Ensemble Model of Visual Transformer and CNN Helps BA Diagnosis for Doctors in Underdeveloped Areas

Zhenghao Wei[1 [0000−0003−3316−3589]]

School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China
weizhh8@mail2.sysu.edu.cn

Abstract. The diagnosis of Biliary Atresia (BA) is still complicated and high resource consumed. Though sonographic gallbladder images can be used as an initial detection tool, lack of experienced experts limits BA infants to be treated timely, resulting in liver transplantation or even death. We developed a diagnosis tool by ViT-CNN ensemble model to help doctors in underdeveloped area to diagnose BA. It performs better than human expert (with 88.1% accuracy versus 87.4%, 0.921 AUC versus 0.837), and still has an acceptable performance on severely noised images photographed by smartphone, providing doctors in clinical facilities with outdated Ultrasound instruments a simple and feasible solution to diagnose BA with our online tool.

Keywords: biliary atresia · visual transformer · medical image processing · ensemble model.

1 Introduction

Biliary atresia (BA) is a pediatric disease affects both intrahepatic and extrahepatic bile ducts, which leads to pathological jaundice and liver failure in early infancy [1, 2]. Though BA only has a prevalence rate of about 1 in 5000–19,000 infants all over the world [3, 4], it is the most common cause for liver transplantation in infants below 1 year old [5]. If not treated timely, the disease will progress to end-stage liver cirrhosis rapidly, and liver transplantation will be necessary. Receiving Kasai portoenterostomy (KPE) surgery before age 2 months can largely extend the infant's native liver survival time [6]. Therefore diagnosis in an early stage is essential. However, it is still hard to distinguish BA from common causes of cholestasis [7], and diagnostic methods of BA including assay of serum matrix metalloproteinase-7 and screening the direct bilirubin concentration [7, 8], are not feasible for medical facilities with underdeveloped conditions especially in rural regions.

In the rural area of developing Asian country like China and India, ultrasound examination is the main approach to diagnose BA in jaundiced infants. Though the method has a specificity over 90% [9], the lack of experienced doctor who are

capable to make diagnosis and perform operation in undeveloped region, delays the appropriate treatment and the patient may miss the optimum therapeutic time. Thus, an AI-assisted diagnosis mechanism is vital to enable clinical staffs in undeveloped region to make preliminary diagnosis timely before the condition worsen.

Deep learning methods, specifically, the convolutional neural networks (CNNs), have been widely used in Biomedical Image Analysis tasks, and have been proved to be superior or comparable to human experts in tasks like diagnosis of lung cancer or Covid-19 [10, 11]. In 2021, teams from Sun Yat-Sen University (SYSU) and The First Affiliated Hospital of SYSU developed a deep learning Model based on Deep Residual Networks, to help diagnose BA, as well as a smartphone application which can help doctors especially in rural area, to diagnose rare disease like BA [12].

Though the model outperforms human experts, both in accuracy and sensitivity, the worst case is still unconsidered. In the case when sonographic system is not connected to Internet and exporting images is impermissible, taking a photo by smartphone may be the simplest solution. Moiré patterns and other noise can catastrophically reduce the accuracy of the classification model, and CNN is proved to be effective in such denoising tasks as Moiré photo restoration [13, 14]. Visual Transformer (ViT) is another state of the art (SOTA) deep learning model for image processing, which holds better performance in multitask than CNN [15], and is also utilized in Biomedical Image Processing (BIP) tasks, like MRI Segmentation [16] and Covid-19 diagnosis [17]. We combined the two techniques and utilize both the high accuracy of ViT models and the denoising ability of CNNs.

In summary, it is of great necessity to develop a diagnosis system of BA to assist unexperienced clinical staff in rural area to make preliminary diagnosis, and the system should have the ability to eliminate noises like Moiré patterns. In this study, we developed an ensemble deep learning model (EDLM) which contains four ViT models and two CNN models. It outperforms human experts in normal cases and have a better performance in noised dataset than ensemble CNN models or ViT models. Our main contributions can be concluded as follow:

1. We proposed an EDLM of both ViT and CNN to diagnose BA for infants aged below 60 days. The model is an image classification model which classifies sonographic gallbladder images, and it contains 4 ViT models (ViT-base, Swin Transformer-base, CvT-24 and CSwin Transformer-base) and 2 CNN models (ResNet-152 and SENet). Our model has a better accuracy and AUC score than human expert (with 88.1% accuracy versus 87.4%, 0.921 AUC versus 0.837).

2. Due to the imbalanced data (as a rare disease, BA-positive sample is lesser), we compared several strategies to enhance the specificity and sensitivity of the model.

3. We discussed several designs of the EDLM, compared the ViT-CNN ensemble model with ensemble CNN model, ensemble ViT model and single models, and two kinds of voting strategy. And we proved our mechanism performs better.

4. In the noised case, our ensemble model evidently outperforms others. And we adopted several few-shot tuning methods to enhance the model's ability to classify noised sonographic images photographed by cellphone.

2 Related works

2.1 Conventional BA diagnosis methods

BA is such a rare and complicated disease, as well as the infantile patients can't afford surgical exploration, an accurate diagnosis of BA is still a challenge. There are several clinical symptoms of BA including persistent jaundice, pale stools, and dark urine [3], for the conjugated hyperbilirubinemia due to cholestasis. However, neonatal cholestasis can be caused by many other diseases. Though screen the direct bilirubin concentration [9, 18] or stool color [4] can yield sensitivities over 97%, and serum gamma-glutamyl transferase (GGT) can be considered to distinguish BA from PFIC, bile acid synthesis or metabolism disorders, but mechanical bile duct obstruction, paucity of interlobular bile ducts or cystic fibrosis are still possible [19]. The specificity of method beyond is not satisfying (that means other disease still can't be excluded).

Those BA-suspected infants need further examinations including ultrasonography (US), hepatobiliary scintigraphy and magnetic resonance cholangiography (MRCP). Hepatobiliary scintigraphy has a low specificity as 70.4% and is time-consuming and radiant [20]. MRCP is also limited by the small body size of infants and is also deprecated for the need of sedation [21]. Therefore, US has been recommended as the preferred imaging tool for the initial detection of BA [3]. Several direct features reflecting the abnormalities of biliary system can be the criterion of BA, including Gallbladder abnormalities [20, 22], Triangular Cord sign [23, 24], and porta hepatic cyst [25, 26], but any single feature can't guarantee extremely high specificity and sensitivity at the same time. Even some infants may still have equivocal US results, that US-guided percutaneous cholecystocholangiography (PCC) may be needed [27]. In conclusion, the diagnosis of BA is a process full of complexity and uncertainty, that's why human experts mentioned in Zhou, W *et al.*[12] have an AUC lower than 0.85.

2.2 CNN models

The convolutional neural networks (CNN) has been the mainstream of Computer Vision (CV) domain since the invention of AlexNet [29], though come up decades ago [30]. In the last decade, deeper and more effective CNN models have been proposed and achieved enormous success, like VGG [31], GoogleNet [32], ResNet [33], DenseNet [34], PNASNet [35], and EfficientNet [36], etc.

ResNet and its variants are still the solid backbone architectures of CV, which introduces the Residual mechanism. Se-ResNet or SENet is an evolved model with a new mechanism called Squeeze-and-Excitation (SE) [37], allows the network to perform feature recalibration. That means it can use global information

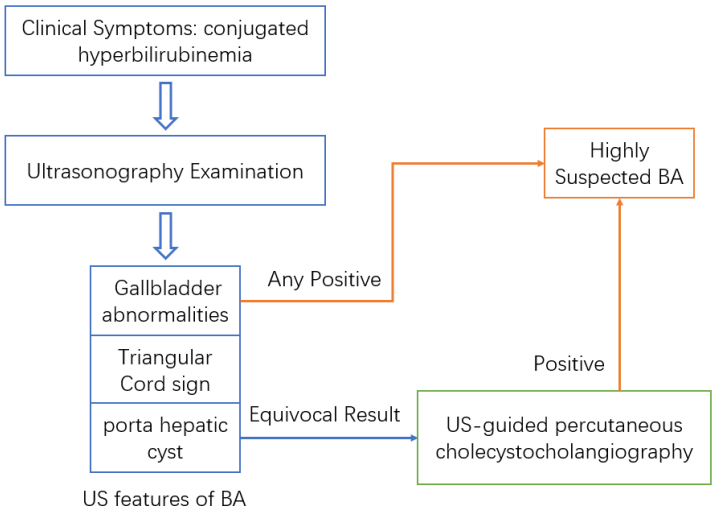


Fig. 1. A sketch map of the conventional BA diagnosis process. To learn a more detailed diagnostic decision flow chart can refer to[28]

to reemphasize valuable local features while restrains the less important ones. It can be likened to attention mechanism, which has a global receptive field on its input. Therefore, it performs better on classification tasks, and it is adopted by the BA-diagnosis application of Zhou, W *et al.*[12] as the main network.

2.3 Visual Transformer models

Visual Transformer (**ViT**) is firstly proposed by Google in 2021 [15], is the SOTA technique in image classification tasks. Since 2017, Transformer has been widely used in natural language processing (NLP) tasks [38], and pretrained models based on Transformer architecture like BERT [39], GPT [40] and UniLM [41] are still the SOTA techniques. Transformer consists of several Encoder blocks and decoder blocks, based on self-attention mechanism, which provide the capability to capture non-local features or encode dependencies between distant pixels [42]. It can be concluded as that CNN performs better on the extraction of local features and Transformer concentrates on non-local features. ViT imported the Transformer method into CV and proposed an approach to embedded patches of image into sequential input which is similar to inputs of NLP models. It outperforms all SOTA CNN models on all popular image classification benchmarks [15], but it doesn't have an equivalent performance on other downstream tasks like object detection and instance segmentation, due to its lack of inductive bias compared to CNNs.

DeiT is a follow-up model of ViT, which didn't change the network architecture but imported Knowledge Distillation method to transfer inductive bias from CNN teacher model to the student model [43].

Swin Transformer is one of the best performed CV backbone in multi-task [44]. It proposed an improved sliding window method (shifted window) to eliminate ViT's shortage caused by fixed scale. Swin Transformer utilized both the hierarchical design and the shifted window approach to transcend the former state of art models in several tasks, as well as reduce the computational complexity. Comparing to ViT, it holds less parameters and have a better speed-accuracy trade-off.

CvT is a combination of ViT and CNN, like Swin Transformer, it contains a hierarchy of Transformers to capture features of different receptive fields but utilizes convolutional layers to produce token embedding [45]. These changes introduce desirable properties of CNN, such as inductive bias.

CSwin Transformer is another evolved model of ViT proposed by Microsoft, the same as Swin Transformer [46]. Instead of the Shifted Window Attention (SWA) mechanism of Swin Transformer, it introduced a Cross-Shaped Window (CSwin) mechanism. SWA allows information exchange through nearby windows, but the receptive field is still enlarged quite slowly, therefore Swin Transformer needs a dozen of Transformer blocks in the pyramid structure. CSwin has a lower computational cost while can achieve the global receptive filed in less steps. CSwin Transformer is still the CV backbone models with highest performance.

3 Methods and Data Materials

3.1 EDLM of CNNs

Due to the Condorcet's Jury Theorem [47], ensemble classification model made up of individual models which have probabilities of being correct greater than 1/2, has a probability of being correct higher than the individual probabilities [48]. Hybrid or ensemble machine learning models have been popular since decades and can trace their history back to random forest [49] and bootstrap aggregating [50].

Our first attempt is to train k different models on k-fold split trainsets. That means we randomly separate the internal dataset into k (*e.g.*, five) complementary subsets, then in the k cases, different single subset is selected as the internal validation set, and other $k-1$ subsets will be combined as the training set. On each training set, a CNN model is trained, and for different models, their performances are evaluated on different validation sets (the remained one subset). It is similar to k-fold cross validation, but every model is kept (in k-fold cross validation, only the best model will be selected, with others abandoned). The ensemble model predicts the label of test data by calculating the average of outputs (predicted score of each category) of k models, and then do SoftMax to identify the final prediction label.

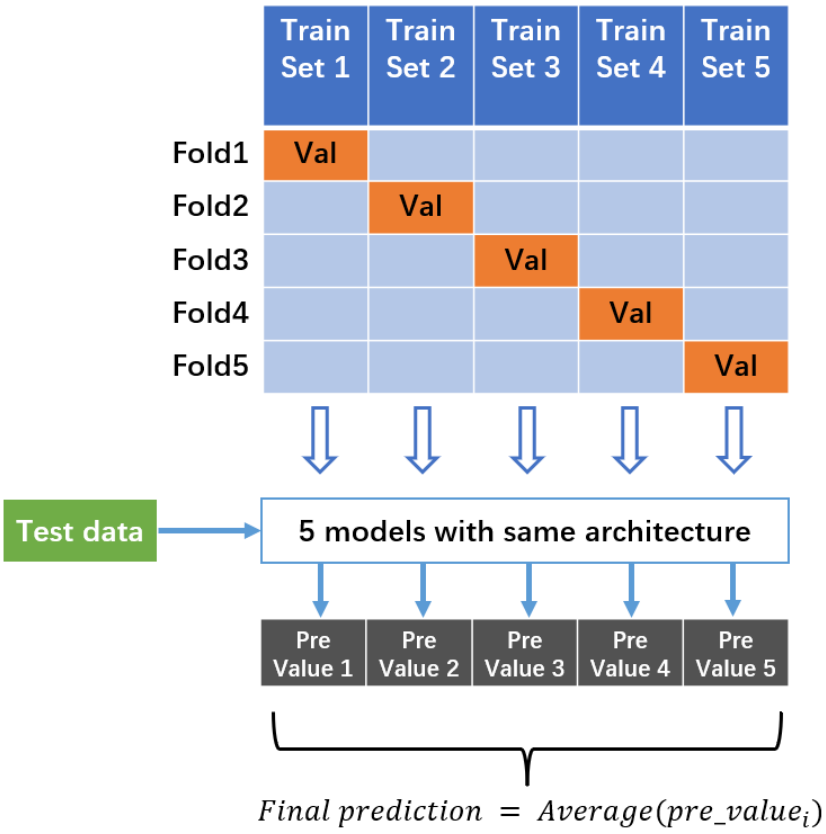


Fig. 2. The framework of an ensemble model. The form upside shows that the training set was divided into complementary subsets (means they are disjoint and the union of them are the universal set), only one of the subsets is selected as validation set in the training of each model. Differ to the classic case of Condorcet's Jury, neither simple majority rule nor unanimity rule is adopted, but we calculate the average of SoftMax outputs of the models (a real value instead of 0-1 value).

To construct the ensemble model, several CNN architectures were considered. Including ResNet [33], DenseNet [34], SeNet [37] and EfficientNet [36]. Zhou, W *et al.*[12] utilized SeNet-154 as the base classification of the EDLM, but we found it not a best choice, due to that the sonographic images are greyscale images de facto, and the motivation of SeNet is to enhance ResNet from the aspect of channel relationship. So, the SE mechanism may not function, and we can see, as a result, SeNet-154 didn't perform better than ResNet-152 (83.57% versus 85.60% on accuracy). Consequently, we adopted EfficientNetB6 as the base CNN model

3.2 ViT-CNN EDLM

Though ensemble model based on single model can promote the performance indeed, those models with same architecture have extremely high correlation, which conflicts with the assumption of independence of the Condorcet's Jury Theorem [47]. That means positive correlation between base models will decrease the accuracy of the ensemble model, and diversity of models will conversely improve the accuracy [51]. Therefore, choosing models with entirely different architectures may perform better than single architecture ensemble model.

As is concluded in Sec. 2.3, Visual Transformer has been becoming the new backbone architecture of computer vision, and it has a very different mechanism from CNNs. Convolutional mechanism endows CNNs the characteristics of translation invariance and inductive bias, while Transformer gives ViT models stronger ability of extracting global features in a patch, and they gain inductive bias by other approaches. So that if we combine CNNs with ViT models, they may concentrate on very different patterns and operates in different ways, and the diversity of the ensemble model will be guaranteed.

Other than EDLM in Sec. 3.1, since base models vary in architecture, it has no necessity to ensure models with different architectures trained on different dataset to produce diversity. Thus, models of same architecture will share a group of k-fold generated sets, while training sets of models with different architectures have no relevancy. The prediction rule is still the average strategy like in Sec. 3.1.

Several Visual Transformer methods were considered in our attempts, including ViT (the origin google version instead of the abbreviation of Vision Transformer), DeiT, Swin Transformer, CvT, and CSwin Transformer. As is shown below, single CSwin Transformer performed best in all these models.

3.3 Datasets

It should be mentioned that, besides the lack of experienced doctors, the sonographic machines are usually not connected to the internet, and export images is not feasible. Therefore, taking a photo by cellphone and upload to remote diagnosis system might be a temporary expedient.

Thus, we adopted two datasets, besides the sonographic gallbladder images provided on https://zenodo.org/record/4445734, we generated more noised pictures, which are photographed by several doctors with different models of smartphones. Differ from experiments in Zhou, W *et al.*[12], we want to explore the generalization ability of the model more, if we only take one type into account, there might be some bias unconsidered.

The first dataset consists of a 3705 images internal training dataset (internal validation set is also split from it), and an 842 images external validation dataset, all the images are segmented. The second dataset has 24150 pictures for the internal dataset, and 840 for the external validation one. 3659 pics of the internal set are original sonographic gallbladder images, and the others are reproduced picture took by smartphone, based on foresaid original images. Images in the internal set were took by five doctors, and there were two new coming

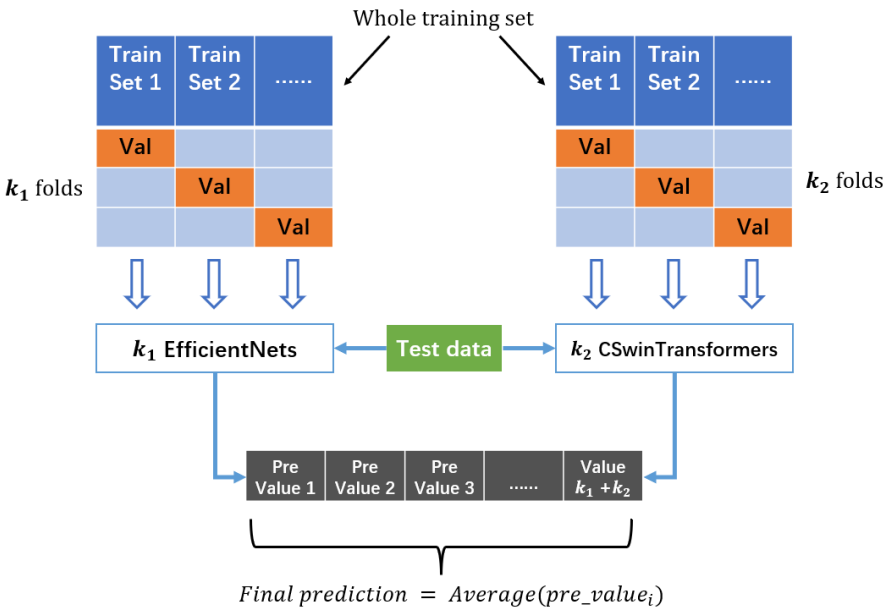


Fig. 3. The framework of a ViT-CNN EDLM (*i.e.*, two architectures, namely CSwin Transformer and EfficientNet). The difference is that, for each architecture, several models are trained on k-fold cross validation rule independently. For models of an architecture, they monopolize a whole training set, because there are enough diversity from different architectures, and models with same architecture are required to gain patterns.

doctors producing the external data. that is because we expect the model to gain adequate generalization ability, and the immunity against random noise.

3.4 Data processing

What should be considered at first is, that the positive samples are much less than the negative ones, due to BA is such a rare disease, as is mentioned, has a low incidence about 1 in 5000–19,000 infants [3]. There are only about 23% of the samples are BA positive in both datasets, so some measures should be applied to solve the imbalance problem. We tried several approaches, like resampling, down sampling, modifying class weights in the loss function, or k fold cross validation. As a result, we notice that resampling is the most practical strategy. Because we have adopted an EDLM method, k fold cross validation is redundant; down sampling made a precision loss, deservedly; changing class weights might be an efficient strategy, but we found it performs worse than simply resampling. It may blame the ensemble model, for in the trainset of a single model, some key

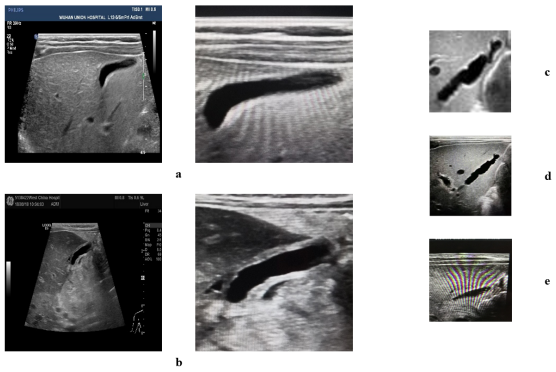


Fig. 4. **a**: an example of the second dataset, the left one is an original sonographic gallbladder image of a non-BA infant, and the right one is the reproduced picture of the left. **b**: BA patient, the same as **a**, original image on the left and reproduced on the left. **c**: a sample of the first dataset, differ from the second one, pictures are pre-segmented. **d**: smartphone reproduced picture in Zhou, W *et al.*[12]. **e**: smartphone reproduced picture of the second dataset. It's obvious that noise in **d** isn't severe, and the performance on **e** can better illustrate the generalization ability of the model.

samples of a pattern may miss, while in the resampling case, it is ensured that the trainset are more likely to contain all key samples.

Secondly, if we aim at improving the generalization ability of the model, even the 24k training dataset won't be sufficient, adopting some data augmentation techniques is necessary. Random rotation, random horizontal flip are adopted, resize and random crop is also applied to the original set, and all the images are turned into greyscale, for color has no significance in sonographic images.

And whether training data need masks, is also worth discussing. At first, the hospital provided manually annotated images, with a mask showed the focus, but finally we found the mask an obstacle to improve the model performance. From our perspective, the mask quality is unsatisfactory, but doctors who made the annotation are not to blame. As is mentioned in Sec. 2.3, multiple features are considered in the US diagnosis of BA, when the doctor label the images, he may only focus on one of those features, and omit some. Thus, ignoring the mask information and just train models on the original data could perform better.

4 Experiments and Results

4.1 Experimental settings

There are two tasks in the experiment, one is the diagnosing BA on original sonographic gallbladder images, the other one is on noised pictures photographed by smartphone. We solved the two problems step by step.

Firstly, we trained the classification model for original images. Models with different architectures were trained on different 5-fold split dataset, that means

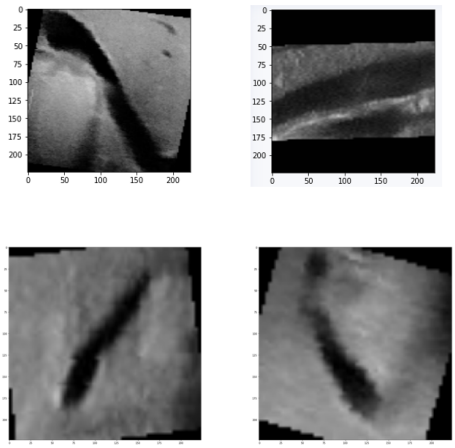


Fig. 5. Samples after data augmentation.

for one architecture, five models were trained at once, the training set was divided into five subsets as discussed in Sec. 3.1, and for each model, different subsets were used as validation set, remained four sets were used as the training cohort.

For every architecture, models are trained by transfer learning, the network loaded pretrained weight on ImageNet-1k classification dataset as the initial weight. And we fine-tuned the model on our training cohort. The loss function was Cross Entropy loss, and we tried weighted Cross Entropy as mentioned in Sec. 3.4, but we found it not better than resampling. The model was trained for 200 epochs and evaluated on internal validation set for every 5 epochs. If the evaluation loss didn't decrease for next 40 epochs, then the training progress would be forced to an early termination, and the parameter model with the lowest loss (*i.e.*, 40 epochs before) will be saved.

At the step of comparing different architectures, like k-fold cross validation, we chose the model with best performance on the test sets in every k-fold training as the representative. But after we chose the base architectures of the ensemble model, the five models will be adopted together to construct the ensemble model.

To control the number of models in different ensemble models consistent, all the ensemble models in the comparison consists of six base models. The ViT-CNN ensemble model has three SWin Transformer Base and three EfficientNet B6, trained on 3-fold cross training cohort.

For the second problem, we also utilized transfer learning, but it was the weight of ensemble model trained previously to be loaded. For each base model, we fine-tuned the model on the noised data, with augmentation mentioned in Sec. 3.4, for 30 epochs, evaluated every 5 epochs, and the best model in evaluation would be selected.

4.2 Results

As is shown below, for the single models, ViT family outperforms CNNs, and CSwin Transformer is the best single model. All the ensemble models have a better performance than human expert, but the ViT-CNN ensemble model do the best. CSwin Transformer has an 87.86% accuracy on the original images (while precision is 88.07% and recall is 87.74%, areas under the receiver operating characteristic curve of 90.78%), and 80.5% accuracy on the smartphone-took images (while precision is 80.49% and recall is 81.32%, areas under the receiver operating characteristic curve of 80.83%).

For ensemble models, ViT-CNN ensemble model has 88.11% accuracy on the original images (while precision is 88.35% and recall is 87.98%, areas under the receiver operating characteristic curve of 92.90%), and 81.11% accuracy on the smartphone-took images (while precision is 82.33% and recall is 81.71%, areas under the receiver operating characteristic curve of 81.04%).

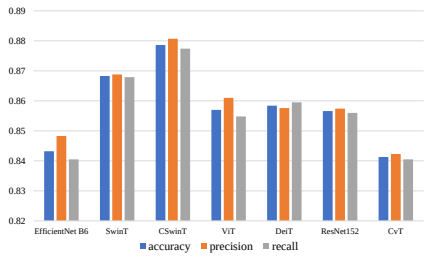


Fig. 6. Comparison between single models.

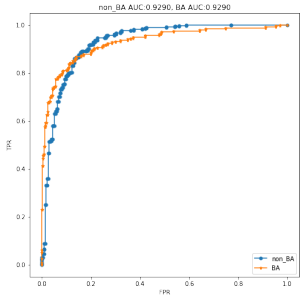


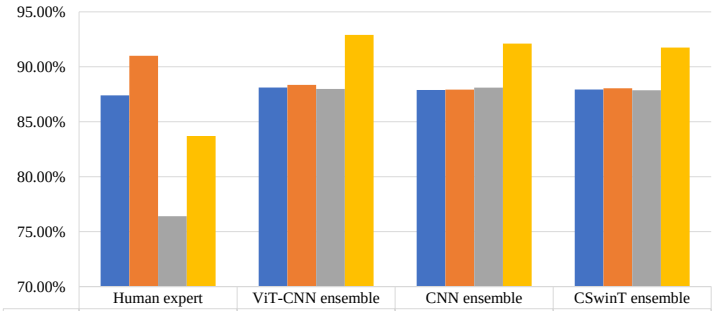
Fig. 7. The areas under the receiver operating characteristic curve (AUC) of ViT-CNN ensemble model.

Other than human experts, the deep learning models have precision and recall very similar, while the human experts have a recall obviously lower than the precision. The phenomenon indicates that our rebalance strategy works, that the deep learning models didn't performs worse on positive samples than negative ones. But human experts are still limited by the rarity of the disease, they can't recognize the atypical samples so that they have a lower sensitivity than machine.

5 Discussion

5.1 Comparison between methods

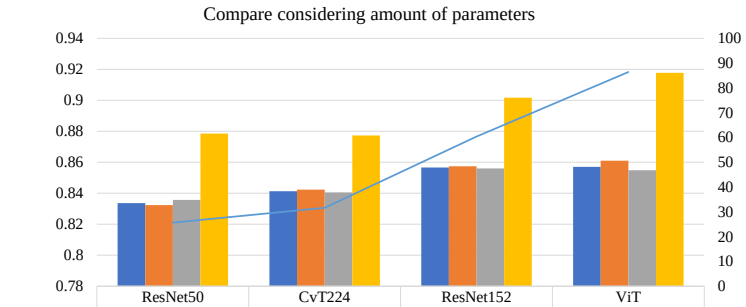
As has been discussed in Sec. 2.3, ViT models have been proved to be better backbone architectures than CNN models like ResNet or EfficientNet. In the



	Human expert	ViT-CNN ensemble	CNN ensemble	CSwinT ensemble
accuracy	87.40%	88.11%	87.89%	87.93%
precision	91.00%	88.35%	87.92%	88.04%
recall	76.40%	87.98%	88.10%	87.86%
auc	83.70%	92.90%	92.11%	91.75%

Fig. 8. Comparison between different ensemble models and human expert.

test of different architectures, generally, ViT models did better, and the order of performance is: CSwin Transformer, Swin Transformer, DeiT, ViT, CNNs. CvT is an exception that we didn't find a version of CvT has comparable amount of parameters to ViT-base (about 80M). So it is predictable that it has a worse performance than ResNet152 and other models



	ResNet50	CvT224	ResNet152	ViT
accuracy	0.8336	0.8413	0.8566	0.857
precision	0.8323	0.8423	0.8574	0.861
recall	0.8357	0.8405	0.856	0.8548
auc	87.85%	87.73%	90.17%	91.77%
Parameters	25.61	31.6	60.36	86.42

Fig. 9. Comparison considered difference in amount of parameters between models

However, it does not absolutely mean that ensemble ViT model will outperform ensemble CNN model. Though single CNN models have a worse performance than ViT models, as is shown in Fig. 8, ensemble CNN model has a comparable or better performance than ensemble ViT model. It can be inter-

preted as that in classification tasks, CNNs are more sensitive to training data. In each fold of training cohort, the performance of CNNs are limited by the lack of some images, but in the ensemble case, the diversity produced by different data, will improve the performance of ensemble model.

The ViT-CNN ensemble model has the best performance as we expected, with the same number of base models as other ensemble models. The diversity of model architectures may account for the better performance, it matches the assumption of independence of the Condorcet's Jury Theorem.

5.2 Ensemble strategy: simple majority or average

Though ensemble model has become a popular technique in improving the upper bound of model performance, few investigators take the difference between classic Condorcet's Jury and EDLM into account.

Assume an ensemble model is made up of several individual models whose prediction score of an input are independent and identically distributed (*i.e.*, suppose it is a random variable and we know the priori probability distribution), *e.g.*, a Gaussian distribution $N(\mu, \sigma^2)$, the difference of two kinds of voting strategy will be obvious. In the simple majority case, the probability of a single model misclassifies a positive sample (*i.e.*, $\mu > 0.5$, and prediction score is less than 0.5) is

$$\Pr\left(\textit{prediction score} < 0.5\right) = \int_0^{0.5} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu > 0.5. \tag{1}$$

For convenience, we record $p_1 = \Pr\left(\textit{prediction score} < 0.5\right)$, and the prediction score of a single model is S_i. Suppose the ensemble model has k base models, the voting game turns into a Bernoulli experiment $b(k, p_1)$. The probability of majority voting predicts incorrectly is the probability of the number of cases $S_i < 0.5$ is less than $k/2$ for $0 < i \leq k$ (k is odd), like this:

$$\begin{gathered}\Pr\left(\textit{simple majority voting is wrong}\right) = \\ p_1^k + C_k^1 p_1^{k-1}\left(1-p_1\right) + \ldots C_k^{\frac{k-1}{2}} p_1^{\frac{k+1}{2}}\left(1-p_1\right)^{\frac{k-1}{2}}\end{gathered} \tag{2}$$

And the average strategy is much simpler. The average output of the ensemble model can be represented as $\bar{X} = (X_1 + X_2 + \ldots + X_k)/k, X_i \sim N(\mu, \sigma^2)$, and all X_i are independent and identically distributed. Thus, $\bar{X} \sim N(\mu, \frac{\sigma^2}{k})$, that means more base models there are, more accurate the prediction is.

Let us quantitatively compare these two strategies. Assume the output of a base model correspond to $N\left(0.6, 0.1\right)$ and there are $k = 5$ base models, and $p_1 = 15.87\%$ can be easily calculated (having a 1σ deviation). Then

$\Pr\left(\textit{simple majority voting is wrong}\right) = 0.0311$, and the probability of the average case is wrong is the probability of having a $\sqrt{5}\sigma$ deviation,

$$\Pr\left(\textit{average score is wrong}\right) = \int_0^{0.5} \frac{\sqrt[4]{k}}{\sqrt{2\pi\sigma}} e^{-\frac{k(x-\mu)^2}{2\sigma^2}} = 0.0127 \tag{3}$$

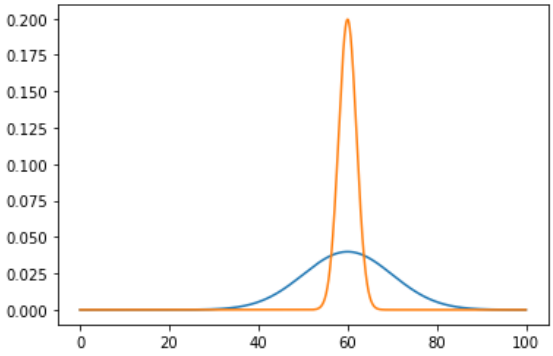


Fig. 10. PDF of X_i and $\bar{X}$.

It is obvious that the average strategy is less possible to mistake than the simple majority strategy.

References

1. J. Wang, Y. Xu, Z. Chen, J. Liang, Z. Lin, H. Liang, Y. Xu, Q. Wu, X. Guo, J. Nie, B. Lu, B. Huang, H. Xian, X. Wang, Q. Wu, J. Zeng, C. Chai, M. Zhang, Y. Lin, L. Zhang, S. Zhao, Y. Tong, L. Zeng, X. Gu, Z. gui Chen, S. Yi, T. Zhang, D. Delfouneso, Y. Zhang, S. L. Nutt, A. M. Lew, L. Lu, F. Bai, H. Xia, Z. Wen, and Y. Zhang, "Liver immune profiling reveals pathogenesis and therapeutics for biliary atresia," *Cell*, vol. 183, no. 7, pp. 1867–1883.e26, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092867420314550
2. A. Asai, A. Miethke, and J. Bezerra, "Pathogenesis of biliary atresia: Defining biology to understand clinical phenotypes," *Nature reviews. Gastroenterology & hepatology*, vol. 12, 05 2015.
3. J. L. Hartley, M. Davenport, and D. A. Kelly, "Biliary atresia," *The Lancet*, vol. 374, no. 9702, pp. 1704–1713, 2009.
4. C.-H. Hsiao, M.-H. Chang, H.-L. Chen, H.-C. Lee, T.-C. Wu, C.-C. Lin, Y.-J. Yang, A.-C. Chen, M.-M. Tiao, B.-H. Lau *et al.*, "Universal screening for biliary atresia using an infant stool color card in taiwan," *Hepatology*, vol. 47, no. 4, pp. 1233–1240, 2008.
5. R. Ohi, "Surgical treatment of biliary atresia in the liver transplantation era," pp. 1229–1232, 1998.
6. M.-O. Serinet, B. E. Wildhaber, P. Broue, A. Lachaux, J. Sarles, E. Jacquemin, F. Gauthier, and C. Chardot, "Impact of age at kasai operation on its results in late childhood and adolescence: a rational basis for biliary atresia screening," *Pediatrics*, vol. 123, no. 5, pp. 1280–1286, 2009.
7. C. Lertudomphonwanit, R. Mourya, L. Fei, Y. Zhang, S. Gutta, L. Yang, K. E. Bove, P. Shivakumar, and J. A. Bezerra, "Large-scale proteomics identifies mmp-7 as a sentinel of epithelial injury and of biliary atresia," *Science translational medicine*, vol. 9, no. 417, p. eaan8462, 2017.
8. S. Harpavat, J. A. Garcia-Prats, C. Anaya, M. L. Brandt, P. J. Lupo, M. J. Finegold, A. Obuobi, A. A. ElHennawy, W. S. Jarriel, and B. L. Shneider, "Diagnostic

yield of newborn screening for biliary atresia using direct or conjugated bilirubin measurements," *Jama*, vol. 323, no. 12, pp. 1141–1150, 2020.
9. T. M. Humphrey and M. D. Stringer, "Biliary atresia: Us diagnosis," *Radiology*, vol. 244, no. 3, pp. 845–851, 2007.
10. X. Qi, Z. Jiang, Q. Yu, C. Shao, H. Zhang, H. Yue, B. Ma, Y. Wang, C. Liu, X. Meng *et al.*, "Machine learning-based ct radiomics model for predicting hospital stay in patients with pneumonia associated with sars-cov-2 infection: A multicenter study," *MedRxiv*, 2020.
11. D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.
12. W. Zhou, Y. Yang, C. Yu, J. Liu, X. Duan, Z. Weng, D. Chen, Q. Liang, Q. Fang, J. Zhou *et al.*, "Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images," *Nature communications*, vol. 12, no. 1, pp. 1–14, 2021.
13. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.
14. Y. Sun, Y. Yu, and W. Wang, "Moiré photo restoration using multiresolution convolutional neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4160–4172, 2018.
15. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2021.
16. Y. Gu, Z. Piao, and S. J. Yoo, "Sthardnet: Swin transformer with hardnet for mri segmentation," *Applied Sciences*, vol. 12, no. 1, p. 468, 2022.
17. L. Zhang and Y. Wen, "A transformer-based framework for automatic covid19 diagnosis in chest cts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 513–518.
18. S. Harpavat, J. A. Garcia-Prats, and B. L. Shneider, "Newborn bilirubin screening for biliary atresia." *The New England journal of medicine*, vol. 375, no. 6, pp. 605–606, 2016.
19. A. G. Feldman and R. J. Sokol, "Recent developments in diagnostics and treatment of neonatal cholestasis," in *Seminars in pediatric surgery*, vol. 29, no. 4. Elsevier, 2020, p. 150945.
20. L. Zhou, Q. Shan, W. Tian, Z. Wang, J. Liang, and X. Xie, "Ultrasound for the diagnosis of biliary atresia: a meta-analysis," *American Journal of Roentgenology*, vol. 206, no. 5, pp. W73–W82, 2016.
21. M.-J. Kim, Y. N. Park, S. J. Han, C. S. Yoon, H. S. Yoo, E. H. Hwang, and K. S. Chung, "Biliary atresia in neonates and infants: triangular area of high signal intensity in the porta hepatis at t2-weighted mr cholangiography with us and histopathologic correlation," *Radiology*, vol. 215, no. 2, pp. 395–401, 2000.
22. P. Farrant, H. Meire, and G. Mieli-Vergani, "Ultrasound features of the gall bladder in infants presenting with conjugated hyperbilirubinaemia." *The British Journal of Radiology*, vol. 73, no. 875, pp. 1154–1158, 2000.
23. H.-J. Lee, S.-M. Lee, W.-H. Park, and S.-O. Choi, "Objective criteria of triangular cord sign in biliary atresia on us scans," *Radiology*, vol. 229, no. 2, pp. 395–400, 2003.

24. W.-H. Park, S.-O. Choi, H.-J. Lee, S.-P. Kim, S.-K. Zeon, and S.-L. Lee, "A new diagnostic approach to biliary atresia with emphasis on the ultrasonographic triangular cord sign: comparison of ultrasonography, hepatobiliary scintigraphy, and liver needle biopsy in the evaluation of infantile cholestasis," *Journal of Pediatric surgery*, vol. 32, no. 11, pp. 1555–1559, 1997.
25. M. Koob, D. Pariente, D. Habes, B. Ducot, C. Adamsbaum, and S. Franchi-Abella, "The porta hepatis microcyst: an additional sonographic sign for the diagnosis of biliary atresia," *European radiology*, vol. 27, no. 5, pp. 1812–1821, 2017.
26. E. Caponcelli, A. S. Knisely, and M. Davenport, "Cystic biliary atresia: an etiologic and prognostic subgroup," *Journal of pediatric surgery*, vol. 43, no. 9, pp. 1619–1624, 2008.
27. L.-y. Zhou, S.-l. Chen, H.-d. Chen, Y. Huang, Y.-x. Qiu, W. Zhong, and X.-y. Xie, "Percutaneous us-guided cholecystocholangiography with microbubbles for assessment of infants with us findings equivocal for biliary atresia and gallbladder longer than 1.5 cm: a pilot study," *Radiology*, vol. 286, no. 3, pp. 1033–1039, 2018.
28. W. Zhou and L. Zhou, "Ultrasound for the diagnosis of biliary atresia: From conventional ultrasound to artificial intelligence," *Diagnostics*, vol. 12, no. 1, p. 51, 2021.
29. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
30. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
31. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.
32. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
33. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
34. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
35. C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.
36. M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
37. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
38. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
39. J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

40. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
41. L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
42. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
43. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.
44. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
45. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.
46. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.
47. P. J. Boland, "Majority systems and the condorcet jury theorem," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 38, no. 3, pp. 181–189, 1989.
48. S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, "Advances in machine learning modeling reviewing hybrid and ensemble methods," in *International Conference on Global Research and Education*. Springer, 2019, pp. 215–227.
49. T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.
50. L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
51. S. Kaniovski and A. Zaigraev, "Optimal jury design for homogeneous juries with correlated votes," *Theory and decision*, vol. 71, no. 4, pp. 439–459, 2011.