Lung Cancer Diagnosis Based on Convolutional Neural Networks Ensemble Model

Lei Lyu*
Electrical Engineering Department
Columbia University
New York, NY, USA
*ll3433@columbia.edu

***Abstract*—Lung cancer is a lethal disease that can be treated efficiently if diagnosed in an early stage. Screening is a technology involving using CT scan to diagnose whether the lung is attacked by malignant tumors. This study proposes a CNN-based framework to help classify if the CT scan detects a cancer or not. In the analysis, several individual CNN models, including AlexNet, VGG, DCNN and DenseNet, are applied to make predictions and their performances are compared. Subsequently, selected individual models are ensembled by voting and stacking strategy that synthesize their predicting results. According to the results, the best individual model is DenseNet with average pooling layers, which gains a 97.48% accuracy and a 0.99019 AUC score. In comparison, the best ensemble model turns out to be assembling predicting results of best three individual models by stacked generalization, which reaches a 99.37% accuracy and a 0.99984 AUC score. These results show that it is useful to apply ensemble algorithm to improving the performance above individual models in this lung cancer diagnosis framework. Moreover, the final ensemble structure is efficient and reliable on figuring out lung scan images with malignant tumors.**

Keywords-CNNs; ensemble methods; medical image; image classification; lung cancer screening

I. INTRODUCTION

Lung cancer is one of the most harmful diseases to human beings on account of its high frequency and high mortality rate, with both coming at the leading positions among all cancer-related diseases [1]. According to American Cancer Society, lung cancer is the second most common cancer in both sexes, which ranks after prostate cancer and breast cancer respectively for men and women [2]. The total deaths caused by lung cancer each year even exceed the aggregation of colon, breast and prostate cancers [2]. As shown in the latest worldwide cancer statistics, cancer has led in the cause of death in most countries and with 36 kinds of cancers taken into consideration, new cases of lung cancer account for 11.4% (2,206,771 cases), which ranks only after female breast cancer (11.7%), but mortality cases of lung cancer surpass others by reaching 18% (1,796,144 cases) out of all with the second lethal cancer liver cancer occupying only 8.3% (830,180 cases) [3]. Patients are expected to live longer if their lung cancers are diagnosed in an early stage. In terms of the related statistics for 8th Tumor, Node, Metastasis (TNM) staging system, five-year survival rate is as high as 92% for very early-stage (T1a) lung cancer and drops to 38%-47% in T4 stage [4]. Furthermore, the age is also a factor that affects the five-year survival rate, with people under 50 are 83.7% likely to survive 5 years longer in stage 1, compared with only 54.6% of patients over 65 making it [5]. Therefore, it is necessary to find lung cancer as early as possible to let patients get treatments in an early stage, i.e., increases the possibility of their survivals from this lethal disease. Screening is a tool to find a disease in bodies before the symptoms occur, i.e., it is very useful to find lung cancer in the early stage which appears not so serious to affect people's bodies fiercely [6]. Recently, a low-dose CAT scan or CT scan (LDCT) is found in good practice for lung cancer screening to lower the possibility for high-risk people to die from lung cancer [6].

However, as LDCT technology for screening is used on people without showing symptoms, there will be far more cases than those who indeed have cancer. Actually only 2%-4% patients in USA get screened, since radiologists analyzing on lung scans fail to meet with the increasing demand [7-9]. As a result, it requires a more efficient and accurate method to differentiate whether the CT scan indicates a lung cancer or not for people taking screening. Artificial Intelligence (AI), machine learning and deep learning, all these terms are hot topics nowadays. Computer-aided methods have significant advantages in efficiency, and the main difficulties lie in the path to teach computer to accurately complete specific jobs like detecting lung cancer from screening CT scan. AI technology is the potential solution to train machines on this purpose [7]. Among all methods, deep learning, especially Convolutional Neural Network (CNN), is particularly adept at the task of image processing. Thus, plenty of big companies, research organizations and university teams are working on using deep learning models to detect potential lung cancer on the CT scan. Researchers in Google jointly with Northwestern University and other institutions designed a CNN region-of-interest model along with a full-volume model to train on LDCT volumes and the results of above models will be processed with another CNN cancer risk prediction model [9,10]. The performance of their model reached 94% accuracy on early-stage lung cancer identification and won over a panel of radiologists with many years' experience in terms of detecting cancer or predicting cancer two years later from a single CT volume and this model will hopefully be available on the Google Cloud Healthcare API after it has been tested more [8,9]. Recently in March 2021, a Virtual Nodule Clinic AI-powered clinical decision support software from the company Optellum received FDA 150(k) Clearance to become the world's first that kind of tool on lung cancer detection, which assigns LCP (Lung Cancer

Prediction) Score calculated from 3D pattern images from CT scans. Generally, this score aids to improve sensitivity and specificity of diagnosis largely and makes readers consistent whatever levels of expertise they are, which makes it useful in real practice on lung cancer diagnosing [11,12].

The main contribution of this research is to achieve an accurate and efficient prediction performance on lung cancer images. First, the AlexNet [13], VGG [14], DCNN [15] and DenseNet [16] models were utilized to make predictions. It turns out that DenseNet with average pooling layers performed best in prediction and reached 97.48% accuracy on testing data. Second, model ensemble technology was adapted through voting, including simple and weighted average soft voting, and stacking to synthesize the results of individual models. The best ensemble model out of all trials was a stacking model using a new classifying neural network to process aggregated scores of images assigned by three best individual models to make a final prediction. Specifically, it gained testing accuracy as high as 99.37%, with only one wrongly predicted testing image among all 159 testing cases.

The rest of the paper is organized as follows. In Section 2, the related work in CNN models, CNN models on image processing and model ensemble strategy research areas are introduced. Section 3 illustrates the data, individual models and model ensemble algorithms used in this study. Section 4 shows results of the individual model performance, ensemble model performance and sample testing images analysis. Eventually, the conclusion and discussion part are given in Section 5.

II. Literature Review

CNN models are well known for their excellent effects on various tasks, which are widely used in many fields contemporarily. Hoseinzade et al. developed a fine-tune CNN-based structure that is better initialized with a layer-wise supervised learning method [17]. Combined with strategy to extract general mechanism of stock market as features, the model can make predictions on directional movements of markets, which outperformed several other state-of-art baseline algorithms in the stock predicting area [17]. Liang et al. successfully embedded a CNN structure using JavaScript inside the browser for applications of NLP technique under a high latency or low connectivity circumstance, which is feasible to run in a promising speed and on any device that has installed a browser [18]. Liu and Qi proposed an innovative ResGCNN structure that integrates functions of BiGRU, TextCNN and residual network to achieve both long-term dependencies and key local features of text in text sentimental analysis jobs, and the performance of this model exceeded other nine text classification models on several English and Chinese text datasets [19]. These two researches promoted the development of CNN on NLP tasks in two different but effective directions to render CNN a common tool in NLP-related research and application. Suo et al. developed a time fusion CNN model from traditional CNN structure to achieve better patient similarity analysis and vectorized original data into similarity ranks, which can be processed with weighted sampling model to make personal predictions on patients' diseases with high and steady accuracy [20]. Besides processing text data, CNN models are adaptable to images which are typically 2D or 3D vectors represented by pixels. Hou et al. adapted a CNN structure to nonlinearly transform input images into targeted output images, including downscaling, decolorization and high dynamic range image tone mapping [21]. It was the first deep learning application on these three image processing jobs, and innovatively added another pre-trained CNN structure for loss function calculation based on output images [21].

However, CNN models are more famous for their contributions into the image classification filed. Since LeCun et al. proposed CNN structure in 1989 on zip code recognition to classify handwritten digits [22], CNN frameworks are under rapid development these years into deeper and more complex structures, involving AlexNet [13], VGG [14] and DenseNet [16], to better classify images into corresponding categories. Besides those well-known CNN models that can be generally used for many datasets, various CNN frameworks were also designed for specific types of image classification. Wei et al. proposed an infrastructure of CNN called Hypotheses-CNN-Pooling, which first arbitrarily segments image into different hypotheses, uses a shared CNN structure to generate an output and finally goes through a max pooling layer to make ultimate predictions [23]. Moreover, this structure gained a promising result on assigning multiple labels to one image [23]. Zhong et al. designed an agile CNN structure called SatCNN with smaller kernel sizes, which aimed at a specific dataset consisting of high-spatial resolution remote-sensing images and achieved most advanced performance on SAT datasets [24]. Classifying medical images is one of those specific tasks that need to be better investigated in. Jiao et al. designed a feature-based deep CNN structure on breast cancer detection that aimed at simulating the process of how doctors diagnose on the CT image having cancer or not by extracting deep features and achieved high predicting accuracy combined with the establishment of a decision mechanism to further process those features [15]. Chen et al. added an auto-encoder to CNN structure to enable learning features from partly labeled lung nodules images due to privacy limitations [25]. Li et al. designed a CLU-CNN framework based on ANCF and BN-IN Net to improve adaptation ability of CNN structures to different source and target domains, which is the problem that occurs in medical images but not an issue for normal images [26].

Researches above involve using only one model for specific purposes. However, as different models may focus on different features of data, another choice is to make trials on assembling a few models to generalize their abilities of processing specific datasets. Zhou summarized in his book several model ensemble algorithms used upon machine learning models that have gained success in many areas, including Boosting, Bagging, Averaging, Voting, Stacking etc. [27]. As neural networks become popular these years, a large number of scholars come to combine ensemble methods with deep learning models. Adhikari applied an artificial neural network structure to determining the weights of the linear combination of Box-Jenkins models, FANN models and SVM models on making time-series predictions and gained better performance compared with individual models as well as other combining schemes [28]. Kuang et al. appended a neural network structure on the end of his model to assemble features extracted by

previous three CNN subnets and achieved good performance on facial expression recognition problems [29]. More specifically, model ensemble techniques combined with CNN structures have been widely used in medical image diagnosis field. Kumar et al. came up with using several fine-tuned CNN models to extract features from medical images and applied 5 classifiers after CNN structures to ensemble their predictions, which made improvements on accuracy [30]. Osowski and Les first applied a pre-trained AlexNet model to extract features from images, next made use of softmax, SVM and RF classifier to generate 9 units to vote for final class of images and finally significantly improved the quality of the diagnosing system on Melanoma [31].

III. METHODOLOGY

A. Data

In this study, the data, called the IQ-OTH/NCCD lung cancer dataset, were extracted from Kaggle [32]. The Iraq-Oncology Teaching Hospital and National Center for Cancer Diseases are responsible to collect the data in 2019 with most cases coming from middle Iraq for analysis. The dataset contains 1097 jpg-format images mostly in the size of 512×512×3. Some images are contributed by the same patient. Original data are classified into three categories: malignant, benign and normal.

In data processing, the function in scipy.misc module was used to reshape the image to size of 256×256×3 and all pixels were normalized. Then, the benign and normal cases were merged into a new category, nonmalignant. Finally, 159 testing images were split out and remaining images were shuffled. The images left over were further divided into 800 training images and 138 validation images.

B. CNN Models

In this research, four types of CNN models were utilized, AlexNet, VGG, DCNN and DenseNet. First, AlexNet has the feature to extract large kernel sizes, 11 and 5, in the first two convolutional layers and then drop the size to 3 [13]. Second, VGG (D, 16 weight layers) sets kernel sizes to be 3 in all convolutional layers and has deeper layers compared with AlexNet [14]. Third, a deep feature-based CNN model is involved in this research, and I called it DCNN here [15]. Finally, this study used the DenseNet characterized by its unique structure Dense Block, which consists of Dense Layers that aggregate not only input of Dense Block but also outputs of all previous Dense Layers [16].

Since these CNN models are typically designed for images in size of 224×224×3 and in type of ordinary images instead of medical images (e.g., lung CT images here), their structures were modified slightly. For AlexNet, its large kernel size was maintained in the beginning, while I added a convolutional layer with kernel size of 3 after the first two convolutional layers respectively and modified pooling size to be 2. For VGG, the kernel size and stride size were changed in the first two convolutional layers to be 5 and 3 and the last 3 convolutional layers were cut off to make the structure shallower. For DCNN, its original structure was preserved. For DenseNet, it involved three Dense Blocks with each Dense Block containing 4 Dense Layers. Furthermore, DenseNet models with only average pooling layers or max pooling layers were used and they were called DenseNet (Avg) and DenseNet(Max) respectively. Based on the DenseNet(Max) model, the last pooling layer before fully-connected layers on the end was altered to be an average pooling layer and it was named DenseNet(Mix).

C. Model Ensemble

After implementing all those individual models, model ensemble technique was used to improve their performances. One of the most common ensemble methods is voting. More particularly, soft voting is used to combine outputs that are indicating the class probability and it can either simply average over all the probabilities of individual models or assign a weight to each probability output before averaging [27]. Another useful strategy to combine individual models is stacking, or stacked generalization. In the stacking procedure, the outputs of individual models are aggregated to be the new input for the classifier in the second stage and these two stages are typically trained separately [27]. In particular, it needs to receive confidence, not prediction, from models in the lower level to make stacked generalization more success. Besides, it is free to choose the types of the first-stage models [33].

In this research, individual models would generate outputs in forms of scores between 0 and 1. If the score is closer to 1, the image will be labeled as malignant, and vice versa. Therefore, the outputs can be treated as class confidence, or class probability, for the ensemble models. First, simple soft voting was applied, which simply took average of outputs of individual models chosen to calculate the new score. For this method, two models were implemented: Voting-6-average and Voting-3-average. For all ensemble models mentioned above and below, “6” indicates using all models and “3” means using only AlexNet, DenseNet(Avg) and DenseNet(Mix). Similarly, I tried to make use of weighted soft voting too, in which I assigned weights to each output according to the corresponding model predicting accuracy and took weighted average of individual models’ scores to gain the ultimate score. The same as above, I implemented two models, Voting-6-weight and Voting-3-weight. Finally, I thought about using stacking algorithm. In the first stage, all individual models were normally trained, instead of using n-fold technique to train models [33]. Then, all outputs were aggregated into one vector as the input of the next stage. In the second stage, I used another neural network with three fully connected layers activated by ReLU function (64, 256 and 1 units in each layer) plus two dropout layers using drop-out rate 0.5 and trained it with new inputs and original labels. Again, two stacking models, Stacking-6 and Stacking-3, were implemented.

IV. RESULTS

A. Individual Model

1) Model Performance Comparison

TABLE I. INDIVIDUAL MODEL PERFORMANCE INDEXES

Model	Testing Accuracy	Testing Loss	Training Time (s/epoch)	Precision	Recall	F_1	F_2	AUC
AlexNet	96.23%	0.0633	1.3	100%	92.59%	96.15%	93.98%	0.99937
VGG	93.08%	0.5536	2.4	97.30%	88.89%	92.90%	90.45%	0.94492
DCNN	91.19%	0.3829	2.5	98.55%	83.95%	90.67%	86.51%	0.94872
DenseNet(Avg)	97.48%	0.1813	3.8	97.53%	97.53%	97.53%	97.53%	0.99019
DenseNet(Max)	93.71%	1.1076	3.8	98.63%	88.89%	93.51%	90.68%	0.96882
DenseNet(Mix)	95.60%	0.1103	3.8	95.12%	96.29%	95.71%	96.06%	0.99224

Table 1 lists information on testing accuracy, testing loss, training time, precision, recall, F_1, F_2 and AUC performance of six individual models, including AlexNet, VGG, DCNN, DenseNet(Avg), DenseNet(Max) and DenseNet(Mix). As shown in Table 1, the top three models in terms of testing accuracy are DenseNet(Avg), AlexNet and DenseNet(Mix). For testing loss, AlexNet, DenseNet(Mix) and DenseNet(Avg) perform best out of all models. AlexNet is the fastest model to train, while all three types of DenseNet need the longest time to train in each epoch. All models gain over 95% performance on precision, and AlexNet, DenseNet(Max) and DCNN is three best models considering precision. Only half of models exceed 90% result on recall, and they are DenseNet(Avg), DenseNet(Mix) and AlexNet from highest to lowest. When it comes to consider the F_1 score, DenseNet(Avg), AlexNet and DenseNet(Mix) are the best three models. Based on the F_2 score, which pays more attention to the performance in recall, DenseNet(Avg), DenseNet(Mix) and AlexNet still outperform other three models. Furthermore, for the result of AUC, AlexNet is the best, DenseNet(Mix) is the second best and DenseNet(Avg) is the third best. In general, with all these evaluation indexes taken into consideration, the best 3 models should be AlexNet, DenseNet(Avg) and DenseNet(Mix), in no particular order.

2) Loss Curves Analysis

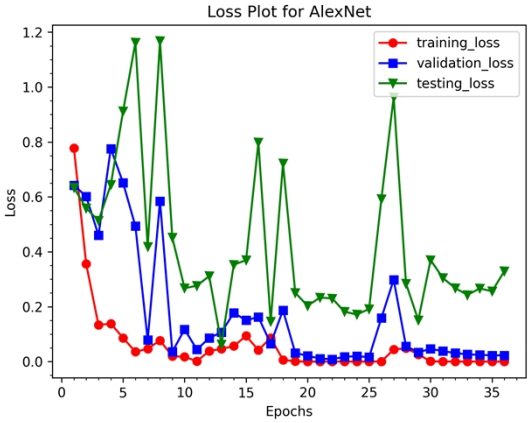

Figure 1. Loss plot of AlexNet Model

Training loss, validation loss and testing loss plot of AlexNet over 36 epochs of training is demonstrated in Figure 1. The training loss and validation loss tend to become very steady and remain in a low level. The testing loss will fluctuate but in an acceptable range. The value of testing loss is slightly larger than that of training loss, and the difference is roughly 0.3 in the end.

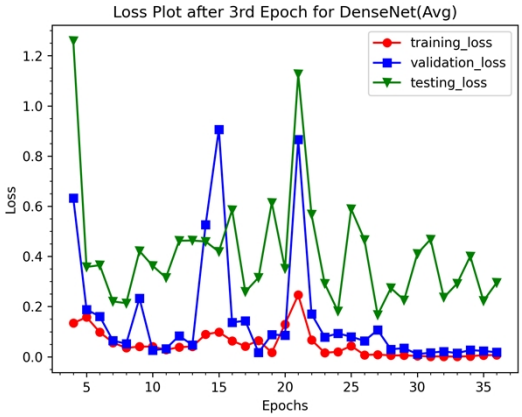

Figure 2. Loss plot of DenseNet(Avg) Model after the third epoch

Figure 2 illustrates how training loss, validation loss and testing loss curves of DenseNet(Avg) Model move over 36 epochs of training. The plot starts from the 4th epoch, as the loss values in the first three epochs are extremely high and it will make the loss curve after the third epoch look like a flat one. The training loss and validation loss tend to converge and remain in a low quantity. The testing loss fluctuates but still in a reasonable range. The value of testing loss is roughly 0.3 higher than that of training loss in the end of training.

B. Ensemble Model

1) Model Performance Comparison

The results of testing accuracy, testing loss, precision, recall, F_1, F_2 and AUC performance of ensemble models, including Voting-6-average, Voting-3-average, Voting-6-weight, Voting-3-weight, Stacking-6 and Stacking-3, are listed in Table 2. The Satcking-6 model makes use of all 6 individual models, while the Stacking-3 model only involves AlexNet, DenseNet(Avg) and DenseNet(Mix) models in stacking process. The Stacking-3 model does a better job in terms of all evaluation indexes, including testing accuracy, testing loss, precision, recall, F_1 score, F_2 score and AUC. The model gains testing accuracy of 99.37% and testing loss as low as 0.0414. The precision and recall calculated for Stacking-3 model are 100% and 98.77% respectively. Moreover, the F_1 score and F_2 score are calculated as 99.38% and 99.01%. As for the AUC score, Stacking-3 model attains a result of 0.99984.

2) Confusion Matrix Comparison

TABLE II. ENSEMBLE MODEL PERFORMANCE INDEXES

Model	Testing Accuracy	Testing Loss	Precision	Recall	F_1	F_2	AUC
Voting-6-average	94.97%	0.3998	100%	90.12%	94.81%	91.94%	0.99620
Voting-3-average	98.11%	0.1183	100%	96.30%	98.11%	97.01%	0.99953
Voting-6-weight	94.97%	0.3958	100%	90.12%	94.81%	91.94%	0.99620
Voting-3-weight	98.11%	0.1186	100%	96.30%	98.11%	97.01%	0.99937
Stacking-6	96.86%	0.1839	96.34%	97.53%	96.93%	97.29%	0.99683
Stacking-3	99.37%	0.0414	100%	98.77%	99.38%	99.01%	0.99984

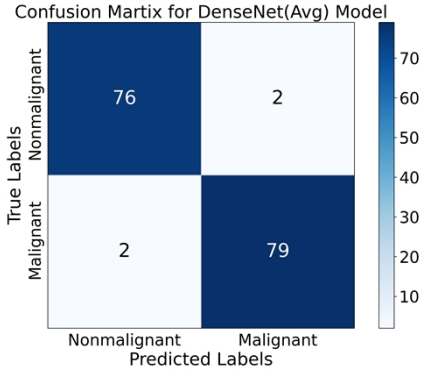

Figure 3. Confusion matrix of the best individual model DenseNet(Avg)

Figure 3 represents the confusion matrix of the best individual model in terms of testing accuracy, which is DenseNet(Avg). According to the results, one can calculate accuracy as 0.975 (155/159), misclassification rate as 0.025 (4/159), true positive rate as 0.975 (79/81), true negative rate as 0.974 (76/78) and precision as 0.975 (79/81).

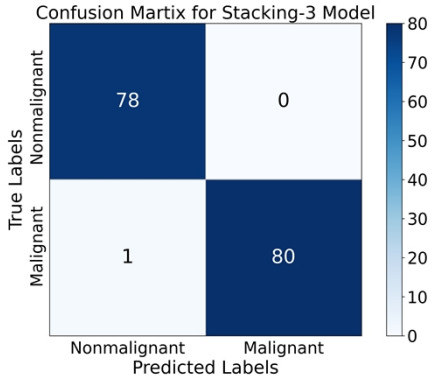

Figure 4. Confusion matrix of the best ensemble model Stacking-3

As illustrated in Figure 4, the confusion matrix of the best ensemble model considering the testing accuracy, Stacking-3, is shown. It can be calculated from this confusion matrix that for Stacking-3 model, accuracy reaches 0.994 (158/159), misclassification rate is only 0.006 (1/159), true positive rate reaches 0.988 (80/81), true negative rate is 1.000 (78/78) and precision reaches 1.000 (80/80).

Comparing two confusion matrixes and corresponding statistics results, Stacking-3 model, which gains the best prediction performance after assembling results of AlexNet, DenseNet(Avg) and DenseNet(Mix) using a neural network, outperforms DenseNet(Avg), the best individual model in terms of predicting accuracy on the testing dataset.

3) Loss Curves Analysis

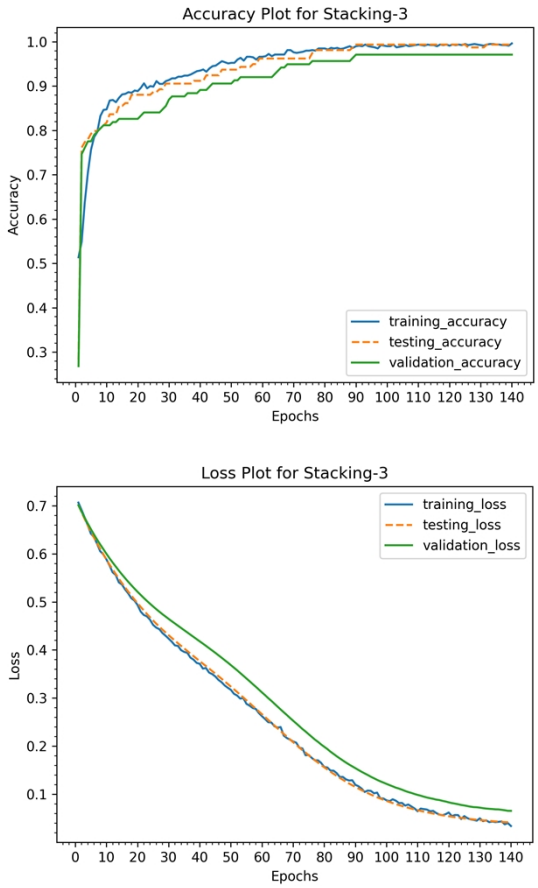

Figure 5. Accuracy plot of Stacking-3 Model in the neural-network classifier training stage (Above). Loss plot of Stacking-3 Model in the neural-network classifier training stage (Below)

The training process of the second-stage neural network used to ensemble the outputs of AlexNet, DenseNet(Avg) and DenseNet(Mix) is demonstrated in the plot with training accuracy, validation accuracy and testing accuracy and the plot with training loss, validation loss and testing loss over 140 epochs of training in Figure 5. In the accuracy plot, the testing accuracy and training accuracy finally tend to be flat and reach over 0.99 in the value. The validation accuracy also converges, but is slightly lower, which is roughly 0.97. For the loss plot, all the loss curves for training, validation and testing data indicate the tendency to converge to a low value. However, the values of training and testing loss are smaller than the quantity of validation loss in the whole training process.

(a) Malignant (b) Nonmalignant (c) Nonmalignant

Figure 6. Stacking-3 model predictions on test sample images (a), (b) and (c) [32]

TABLE III. SEVERAL MODEL SCORES ON TEST SAMPLE IMAGES (a), (b) AND (c) AND CORRESEPONDING CLASSIFICATION RESULTS

		AlexNet	DenseNet(Avg)	DenseNet(Mix)	Voting-3-weight	Stacking-3
Figure 6(a)	Score	0.1583	0.4784	0.6119	0.4160	0.5132
	Prediction	Nonmalignant	Nonmalignant	Malignant	Nonmalignant	Malignant
Figure 6(b)	Score	0.1628	0.5111	0.5140	0.3962	0.4506
	Prediction	Nonmalignant	Malignant	Malignant	Nonmalignant	Nonmalignant
Figure 6(c)	Score	0.0016	0.9996	0.1801	0.3968	0.1642
	Prediction	Nonmalignant	Malignant	Nonmalignant	Nonmalignant	Nonmalignant

C. Testing Samples Analysis

In Figure 6, three sample images (a), (b) and (c) from the testing set and their corresponding results predicted by the Stacking-3 model are displayed. In Table 3, it lists the scores that AlexNet, DenseNet(Avg), DenseNet(Mix), Voting-3-weight and Stacking-3 models assign to these three sample images. If the score is over 0.5, the image will be labeled as malignant case, or it will be classified into nonmalignant class.

For Figure 6(a), which is actually malignant, the Stacking-3 model predicts it as malignant. AlexNet scores it as 0.1583, DenseNet(Avg) scores it as 0.4784, DenseNet(Mix) scores it as 0.6119. For the ensemble model Voting-3-weight, it gains a score of 0.4160, which misclassifies it as nonmalignant. For the Stacking-3 model, it scores Figure 6(a) with 0.5132 and correctly labels it as malignant.

For Figure 6(b), it is actually a malignant case, but the Stacking-3 model misclassifies it as nonmalignant. AlexNet, DenseNet(Avg) and DenseNet(Mix) score this image with 0.1628, 0.5111 and 0.5140 correspondingly. The Voting-3-weight model assembles these scores to gain a new score as 0.3962, which fails to label Figure 6(b) correctly as malignant. For the Stacking-3 model, it generates the score as 0.4506, which also mislabels this image as nonmalignant, but the score is higher than that of the Voting-3-weight model.

When it comes to Figure 6(c), it is a nonmalignant image and the Stacking-3 model predicts it as nonmalignant. AlexNet scores it as 0.0016, which is very low, while DenseNet(Avg) assigns an extremely high score of 0.9996 to it. DenseNet(Mix) scores 0.1801 to this image. The Voting-3-weight model calculates the weighted average of three scores to attain the final score of 0.3968 and succeeds in predicting its label as nonmalignant. For the Stacking-3 model, it also correctly classifies Figure 6(c) as a nonmalignant case, although with a lower score of 0.1642.

V. CONCLUSION AND DISCUSSION

Cancer has become a serious disease and lung cancer is one of the most frequent and the most lethal one among all cancers. Screening is proven to be useful to reduce the mortality of lung cancer. Nevertheless, due to the limitation on the ability and number of radiologists, only a small portion of patients get screening. CNN models come as a suitable solution to this problem. In this research, I focused on using CNN models to efficiently and accurately predict lung cancer based on lung CT images. First, the dataset was accessed on the Kaggle and image preprocessing was conducted. Six individual models were adapted from AlexNet, VGG, DCNN and DenseNet with some modifications. Next, the performances of 6 individual models were compared in terms of several common evaluation indexes. With these models verified to be trained properly after analyzing on the loss plot, several schemes were proposed to ensemble models, including average (soft) voting, weighted (soft) voting and stacking. Afterward, those ensemble models were compared and the best ensemble strategy was the stacking model using AlexNet, DenseNet(Avg) and DenseNet(Mix), which reached a prediction accuracy of 99.37%. Furthermore, I compared the most predicting-accurate individual model with the best ensemble model using confusion matrix. Additionally, the accuracy plot and loss plot of the stacking model were analyzed to validate the proper training process in the second stage. Finally, three sample images in the testing set were chosen to demonstrate how models predict a malignant or nonmalignant case. In general, this study proposes a CNN-ensemble-model-based framework to aid with diagnosis of patients' CT scans with a high accuracy and efficiency, which makes finding lung cancer in an early stage more possible. By implementing a more generalized and powerful model, I believe that CNN-based ensemble models could outperform radiologists on diagnosing lung cancer from CT scans.

However, the current findings still have some limitations. First, in this study, the data size is not big enough and only

involves participants in one country, i.e., it will affect its generalization to more cases world-wide. Therefore, the next step to improve the research is to find more available data, transfer data into proper format and train the model on them. Second, the distribution of categories is not balanced in the original dataset, with only 120 out of 1097 images being benign cases. To address this problem, benign cases were integrated with normal cases to constitute nonmalignant cases, which are roughly the same in size with malignant cases. It is also possible to further improve the model by collecting more benign cases, i.e., one can try to classify those images into three categories. Finally, as for the stacking strategy utilized in the model, it adapted the simplest version of 1-fold training scheme to get a promising result, and it deserves trying n-fold training procedure if more data are available. Moreover, while some models behave better on predicting malignant cases and other models are more accurate on nonmalignant-class prediction, it also needs further research to understand the effect of similar or different patterns of each individual model's prediction performance on the final result of the stacking model.

REFERENCES

[1] Malhotra, J. , Malvezzi, M. , Negri, E. , La Vecchia, C. , & Boffetta, P. . (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, 889.

[2] American Cancer Society. (2021). Key statistics for lung cancer. *American Cancer Society*. Retrieved August 2, 2021, from https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html#written_by

[3] Sung, H. , Ferlay, J. , Siegel, R. L. , Laversanne, M. , & Bray, F. . (2021). Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer Journal for Clinicians, 71*(3).

[4] Kay, F. U. , Kandathil, A. , Batra, K. , Saboo, S. S. , Abbara, S. , & Rajiah, P. . (2017). Revisions to the tumor, node, metastasis staging of lung cancer (8th edition): rationale, radiologic findings and clinical implications. *World Journal of Radiology*(06), 269-279.

[5] Eldridge, L. . (2020). Stage 1 lung caner life expectancy. *Verywell Health*. Retrieved August 2, 2021, from https://www.verywellhealth.com/stage-1-lung-cancer-life-expectancy-2249418

[6] American Cancer Society. (2021). Can lung cancer be found early? *American Cancer Society*. Retrieved August 2, 2021, from https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/detection.html

[7] Kent, J. . (2019). Google develops deep learning tool to enhance lung cancer detection. *Health IT Analytics*. Retrieved August 2, 2021, from https://healthitanalytics.com/news/google-develops-deep-learning-tool-to-enhance-lung-cancer-detection

[8] Johnson, K. . (2019). Google's lung cancer detetion AI outperforms 6 human radiologists. *VentureBeat*. Retrieved August 2, 2021, from https://venturebeat.com/2019/05/20/googles-lung-cancer-detection-ai-outperforms-6-human-radiologists/

[9] Svoboda, E. . (2020). Artificial intelligence is improving the detection of lung cancer. *Nature, 587*(7834), S20-S22.

[10] Ardila, D. , Kiraly, A. P. , Bharadwaj, S. , Choi, B. , & Shetty, S. . (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine, 25*(5), 1.

[11] Sternberg A. . (2021). Diagnostic AI tool for identifying early-stage lung cancer earns FDA 150(k) Clearance. *Cancer Network*. Retrieved August 2, 2021, from https://www.cancernetwork.com/view/diagnostic-ai-tool-for-identifying-early-stage-lung-cancer-earns-fda-150-k-clearance

[12] Lingenbrink, L. . (2021). Optellum receives FDA clearance for the world's first AI-powered clinical decision support software for early lung cancer diagnosis. *Business Wire*. Retrieved August 2, 2021, from https://www.businesswire.com/news/home/20210323005236/en

[13] Krizhevsky, A. , Sutskever, I. , & Hinton, G. E. . (2012). Imagenet classification with deep convolutional neural networks. *NIPS*.

[14] Simonyan, K. , & Zisserman, A. . (2014). Very deep convolutional networks for large-scale image recognition. *Computer Science*.

[15] Jiao, Z. , Gao, X. , Wang, Y. , & Li, J. . (2016). A deep feature based framework for breast masses classification. *Neurocomputing, 197*(jul.12), 221-231.

[16] Huang, G. , Liu, Z. , Laurens, V. , & Weinberger, K. Q. . (2016). Densely Connected Convolutional Networks. *IEEE Computer Society*. IEEE Computer Society.

[17] Hoseinzade, E. , Haratizadeh, S. , & Khoeini, A. . (2019). U-cnnpred: a universal cnn-based predictor for stock markets. Papers.

[18] Liang, Y. , Tu, Z. , Huang, L. , & Lin, J. . (2018). CNNs for NLP in the Browser: Client-Side Deployment and Visualization Opportunities. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.

[19] Liu, C. , & Qi, J. . (2019). Text Sentiment Analysis Based on ResGCNN. *2019 Chinese Automation Congress (CAC)*. IEEE.

[20] Suo, Q. , Ma, F. , Ye, Y. , Huai, M. , & Jing, G. . (2017). Personalized disease prediction using a CNN-based similarity learning method. *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

[21] Hou, X. , Gong, Y. , Liu, B. , Sun, K. , Liu, J. , & Xu, B. , et al. (2018). Learning based image transformation using convolutional neural networks. *IEEE Access, PP*, 1-1.

[22] LeCun, Y. , Boser, B. , Denker, J. S. , Henderson, D. , Howard, R. E. , Hubbard, W. , et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, vol. 1, no. 4, pp. 541-551.

[23] Wei, Y. , Wei, X. , Min, L. , Huang, J. , Ni, B. , & Jian, D. , et al. (2016). Hcp: a flexible cnn framework for multi-label image classification. *IEEE Transactions on Software Engineering, 38*(9), 1901-1907.

[24] Zhong, Y. , Fei, F. , Liu, Y. , Zhao, B. , Jiao, H. , & Zhang, L. . (2017). Satcnn: satellite image dataset classification using agile convolutional neural networks. *Remote Sensing Letters, 8*(2), 136-145.

[25] Chen, M. , Shi, X. , Zhang, Y. , D Wu, & Guizani, M. . (2017). Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, 1-1.

[26] Li, Z. , Dong, M. , Wen, S. , Hu, X. , Zhou, P. , & Zeng, Z. . (2019). Clu-cnns: object detection for medical images. *Neurocomputing, 350*(JUL.20), 53-59.

[27] Zhou, Z. H. . (2012). *Ensemble Methods: Foundations and Algorithms*. Taylor & Francis.

[28] Adhikari, R. . (2015). A neural network based linear ensemble framework for time series forecasting. *Neurocomputing, 157*(jun.1), 231-242.

[29] Kuang, L. , Zhang, M. , & Pan, Z. . (2016). Facial Expression Recognition with CNN Ensemble. *International Conference on Cyberworlds*. IEEE Computer Society.

[30] Kumar, A. , Kim, J. , Lyndon, D. , Fulham, M. , & Feng, D. . (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical & Health Informatics, 21*(1), 31-40.

[31] Osowski, S. , & Les, T. . (2020). Deep Learning Ensemble for Melanoma Recognition. *2020 International Joint Conference on Neural Networks (IJCNN)*.

[32] Kareem, H. F. . (2020). The IQ-OTH/NCCD lung cancer dataset. *Kaggle*. Retrieved July 2, 2021, from https://www.kaggle.com/hamdallak/the-iqothnccd-lung-cancer-dataset

[33] Ting, K. M. , & Witten, I. H. . (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research, 10*.