Research on Data Science Ensembles for Covid-19 Detection and Length of Stay Prediction

2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) | 978-1-7281-8529-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCIS51004.2021.9397218

SHUBHAM SINHA
Department of CSE
HMRITM, GGSIP University
New Delhi, India
shubhamsinha1320@gmail.com

TUSHAR
Department of CSE
HMRITM, GGSIP University
New Delhi, India
tushardhiman1999@gmail.com

DR. SHALINI GOEL
Associate Professor
Department of CSE
HMRITM, GGSIP University
New Delhi, India
dr.sgoel1803@gmail.com

***Abstract*—We have demonstrated the use of an iteratively severed model of deep learning which associates for diagnosing Covid-19 pulmonary demonstration of using chest X-rays. In this paper, a customized convolutional neural network model is trained and analyzed on publicly available chest X-rays to grasp modality-strict feature demonstrations. Since the best performing models learn iteratively to make the model memory efficient, this model also learns and tries to improve the results with each step and classify the chest X-rays in their categories accurately. Then another model which predicts the length of stay of a patient at the hospital is created using multi-layered data processing approach. This model will empower hospitals for on time interference to prevent confusions and better management of hospital resources. We propose a method that uses catboost model which generally classifies the data in multiple classes. As a result, this study provides modality strict iterative and knowledge reusable model which influences Covid-19 detection and length of stay prediction.**

Keywords—Covid-19; Length of Stay; Catboost; Convolutional Neural Networks; Maxpooling; chest X-rays

I. INTRODUCTION

Novel Coronavirus 2019 (Covid-19) is a disease that pioneered in Hubei district in Wuhan city of China and has expanded globally. It is prompted by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1]. Covid-19 was announced pandemic by the World Health Organization (WHO) on March 11, 2020. Since its outbreak, it has been affecting the human population all over the world. It has affected millions of people and also has taken a lots of lives. There are some symptoms and diseases which indicate Covid-19 infection in a person such as fever, cough, muscle aches, difficulty in breathing, headache and so on [2]. Lung disease which is the primary cause for difficulty in breathing has been reported a symptom like hyperthermia for Covid-19. To detect Covid-19 infection, several tests are performed. One of those tests is reverse transcription- polymerase chain reaction (RT- PCR) which is contemplated superlative [3]. Though they have shown variable sensitivity and their availability is also in check in some geographical locations.

The computed tomography scans (CT scans) and chest X-rays (CXRs) have been exercised to diagnose Covid-19 and progression of the disease [4, 5]. However, they are not currently considered primary diagnostic tools. However, there are various challenges to the use of chest CT. Some of them are sanitization of room between patients which leads to delay in time, non-portability and risk of exposed doctors, hospital staff and other patients to the virus. Portable CXRs which are not as sensitive are reckoned feasible alternative since patients can be diagnosed in isolated rooms and also sanitization complexity is lesser than CT.

Automated artificial intelligence (AI) driven computer-aided diagnostic (CAD) tools designed to segregate and diagnose thoracic abnormalities related to Covid-19 [6]. These tools merge the components of computer vision with radiological image processing for analyzing disease demonstration and concentrating suspicious regions of interest (ROI) [7]. The neoteric advancements in deep learning and machine learning methods such as convolutional neural networks (CNN) have been promising in detection of patterns of disease in medical images [8].

Since Covid-19 is spread all over the world and patients who stay in a hospital are at higher risk of getting infected from the virus. The community spread of Covid-19 has led to major fatalities and all the governments are trying to control this pandemic by reducing community spread of this virus [9]. The healthcare system also combines many units of services according to the patient requirements. The time spent by patients in the hospital is known as length of stay (LOS) [10]. This measurement of duration is done in number of days. The prediction of length of stay can be helpful for the hospitals as they will be allowed to plan for preventive procedures, manage resources of hospital better and provide better health services. The management efficiency of hospital goes down significantly in a couple of situations: First, when the hospital has supply more than demand. Second, when the hospital is short on resources and facilities. With the help of better estimation of length of stay of patients, hospital can have better resource utilization and efficient bed management. The hospitals are always anticipated to do best with continuous dwindling resources. As healthcare remuneration trends for ‘pay for performance’, hospitals can have large financial loss if they do not predict and prevent long LOS.

In our approach, we first create a CNN model which takes the input of the preprocessed image of CXR and detects whether the patient is infected or not. If the patient is tested positive for Covid-19, then the patient must be admitted to the hospital for the treatment of Covid-19 for at least 2 weeks. Then we take the data of the patient using a form and then process it on a catboost model to predict the range of days for which the patient should be admitted [11]. We extracted some performance measures such as recall and accuracy. However the classification models may have different performances and it could be rather hard to identify the best.

The rest content of this paper is organized in the following manner: in section II we have discussed related work, in section III and IV, we have outlined our datasets and models, in section V we have discussed results and finally in section VI we have presented our conclusions and future scope.

II. PRIOR WORK

The section is mainly focused on the research work on Covid-19 detection using CNN and the Length of Stay (LOS).

A. Covid-19 Detection

Various AI efforts for Covid-19 study are evident by a study of the literature. A distinguishing proof about other viral pneumonia from Covid-19 viral demonstrated on chest Computed Tomography scans by high predominance [12]. It was inspected that Covid-19 viral was founded superficially dispensed with vascular thickening and GGO (ground glass opacities). A set of 275 Computed Tomography scans evidencing Covid-19 viral demonstrations and a trained deep Convolutional Neural Network model to get an F-score of 0.85 by categorizing CT scans in Covid-19 viral-related opacities or showing normal, is publicly available by the author of [9]. A pre-trained AlexNet model and customized Convolutional Neural Network model are used to categorize Chest X-rays as Covid-19 pneumonia or normal by an accuracy of 98% and 94.1% respectively by the author of [13]. A ResNet-50 Convolutional Neural Network is utilized to categorize Covid-19 viral demonstrations, pneumonia, and normal in Chest X-Rays and obtained an F-score of 98.19 and an accuracy of 98.18% by the author of [14]. Non-Covid-19, bacterial viral and others type of pneumonia can be frequently analyzed to distinguish and diagnose using Chest X-rays (CXRs) [5]. A customized CNN model which was formed by adding a machine driven designing with manual design prototyping approached to label Chest X-rays as Covid-19 or normal with an accuracy of 92.4% proposed by the author of [8].

B. Modality-strict Knowledge Transfer

To enhance performance, traditionary transfer learning techniques offer obligation where the presentation of learned features are fine-tuned with limited or we can say less amount of Covid-19 pneumonia CXR information [6]. Nevertheless, including lower intra-class variance and a higher inter-class similarity mannered in such a way that the appearance of medical photos lead to overfitting and model bias which results in reduced generalization and performance. By retraining Convolutional Neural Network models on a larger Chest X-ray photos set to understand modality-specific feature representation, the issues of overfitting and model bias can be reduced through modality-specific knowledge transfer [15].

C. Ensemble Classification

CNNs are anomalous models which learns difficult relationships from the information through stochastic optimization and error backpropagation [16], building them intensively sensitive to the statistical noise and random weight initializations existing in the training data [17]. Using ensemble learning these type of problems can be removed through training various models and concatenating the predictions where a specific model’s debilitation are balanced by predictions of variant models. The combinational predictions of specific models exhibit to be better to them.

D. Length of Stay (LOS)

The duration of time from which a patient takes admission in a hospital until the patient’s discharge is known as Length of Stay. Its biggest advantage is knowing how long a patient have to stay in a hospital. It has been studied broadly and some of the studies states that since 1960’s, prediction models are being build trying to address the problem of LOS.

For predicting Length of Stay Gustafson proposed five methodologies, in which three of them are distribution estimators which uses Bayesian theorem, that produces educts based on subjective judgements and empirical data and the other two of them produces point educts based on surgical residents and subjective judgement of surgeons [18]. From eight inguinal herniotomy patients the information is gathered, and solders demographic information and symptomatic information. The less number of samples and the data about the model predicts that the model performance is not effective, causing apprehension as to whether the outputs can be generalized. LOS and mortality prediction equations known as Acute Physiology and Chronic Health Evaluation III (APACHE III) was appraised by Woods el at [19]. Based on many symptomatic variables and some demographics like heart rate, body temperature, blood pressure, etc., a score ranging from 0 to 299 is given to the patients by the use of APACHE III which is a disease intensity or rapidity ranking classification.

III. DATA COLLECTION AND PREPROCESSING

We used dataset from kaggle and GitHub in which one is used for determining whether a patient is suffering from Covid-19 or not. In this dataset, there are around 500 images in training set and 100 images on testing set and these sets are furthered divided into two categories one is ‘Covid-19’ as shown in figure 1 (A) and other is ‘Normal’ as shown in figure 1 (B). Basically these datasets contain X-rays images of chest which is helpful in predicting the results using the model which is built.

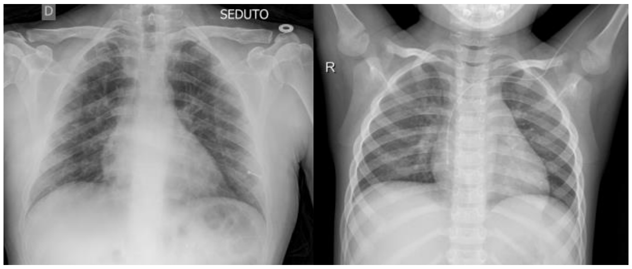

Fig. 1. (A) Covid-19 CXR (B) Normal CXR

And the other dataset is used to predict the length of stay of a patient in the hospital. In this dataset, the data of more than 3 hundred thousand patients is present for training and prediction of LOS. A brief sample of LOS dataset is reviewed in Table I.

For detection for Covid-19 the dataset is pre-processed in a manner that it will pass through some Convolutional layers followed by some Maxpooling which is followed by dropout layers which helps us to pre-process the image and layer by layer it gives a refined image containing only essential features which is mainly required to predict whether a patient is suffering from Covd-19 or not [20]. Whereas in LOS we used OneHot encoding to make the data standardized for every input followed by Label encoding on the range of days column [21]. It is a supervised statistical technique which is used to check the inter-relation between a set of input data or variables. These two techniques help to pre-process data to transform the data in a model-acceptable format to get a desired result [22].

TABLE I. A PART OF LOS DATASET

case_id	Hospital_code	Hospital_type_code	…	Admission	Stay
1	8	c	…	4911	0-10
2	2	c	…	5954	41-50
…	…	…	…	…	…
318438	19	a	…	4752	0-10

IV. EVEDENTIAL MODELS

In this section, we are going to explore our approach towards the solution and explain our models.

A. Covid-19 Detection Model

Our CNN model is a linear stack of convolutional layers, Maxpooling, Dense layer with relu activation. The architecture of the customized CNN model is shown in figure 1. We also used Dropout which is used to reduce issues because of overfitting by improving generalization and providing restricted regularization by minimizing the sensitivity of model for the specific training input [23]. To minimize the model parameters we used separable convolutional and it enhance the performance of the model as compared to conventional convolutional operations [24]. The initialized number of separable convolutional filters are 32 and by the factor of 2 they are increased in successive convolutional layers. These separable convolutional layers are followed by MaxPooling2D and it is used to reduce the dimensions of the input data and allows us to make assumptions for features binned in the sub-regions, and 2D means the dimensions of the input data is in two dimensions like (x-axis, y-axis) [25]. And to optimize the parameters we have used Adam optimizer [26]. For classification of Chest X-rays to their respective categories we used optimal parameters for evaluation and training of model. And after the training of the model we saved the trained model in ".h5" file format for further use with an accuracy of 91.41% approx.

B. Length of Stay Prediction Model

Here we have used the catboost model for the prediction of length of stay of a patient at the hospital [27]. The catboost model is used for prediction because it facilitates the data to learn from previous steps and helps improve the accuracy of the model. We can also set variable number of parameters such as iterations, learning rate, depth, loss function and many more according to the requirements of the program [28].
We are going to explain all the settings, we have used in our model. These are explained as follows.

1) Iterations: The iterations setting is used to optimize the model as per the needs. The model then trains for the specified number of iterations and in each iteration the learning rate of the model increases.

2) Learning rate: This parameter is used to reduce gradient step. The value of learning rate is inversely proportional to the duration of training of the model. We also protected our model from overfitting using the optimized value for this setting.

3) Depth: The value of the depth parameter depends a lot on the processing unit type (CPU or GPU) and its value is used to determine the depth of the tree.

4) Loss function: The value of this metric is used to represent the goals of the model [29]. The value we have used for this setting is 'MultiClass' which represents that we are going to have more than two categories in our prediction results.

5) Eval_metric: This metric is used to detect whether the model is overfitting or not [30]. We used 'accuracy' metric in our model and it provided the accuracy of our model after each step which increased gradually.

When we successfully created our model with the use of above-mentioned settings and parameters, we fit our model on the training part of our dataset so that our model could learn from this data and get ready for the predictions. This process originally takes time to complete because of the size of the dataset and also the number of iterations.

When the process of training our model is finished, the

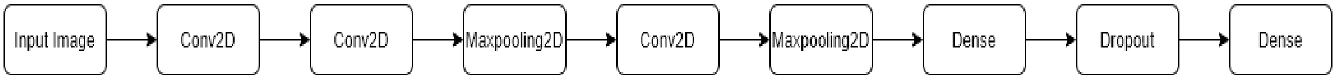

Fig. 2. Architecture of customized CNN

model can be used for predictions later. To save the time of training our model every time, we have saved the trained model in '.h5' file which can be accessed by the program whenever it runs again. This process saves a lot of time and increases the reusability and efficiency of the program as well.

V. RESULTS AND VIEWS

In the section, we have discussed the results achieved from each of the models. The optimized values for all the settings and metrics are attained using the research works. We tuned several parameters to optimize the model and got a great accuracy metric. The average accuracy of prediction of the trained model with different inputs is 92.33% approximately. This score of accuracy can be translated as a fact that the model is reliable and expected to produce good results and the model is learning iteratively so it can get better with application.

VI. CONCLUSION

The impact of Covid-19 on people all over the world and national economies is enormously negative. Chest X-rays take a revolution in diagnostic approach as they play a massive role in the detection of a person whether that person is suffering from Covid-19 or not. When we performed modality-strict learning on the collected CXRs, we learnt about the various modality-strict features of them. This model learns about the data over a number of steps and the loss experienced by the model at each step is rectified and fed to the model in the next step to increase the accuracy metric of the model. When the model predicts whether the patient is Covid-19 positive or negative, our next model will come in to action. The LOS model which is supposed to make the management of staff and resources of a hospital efficient. Using this model, the hospitals can do early interference in alien cases and reduce any complications. We have used an openly available from kaggle to treat the model and created the model using the catboost library. It is also a self-learning model and learns from the data over the steps. For future studies, this study is proposed to extend and predict time of operations and other diseases. This study can make the functionality of hospitals systematic and productive as they can find out which person is infected and they can proceed for the cure of the disease faster and predict how much time patient needs to spend in hospital.

REFERENCES

[1] Y. Xie *et al*., "Revealing the Mechanism of SARS-CoV-2 Spike Protein Binding With ACE2," in *Computing in Science & Engineering*, vol. 22, no. 6, pp. 21-29, 1 Nov.-Dec. 2020, doi: 10.1109/MCSE.2020.3015511.

[2] Rohaun, S. K. (2020). The Emergence Of Covid-19 And Its Spread Along With Symptoms. *COVID-19 Pandemic Update 2020,* 54-72. doi:10.26524/royal.37.4

[3] S. Hu et al., "Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification From CT Images," in IEEE Access, vol. 8, pp. 118869-118883, 2020, doi: 10.1109/ACCESS.2020.3005510.

[4] Ter-Sarkisov, A. (2020). COVID-CT-Mask-Net: Prediction of COVID-19 from CT Scans Using Regional Features. doi:10.21203/rs.3.rs-104621/v1

[5] Deep Transfer Learning-based COVID-19 prediction using Chest X-rays. (n.d.). doi:10.37473/dac/10.1101/2020.05.12.20099937

[6] Punitha, S., Al-Turjman, F., & Stephan, T. (2020). Genetically Optimized Computer-Aided Diagnosis for Detection and Classification of COVID-19. *AI-Powered IoT for COVID-19,* 105-122. doi:10.1201/9781003098881-5

[7] Cai, C., Chen, L., Zhang, X., & Gao, Z. (2020). End-to-End Optimized ROI Image Compression. *IEEE Transactions on Image Processing, 29*, 3442-3457. doi:10.1109/tip.2019.2960869

[8] Gao, T. (2020). Chest X-ray image analysis and classification for COVID-19 pneumonia detection using Deep CNN. doi:10.21203/rs.3.rs-64537/v2

[9] Preparedness and response to community spread of COVID-19 governmental and national recommendations for COVID-19 by the Pan-Academic Action Committee. (2020). *Epidemiology and Health, 42*. doi:10.4178/epih.e2020020

[10] Nouaouri, I., Samet, A., & Allaoui, H. (2015). Evidential data mining for length of stay (LOS) prediction problem. *2015 IEEE International Conference on Automation Science and Engineering (CASE)*. doi:10.1109/coase.2015.7294296

[11] Ibrahim, A. A., L., R., M., M., O., R., & A., G. (2020). Comparison of the CatBoost Classifier with other Machine Learning Methods. *International Journal of Advanced Computer Science and Applications, 11*(11). doi:10.14569/ijacsa.2020.0111190

[12] Lee, J. (2020). COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image (Preprint). doi:10.2196/preprints.19407

[13] T. Wang, Y. Zhao, L. Zhu, G. Liu, Z. Ma and J. Zheng, "Lung CT image aided detection COVID-19 based on Alexnet network," 2020 5th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, 2020, pp. 199-203, doi: 10.1109/CCISP51026.2020.9273512.

[14] He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770-778, Jun. 2016.

[15] Figure 5: Modality-specific knowledge transfer workflow. (n.d.). doi:10.7717/peerj.8693/fig-5

[16] Asari, V. (2001). Training of a feedforward multiple-valued neural network by error backpropagation with a multilevel threshold function. *IEEE Transactions on Neural Networks, 12*(6), 1519-1521. doi:10.1109/72.963789

[17] Error Backpropagation. (n.d.). *Neural Networks,* 69-87. doi:10.4135/9781412985277.n5

[18] D. H. Gustafson, "Length of stay prediction and explanation," Health Services Research, vol. 37, no. 3, pp. 631–645, 2002

[19] A. W. Woods, F. N. MacKirdy, B. M. Livingston, J. Norrie, and J. C. Howie, "Evaluation of predicted and actual length of stay in 22 scottish intensive care units using the apache iii system." Anaesthesia, vol. 55, no. 11, pp. 1058 – 1065, 2000

[20] Zhang, F., Dvornek, N., Yang, J., Chapiro, J., & Duncan, J. (2020). Layer Embedding Analysis in Convolutional Neural Networks for Improved Probability Calibration and Classification. *IEEE Transactions on Medical Imaging, 39*(11), 3331-3342. doi:10.1109/tmi.2020.2990625

[21] Bharitkar, S. (2019). Encoding in Neural Networks. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). doi:10.1109/icmla.2019.00065

[22] Lin, Z., Ding, G., Han, J., & Shao, L. (2018). End-to-End Feature-Aware Label Space Encoding for Multilabel Classification With Many Classes. *IEEE Transactions on Neural Networks and Learning Systems, 29*(6), 2472-2487. doi:10.1109/tnnls.2017.2691545

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting", Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014

[24] . F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions" in arXiv: 1610.02357, 2018, [Online] Available: https://arxiv.org/abs/1610.02357

[25] Bachtiar, Y., & Adiono, T. (2019). Convolutional Neural Network and Maxpooling Architecture on Zynq SoC FPGA. *2019 International Symposium on Electronics and Smart Devices (ISESD)*. doi:10.1109/isesd.2019.8909510

[26] Khan, A. H., Cao, X., Li, S., Katsikis, V. N., & Liao, L. (2020). BAS-ADAM: An ADAM based approach to improve the performance of beetle antennae search optimizer. *IEEE/CAA Journal of Automatica Sinica, 7*(2), 461-471. doi:10.1109/jas.2020.1003048

[27] Postnikov, E. B., Esmedljaeva, D. A., & Lavrova, A. I. (2020). A CatBoost machine learning for prognosis of pathogen's drug resistance in pulmonary tuberculosis. *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*. doi:10.1109/lifetech48969.2020.1570619054

[28] Zhu, T., Luo, L., Zhang, X., & Shen, W. (2017). Modeling the Length of Stay of Respiratory Patients in Emergency Department Using Coxian Phase-Type Distributions with Covariates. *IEEE Journal of Biomedical and Health Informatics,* 1-1. doi:10.1109/jbhi.2017.2701779

[29] Boţ, R., Heinrich, A., & Wanka, G. (2014). Employing different loss functions for the classification of images via supervised learning. *Open Mathematics, 12*(2). doi:10.2478/s11533-013-0342-5

[30] Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys, 27*(3), 326-327. doi:10.1145/212094.21211