Stacking Deep Learning for Early COVID-19 Vision Diagnosis

Ahmed A. Hammam, Haytham H. Elmousalami, and Aboul Ella Hassanien

Abstract early and accurate COVID-19 diagnosis prediction plays a crucial role for helping radiologists and health care workers to take reliable corrective actions for classify patients and detecting the COVID 19 confirmed cases. Prediction and classification accuracy are critical for COVID-19 diagnosis application. Current practices for COVID-19 images classification are mostly built upon convolutional neural network (CNNs) where CNN is a single algorithm. On the other hand, ensemble machine learning models produce higher accuracy than a single machine leaning. Therefore, this study conducts stacking deep learning methodology to produce the highest results of COVID-19 classification. The stacked ensemble deep learning model accuracy has produced 98.6% test accuracy. Accordingly, the stacked ensemble deep learning model produced superior performance than any single model. Accordingly, ensemble machine learning evolves as a future trend due to its high scalability, stability, and prediction accuracy.

Keywords Deep learning · Ensemble learning · Stacking · Classification · COVID-19 · Biomedical image processing

1 Introduction

The world health organization (WHO) declared that the COVID-19 outbreak is an international pandemic on 11th March 2020. On 24th April 2020, the number of confirmed cases has increased up to 3,042,444 cases and 211,216 deaths around more than 209 countries. The number of confirmed cases and deaths grow exponentially in our world. Therefore, all efforts should be integrated to fight COVID 19 pandemic. In the age of digital transformation, and machine learning (ML) play a key role in
processing the data to be converted into knowledge and decisions [1]. In the age of
digital transformation, big data and machine learning (ML) play a key significant in
processing the data to be converted into knowledge and decisions [2]. Several means
of COVID-19 diagnosis can be applied to identify the confirmed cases of COVID-
19. Radiologist’s diagnosis includes computed tomography (CT) scans; chest X-ray
(CXR) radiographs [3]. In any case, CT scans and X-ray pictures are time expending
and exhausted indeed for master radiologists. Computer vision and deep learning
computing such as convolutional neural networks (CNN) can viably offer assistance
radiologists for identifying COVID-19 affirmed cases.

Based on the chest CT scans, radiologists can detect the (COVID-19) pneumonia
and the arrange of persistent recuperation or weakening. Computerized insights
models can precisely produce early location for the conclusion of the patients of
COVID-19 by detecting the early lung thermal signs within the X-ray images. A
beginning inception neural model has been connected for two-fold classification for
tainted with COVID-19 or wellbeing people utilizing 1119 CT pictures [4]. Using
6000 CT, U-Net++ system can distinguish the COVID-19 patients with 93.55 and
100% for specificity and sensitivity [5]. Feature Pyramid Network have been used
for to distinguish the COVID-19 cases with an acceptable total accuracy of 86.7%.
Accordingly, the fully-connected layers produced better results with sensitivity of
0.93 and AUC of 0.99 [6, 7].

2 Deep Learning Models

A convolutional neural network (CNN) is a deep neural network that is conducted for
computer processing vision applications [8]. CNN architecture is commonly applied for biomed-
ical processing, analysis and classifications. CNNs are consisted of regularized multi-
layer perceptron and filters. The hidden layers of CNNs typically are represented of a
series of convolutional layers where these convolutional layers convolve based on dot
product to extract the features of each sample of images pixels. Each convolutional
layer has the following parameters: Convolutional kernels, the number of input chan-
nels and output channels, the depth of the convolution filter. CNNs includes pooling
layers to reduce the dimensions of the data by converting the neuron outputs to one
layer based on mathematical voting such as average. Multi-layer perception neural
network (MLP) can be used as a Fully connected to classify the images as displayed
in Fig. 1.

MobileNets proved its efficiency in several applications for embedded vision
applications and mobile applications. MobileNets applies streamlined architecture
based on depthwise convolutions for developing light weight model. MobileNets
optimizes the global parameters by trading off accuracy and latency [9]. The key
advantage of MobileNets is using limited hardware resources for computing by
reducing the network parameters and maintains the model accuracy.

MobileNets needs 1/33 of parameters needed for VGG-16 to achieve the same
classification accuracy [10]. Inception is a deep convolutional neural network (CNN)
architecture used for detection and classification applications. Inception architecture expands width and depth of the network with saving the computational cost. The Inception architecture optimizes quality of the architecture design using multi-scale processing and Hebbian principle [11]. Training deeper neural networks is a high computational process. Therefore, residual learning model has been developed to substantially train deeper models using residual functions related to the input layer as showed in Fig. 2. The key advantage of Residual nets is improving accuracy with depth of neural model and easier for optimization. Residual nets model has achieved a first place on the ILSVRC 2015 classification task [12].

3 Research Methodology

Model ensemble is a technique in which the predictions of a collection of models are given as inputs to a second-stage learning model. Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model.
3.1 Stacking Ensemble Deep Learning

The ensemble ML algorithms are depending on ensemble voting such as majority, plurality voting, “hard”, or “soft” voting. In hard voting, the final class label is predicted as the class label that has been predicted most frequently by the classification models. In soft voting, the class labels are predicted by averaging the class-probabilities. The soft voting is only recommended if the classifiers are well-calibrated [13, 14]. In majority voting (Hard Voting) can be formulated as the following formula:

\[ \hat{y}_i = \text{mode} \{C_1(X), C_1(X), \ldots, C_m(X)\} \]  

(1)

The weighted majority vote can be computed by associating a weight \( w_j \) with classifier \( C_j \) as the following formula:

\[ \hat{y}_i = \arg \max_m \sum_{j=1}^m W_j(C_j(X)) \]  

(2)

where (i) is the outcome of the classifier or class labels. In soft voting, the class labels are predicted based on the predicted probabilities \( p_j \) for classifier. This approach is only recommended if the classifiers are well-calibrated.

\[ \hat{y}_i = \arg \max_m \sum_{j=1}^m W_j p_j(X) \]  

(3)

Stacking is to ensemble several classifications algorithms such as Bagging or Boosting techniques. Stacking (Stacked Generalization) is applying different algorithms to learn part of the problem space and combining these different algorithms. Stacking paradigm improves the overall accuracy than any other individual based learner [15]. As shown in Fig. 3, implementation of stacking models consists of two main levels; Level-0 is training base learners (model A, model B and model C) where each model produces different classifications. Level-1(generalizer) is collecting the classification of each based learner to make final classification.

3.2 Model

The scope of the current research will focus on development stacking ensemble deep learning model for early COVID-19 diagnosis prediction using deep stacking model to boost the accuracy of single computer vision algorithms. As illustrated in the Fig. 5. The research methodology can be conducted through the Algorithm 1 and the following steps:
Collecting data set of X-ray or CT scans for COVID-19 confirmed cases.
Applying pre-processing techniques to remove missing data.
Applying different classification computer vision models.
Compare the results and performance of the applied model algorithms to rank using classification evaluation techniques.
Apply the stacking ensemble deep learning using the best performance models.
Compare the results of the stacking ensemble deep learning against the best single model using classification evaluation techniques.
Algorithm 1: Deep stacking Algorithm

OUTPUT $P_f$
INPUT $D_{train}, D_{test}, D_{Valid}, X_{Model\_Number}$
$I=0$

While ($I < X_{Model\_Number}$):
$I=I+1$
Train Model$_I$ ($D_{train}$)
Valid Model$_I$ ($D_{Valid}$)

MODELS = SELECT TOP 3 MODEL

$X_{stack} = \text{None}$

For Model in MODELS:
  # make prediction
  $P_M = \text{Model}. \text{predict} (D_{test})$
  if $X_{stack}$ is None:
    $X_{stack} = P_M$
  else:
    $X_{stack} = \text{data\_stack}(X_{stack}, P_M)$

$P_f = \text{ModelStack}.\text{fit}(X_{stack}, D_{test})$

4 Result

The data set consist of 500 X-ray images where the data set has been divided into three subsets: training set (80%), validation set (10%) and testing set (10%). The whole X-ray images have two labels: 0 for positive COVID-19 case and 1 for negative COVID-19 case as shown in Fig. 4.

To compare machine learning algorithms, the identical blind validating cases used to test the performance of the algorithm. The data set has been dividing into a training set (80%), validation set (10%) and testing (10%) where the validation cases are excluded from the training data to ensure the generalization capability. Classification accuracy (Acc), specificity and sensitivity are scaler measures for the classification performance. Moreover, receiver operating characteristics (ROC) is a
graphical measure for classification algorithm [16]. The receiver operating characteristics (ROC) curve is a two-dimensional graph in which the true positive rate (TPR) represents the y-axis and a false positive rate (FPR) is the x-axis. Classification accuracy (Acc) computes the ratio between the correctly classified instances to the total number of samples as the following equations:

\[
TPR = \frac{TP}{TP + FN} \tag{4}
\]

\[
FPR = \frac{FP}{TN + FP} \tag{5}
\]

\[
Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}
\]

where: true positive (TP); false positive (FP); true negative (TN); false negative (FN).

Based on ROC, the perfect classification happens when the classifier curve possesses through the upper left corner of the graph.

The study applies six different deep learning models for COVID-19 classification. These models are MobileNet, InceptionResNetV2, ResNet50, ResNet50V2, InceptionV3, and VGG16. Table 1 shows the setting for each model used for training process; the number of epochs was 100, learning rate was 0.001, and the optimization algorithm was Adam.

Figures 6 and 7 illustrates the training accuracy and validation accuracy during for the applied models. Moreover, the corresponding confusion matrix to each frame.
Table 1  Setting of algorithms training

| Parameter                      | Setting |
|-------------------------------|---------|
| Epoch number                  | 100     |
| Learning rate                 | 0.001   |
| Optimization algorithm        | Adam    |
| Training sample size (%)      | 80      |
| Testing sample size (%)       | 10      |
| Validation sample size (%)    | 10      |

Fig. 6  Accuracy Comparison For each model a Mobile Net, b ResNet50V2, c Inception ResNetV2, d ResNet50V2, e VGG16 and f InceptionV3

woke have been displayed. All models presented acceptable accuracies ranging from 0.971 to 0.943 based on Testing accuracy. As shown in Table 2 and Fig. 8, the training, validation and testing accuracy for each model have been displayed. The results show that the MobileNet, InceptionResNetV2 and ResNet50 have 0.971 for testing accuracy. ResNet50V2, and InceptionV3 have 0.957 for testing accuracy. VGG16 have 0.943 for testing accuracy. As a result, the first three accurate models were MobileNet, InceptionResNetV2 and ResNet50. The next stage is applying stacking ensemble deep learning model to collect the results of MobileNet, InceptionResNetV2 and ResNet50. The stacked ensemble deep learning model accuracy has produced 0.986 test accuracy. Accordingly, the stacked ensemble deep learning model produced
Fig. 7  Heat-map comparison for each model a MobileNet, b ResNet50V2, c InceptionResNetV2, d ResNet50V2, e VGG16 and f InceptionV3 where 0 = Positive and 1 = Negative
Table 2  Deep learning models results

| Model           | Training accuracy | Validation accuracy | Testing accuracy |
|-----------------|-------------------|---------------------|-----------------|
| MobileNet       | 1                 | 0.985               | 0.971           |
| InceptionResNetV2 | 1              | 1                   | 0.971           |
| ResNet50        | 1                 | 0.985               | 0.971           |
| ResNet50V2      | 1                 | 0.9857              | 0.957           |
| InceptionV3     | 1                 | 1                   | 0.957           |
| VGG16           | 0.977             | 0.971               | 0.943           |

Fig. 8  Deep learning models results

superior performance than any single model. The stacked model improves the accuracy of COVID-19 classification by 1.54% than other the highest accurate applied models.

5  Conclusion

We proposed stacked ensemble deep learning model by combining the predictions from multiple deep learning models on the same dataset, the models are typically different in architecture that are skilled on the dataset, but in different ways. Stacked is an ensemble method where model figures out how to best join the predictions from numerous current models. The stacked ensemble deep learning model accuracy has produced 0.986 test accuracy. The stacked ensemble deep learning model produced superior performance compared to any single model (top single model one is 0.971). The ensemble learning algorithms could be a future trend for prediction and classification applications where dataset have limited size. Moreover, we plan to
further explore more ensemble deep learning approaches to produce higher predictive accuracy than single computer vision algorithm for classification.

References

1. Elmousalami, H.H., Hassanien, A.E.: Day level forecasting for coronavirus disease (COVID-19) spread: analysis, modeling and recommendations, arXiv preprint arXiv:2003.07778 (2020)
2. Elmousalami, H.H.: Comparison of artificial intelligence techniques for project conceptual cost prediction: a case study and comparative analysis. IEEE Trans. Eng. Manage. 1–14 (2020)
3. Hassanien, A.E., Mahdy, L.N., Ezzat, K.A., Elmousalami, H.H., Ella, H.A.: Automatic X-ray COVID-19 lung image classification model based on multi-level thresholding and support vector machine. medRxiv (2020)
4. Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., et al.: A deep learning algorithm using CT images to screen for corona virus disease (COVID-19), medRxiv preprint medRxiv:2020.02.14.20023028 (2020a)
5. Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., Hu, S., Wang, Y., Hu, X., Zheng, B., et al.: Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. medRxiv preprint medRxiv:2020.02.25.20021568 (2020b)
6. Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Chen, Y., Su, J., Lang, G., et al.: Deep learning model to screen coronavirus disease 2019 pneumonia. arXiv preprint arXiv:2002.09334 (2020)
7. Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Zhao, H., Jie, Y., Wang, R., et al.: Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. medRxiv preprint medRxiv:2020.02.23.20026930 (2020)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Models, pp. 1097–1105 (2012)
9. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
10. Li, Y., Huang, H., Xie, Q., Yao, L., Chen, Q.: Research on a surface defect detection algorithm based on MobileNet-SSD. Appl. Sci. 8(9), 1678 (2018)
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … & Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Models, pp. 1–15 (2000)
14. Goyal, A., Sardana, N.: Empirical analysis of ensemble machine learning techniques for bug triaging. In: 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1–6 (2019)
15. Deng, L., Yu, D., Platt, J.: Scalable stacking and learning for building deep architectures. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2133–2136. IEEE (2012)
16. Tharwat, A.: Classification assessment methods. Appl. Comput. Inf. (2018)