proposed ensemble-ALL model with those obtained in related studies that used the same dataset. The comparison includes findings from Duggal et al. [18] and Mondal et al. [7], where some ensemble models were employed. Notably, the results highlight that the proposed ensemble-ALL model outperforms all other models considered in this analysis on the C-NMC 2019 dataset. This underscores the model's effectiveness and its capacity to yield superior performance in the domain of medical image classification.

5. Conclusions

The ensemble-ALL model developed in this study represents a significant advancement in the field of medical image classification, particularly in the context of ALL. This ensemble model is designed to improve the early diagnosis and classification of ALL, enhancing the efficiency of medical professionals in the diagnostic and treatment process. When compared to related state-of-the-art studies, the proposed ensemble-ALL model stands out by achieving remarkable performance metrics, including accuracy, precision, recall, F1-score, and kappa, with values of 96.26, 96.26, 96.26, 96.25, and 91.36%, respectively. These results underscore the efficacy of the ensemble model in the accurate classification of ALL images. Key highlights and components of the ensemble-ALL model include:

- 1). Data Splitting: The original dataset is divided into training, validation, and test sets.
- 2). Individual Models: Several deep learning models, including InceptionV3, EfficientNetB4, ResNet50, ALL-CNN, CONV_POOL-CNN, Network in Network (NIN-CNN), and AlexNet, are individually applied to the C-NMC 2019 dataset.
- 3). Model Selection and Enhancement: The top-performing individual models are selected and enhanced with the SE module to boost performance.
- 4). Ensemble Formation: The two best-performing SE models, SE-InceptionV3 and SE-CONV_POOL-CNN, are combined to create the ensemble-ALL model.
- 5). Feature Extraction and Optimization: Feature extraction from ROI and Bayesian optimization techniques are utilized within the ensemble-ALL model to enhance its performance.
- 6). Cross-Validation: A five-fold cross-validation strategy is employed to rigorously assess the model's performance.
- 7). Performance Metrics: Performance indices, including accuracy, precision, recall, F1-score, and kappa, are recorded for evaluation.

Table 9. Performance comparison of the proposed method with state-of-the-art methods on C-NMC 2019 dataset.

Literatures	Class	Methods	Accuracy	Recall	F1-score	Precision
Mondal et al. [7]	2	VGG-16	84.20%	—	—	—
		Xception	85.30%	—	—	—
		MobileNet	84.10%	—	—	—
		InceptionResNetV2	84.30%	—	—	—
		DenseNet-121	82.10%	—	—	—
		Ensemble models (WEN-auc)	88.60%	—	—	—
		Ensemble models (WEN-f1)	88.70%	—	—	—
Duggal et al. [18]	2	AlexNet	88.50%	—	—	—
		Texture-CNN	93.20%	—	—	—
		ResNet50	82.83%	82.84%	83.27%	84.96%
		InceptionV3	84.82%	84.82%	85.06%	85.63%
		EfficientnetB4	83.26%	83.26%	82.22%	83.57%
		CONV_POOL-CNN	90.78%	90.78%	90.69%	90.70%
		ALL-CNN	90.35%	90.35%	90.34%	90.33%
This study	2	NIN-CNN	89.21%	89.22%	88.69%	89.97%
		AlexNet	84.11%	84.11%	83.86%	83.83%
		SE-CONV_POOL-CNN	89.99%	89.99%	89.99%	90.18%
		SE-InceptionV3	94.89%	94.89%	94.88%	95.05%
		ensemble-ALL	96.26%	96.26%	96.25%	96.26%

The ensemble-ALL model, as demonstrated in the study, has shown remarkable results, outperforming existing state-of-the-art models in image classification for ALL.

While the proposed ensemble-ALL model demonstrates notable performance, there are several limitations to the study that should be acknowledged:

- 1) Limited datasets for ALL: Public datasets specifically tailored for ALL are relatively scarce compared to datasets for other diseases. Consequently, the ensemble-ALL model in this study is exclusively applied to a single ALL dataset, which may not capture the full spectrum of variability that exists across different datasets.
- 2) Hardware limitations: The study is constrained by hardware limitations, particularly in terms of graphics card capabilities and memory. Large datasets, extensive training times, and models with substantial parameters (e.g., EfficientNetB5 to EfficientNetB7, ConvNeXt, etc.) can strain hardware resources. This limitation may hinder the exploration of larger and more complex models.
- 3) Model parameter size vs. accuracy: Our focus of this study centers on optimizing model accuracy. However, the trade-off between model accuracy and the size of model parameters is not deeply explored. Future research could consider achieving high accuracy while minimizing the size of model parameters, which is crucial for real-world applications where resource constraints may be a concern.

These limitations provide valuable insights for future research endeavors, and they underscore the need for ongoing exploration and development in the field of medical image classification.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors gratefully acknowledge the financial support of the Ministry of Science and Technology of Taiwan, R.O.C., through the grant MOST 111-2221-E-167-007-MY3.

Conflict of interest

The authors declare that there are no conflicts of interest.

Data and code availability

The data used to support the findings of this study are available from <https://doi.org/10.7937/tcia.2019.dc64i46r>. The code to generate the results and figures is available of the corresponding author upon request.

References

1. F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, et al., Artificial intelligence in healthcare: Past, present and future, *Stroke Vasc. Neurol.*, **2** (2017), 230–243. <https://doi.org/10.1136/svn-2017-000101>
2. J. Kang, Z. Ullah, J. Gwak, Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers, *Sensors*, **21** (2021), 1–21. <https://doi.org/10.3390/s21062222>

3. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy, *Procedia Comput. Sci.*, **90** (2016), 200–205. <https://doi.org/10.1016/j.procs.2016.07.014>
4. P. Zhai, Y. Tao, H. Chen, T. Cai, J. Li, Multi-task learning for lung nodule classification on chest CT, *IEEE Access*, **8** (2020), 180317–180327. <https://doi.org/10.1109/ACCESS.2020.3027812>
5. American Cancer Society, Leukemia in Children, 2023. Available from: <https://www.cancer.org/cancer/types/leukemia-in-children.html>
6. D. Bhojwani, J. J. Yang, C. Pui. Biology of childhood acute lymphoblastic leukemia, *Pediatr. Clin. North Am.*, **62**(2015), 47–60. <https://doi.org/10.1016/j.pcl.2014.09.004>
7. C. Mondal, M. K. Hasan, M. Ahmad, M. A. Awal, M. T. Jawad, A. Dutta, et al., Ensemble of convolutional neural networks to diagnose acute lymphoblastic leukemia from microscopic images, *Inf. Med. Unlocked*, **27** (2021), 100794. <https://doi.org/10.1016/j.imu.2021.100794>
8. M. M. Hasan, M. M. Hossain, M. M. Rahman, A. Azad, S. A. Alyami, M. A. Moni, FP-CNN: Fuzzy pooling-based convolutional neural network for lung ultrasound image classification with explainable AI, *Comput. Biol. Med.*, **165** (2023), 107407. <https://doi.org/10.1016/j.combiomed.2023.107407>
9. O. Uparkar, J. Bharti, R. K. Pateriya, R. K. Gupta, A. Sharma, Vision transformer outperforms deep convolutional neural network-based model in classifying X-ray images, *Procedia Comput. Sci.*, **218** (2023), 2338–2349. <https://doi.org/10.1016/j.procs.2023.01.209>
10. A. Shakarami, L. Nicolè, M. Terreran, A. P. D. Tos, S. Ghidoni, TCNN: A transformer convolutional neural network for artifact classification in whole slide images, *Biomed. Signal Process. Control*, **84** (2023), 104812. <https://doi.org/10.1016/j.bspc.2023.104812>
11. K. K. Lella, A. Pja, Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath, *Alexandria Eng. J.*, **61** (2022), 1319–1334. <https://doi.org/10.1016/j.aej.2021.06.024>
12. M. Liu, A. N. J. Raj, V. Rajangam, K. Ma, Z. Zhuang, S. Zhuang, Multiscale-multichannel feature extraction and classification through one-dimensional convolutional neural network for Speech emotion recognition, *Speech Commun.*, **156** (2024), 103010. <https://doi.org/10.1016/j.specom.2023.103010>
13. V. Singh, S. Prasad, Speech emotion recognition system using gender dependent convolution neural network, *Procedia Comput. Sci.*, **218** (2023), 2533–2540. <https://doi.org/10.1016/j.procs.2023.01.227>
14. F. Adolphi, J. S. Bowers, D. Poeppel, Successes and critical failures of neural networks in capturing human-like speech recognition, *Neural Networks*, **162** (2023), 199–211. <https://doi.org/10.1016/j.neunet.2023.02.032>
15. M. Jawahar, S. H, J. A. L, A. H. Gandomi, ALNett: A cluster layer deep convolutional neural network for acute lymphoblastic leukemia classification, *Comput. Biol. Med.*, **148** (2022), 105894. <https://doi.org/10.1016/j.combiomed.2022.105894>
16. P. K. Das, S. Meher, An efficient deep convolutional neural network based detection and classification of Acute Lymphoblastic Leukemia, *Expert Syst. Appl.*, **183** (2021), 115311. <https://doi.org/10.1016/j.eswa.2021.115311>
17. K. K. Anilkumar, V. J. Manoj, T. M. Sagi, Automated detection of B cell and T cell acute lymphoblastic leukaemia using deep learning, *IRBM*, **43** (2022), 405–413. <https://doi.org/10.1016/j.irbm.2021.05.005>

18. R. Duggal, A. Gupta, R. Gupta, P. Mallick, SD-Layer: Stain deconvolutional layer for CNNs in medical microscopic imaging, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, **10435** (2017), 435–443. https://doi.org/10.1007/978-3-319-66179-7_50
19. M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, H. Abolghasemi, A fast and efficient CNN model for B-ALL diagnosis and its subtypes classification using peripheral blood smear images, *Int. J. Intell. Syst.*, **37** (2022), 5113–5133. <https://doi.org/https://doi.org/10.1002/int.22753>
20. A. Panthakkan, S. M. Anzar, S. Jamal, W. Mansoor, Concatenated Xception-ResNet50-A novel hybrid approach for accurate skin cancer prediction, *Comput. Biol. Med.*, **150** (2022), 106170. <https://doi.org/10.1016/j.compbimed.2022.106170>
21. M. A. Elashiri, A. Rajesh, S. N. Pandey, S. K. Shukla, S. Urooj, A. Lay-Ekuakille, Ensemble of weighted deep concatenated features for the skin disease classification model using modified long short term memory, *Biomed. Signal Process. Control*, **76** (2022), 103729. <https://doi.org/10.1016/j.bspc.2022.103729>
22. L. F. D. J. Silva, O. A. C. Cortes, J. O. B. Diniz, A novel ensemble CNN model for COVID-19 classification in computerized tomography scans, *Results Control Optim.*, **11** (2023), 100215. <https://doi.org/10.1016/j.rico.2023.100215>
23. N. F. Aurna, M. A. Yousuf, K. A. Taher, A. K. M. Azad, M. A. Moni, A classification of MRI brain tumor based on two stage feature level ensemble of deep CNN models, *Comput. Biol. Med.*, **146** (2022), 105539. <https://doi.org/10.1016/j.compbimed.2022.105539>
24. F. Su, Y. Cheng, L. Chang, L. Wang, G. Huang, P. Yuan, et al., Annotation-free glioma grading from pathological images using ensemble deep learning, *Heliyon*, **9** (2023), 14654. <https://doi.org/10.1016/j.heliyon.2023.e14654>
25. K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, et al., The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository, *J. Digit. Imaging*, **26** (2013), 1045–1057. <https://doi.org/10.1007/s10278-013-9622-7>
26. S. Mourya, S. Kant, P. Kumar, A. Gupta, R. Gupta, C-NMC 2019 | C_NMC_2019 Dataset: ALL Challenge dataset of ISBI 2019, 2019. Available from: <https://doi.org/10.7937/tcia.2019.dc64i46r>
27. J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, preprint, arXiv:1412.6806.
28. M. Lin, Q. Chen, S. Yan, Network in network, preprint, arXiv:1312.4400.
29. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 2011–2023. <https://doi.org/10.1109/TPAMI.2019.2913372>
30. W. Zhou, H. Wang, Z. Wan, Ore image classification based on improved CNN, *Comput. Electr. Eng.*, **99** (2022), 107819. <https://doi.org/10.1016/j.compeleceng.2022.107819>
31. M. M. Khan, M. S. Uddin, M. Z. Parvez, L. Nahar, A squeeze and excitation ResNeXt-based deep learning model for Bangla handwritten compound character recognition, *J. King Saud Univ.-Comput. Inf. Sci.*, **34** (2022), 3356–3364. <https://doi.org/10.1016/j.jksuci.2021.01.021>
32. D. Jin, J. Xu, K. Zhao, F. Hu, Z. Yang, B. Liu, et al., Attention-based 3D convolutional network for Alzheimer’s disease diagnosis and biomarkers exploration state key laboratory of management and control for complex systems, institute of automation, in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (2019), 1047–1051. <https://doi.org/10.1109/ISBI.2019.8759455>

33. X. Li, H. Zhao, T. Ren, Y. Tian, A. Yan, W. Li, Inverted papilloma and nasal polyp classification using a deep convolutional network integrated with an attention mechanism, *Comput. Biol. Med.*, **149** (2022), 105976. <https://doi.org/10.1016/j.combiomed.2022.105976>
34. H. Xu, Y. Liu, L. Wang, X. Zeng, Y. Xu, Z. Wang, Role of hippocampal subfields in neurodegenerative disease progression analyzed with a multi-scale attention-based network, *NeuroImage Clin.*, **38** (2023), 103370. <https://doi.org/10.1016/j.nicl.2023.103370>
35. E. Brochu, V. M. Cora, N. deFreitas, A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, preprint, arXiv:1012.2599.



AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)