A Novel Deep Learning Approach for Detection of Pneumonia from Chest X-rays

Andrew Yuan from Lynbrook High School

Statement of Purpose

- Pneumonia is an infection that inflames the air sacs in lungs. It is the leading cause of death for children under 5. In 2017, 2.56 million people died from pneumonia worldwide, of which almost a third were children younger than 5 years old.
- Previous studies show that detecting pneumonia with deep learning from chest X-rays give one of the lowest diagnosis accuracies among 14 common lung diseases.
- This study proposes a novel deep learning approach that shows substantial improvement on pneumonia diagnosis accuracies.
- The deep learning model built with this approach can be continuously improved over time and used for other types of image classifications.

Background

- Previous studies show that, with NIHCC dataset, pneumonia detection AUC scores were second lowest among 14 lung diseases
- All these studies use transfer learning with individual pre-trained CNNs only
- One of my goals is to prove that with my approach will give better performance

	Wang et al. 2017 [2]	Yao et al. 2017 [3]	Rajipurkar et al. 2017 [7]	Li et al. 2018 [8]	Li et al. 2018 [8]
14 lung diseases	ResNet-50	DenseNet-based	DenseNet-121	DenseNet-121	DenseNet-RNN
atelectasis	0.716	0.772	0.8094	0.811	0.776
cardiomegaly	0.807	0.904	0.9248	0.882	0.806
effusion	0.784	0.859	0.8638	0.884	0.86
infiltration	0.609	0.695	0.7345	0.714	0.691
mass	0.706	0.792	0.8676	0.846	0.878
nodule	0.671	0.717	0.7802	0.77	0.826
pneumonia	0.633	0.713	0.768	0.745	0.751
pneumothorax	0.806	0.841	0.8887	0.889	0.727
consolidation	0.708	0.788	0.7901	0.802	0.86
edema	0.835	0.882	0.8878	0.899	0.78
emphysema	0.815	0.829	0.9371	0.915	0.84
fibrosis	0.769	0.767	0.8047	0.812	0.892
Pleural Thickenin	0.708	0.765	0.8062	0.807	0.763
hernia	0.767	0.914	0.9164	0.831	0.77

My Approach Overview

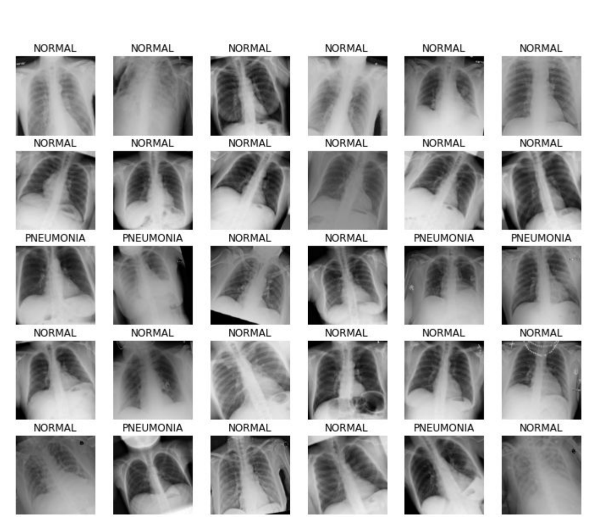
- Use the same raw dataset from NIHCC and pre-process the raw data for deep learning use
- Determine three most suitable pre-trained CNN candidates for this study
- Train each of the three selected CNNs with the pre-processed dataset separately using deep transfer learning
- Build the neural network model to combine all the three trained CNNs and train it to produce the final prediction

Pre-Processing Raw Data

- Selected 60,683 images (normal and pneumonia) from total of 121,120 chest x-ray images downloaded from NIHCC, then further divide them into 3 datasets randomly: Train (80%), Validation (10%) and Test (10%) set
- Downsized all images in 3 datasets to 224 x 224 (x3) as deep learning model input requires
- Fixed the data imbalance issue via oversampling and undersampling
- Applied data augmentation to train dataset
- Applied data normalization to all 3 datasets

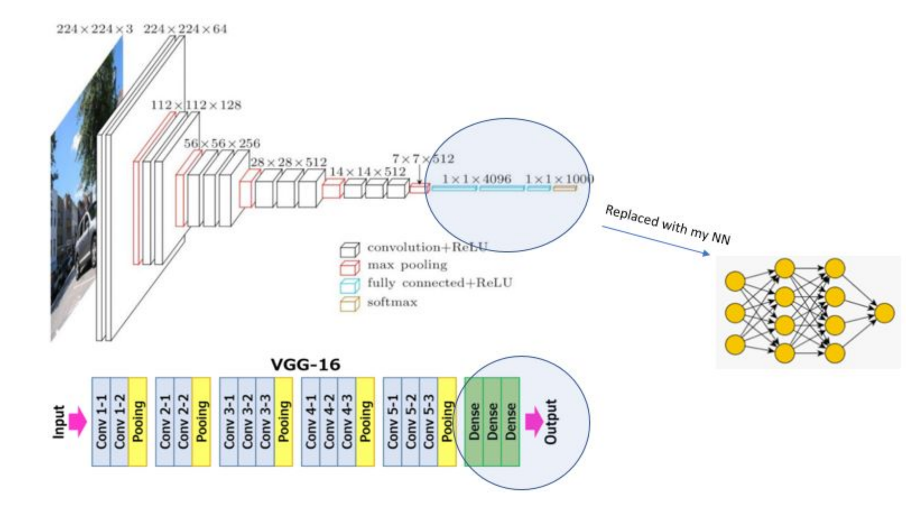
NIH Clinical Center America's Research Hospital

https://nihcc.app.box.com/v/ChestXray-NIHCC



Applying Deep Transfer Learning

- Use an ImageNet-trained convolutional neural network (CNN) as a starting point
- Replace the top of the model with my own neural network
- Train it with my pre-processed chest X-ray image dataset for detection of pneumonia



Selecting Three Best Pre-Trained CNNs

- Past and current studies from others show that **ResNet** and **DenseNet** give the best performances

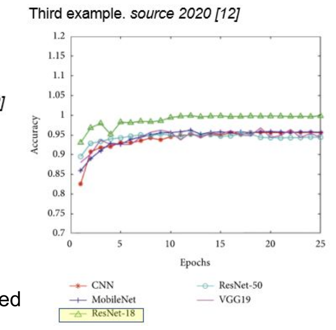
Real example of detecting pneumonia, *source 2020 [5]*

Architecture	Testing Accuracy	Testing Loss
ResNet18	97.29	0.096
DenseNet121	98.00	0.064
Inception	97.00	0.098
Xception	96.57	0.101
MobileNetV2	96.71	0.096

another example (numbers are AUC score). *Source 2017 [2]*

Pre-trained CNNs used by Wang et al. 2017 [2]	
AlexNet	0.5493
GoogLeNet	0.599
VGGNet-16	0.51
ResNet-50	0.6333

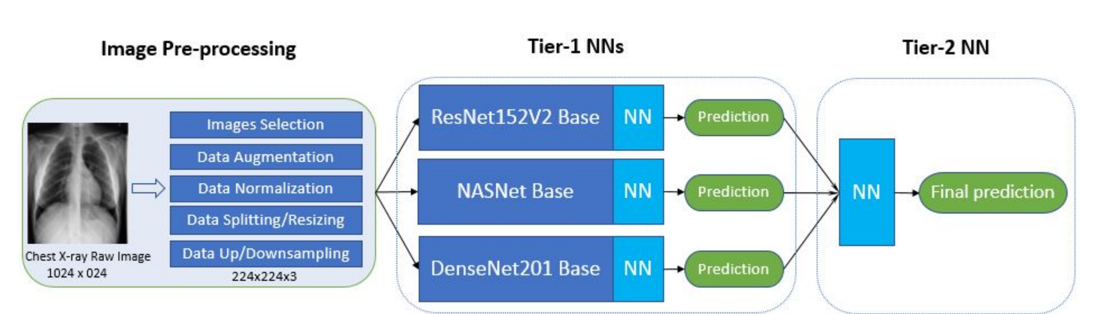
Third example. *source 2020 [12]*



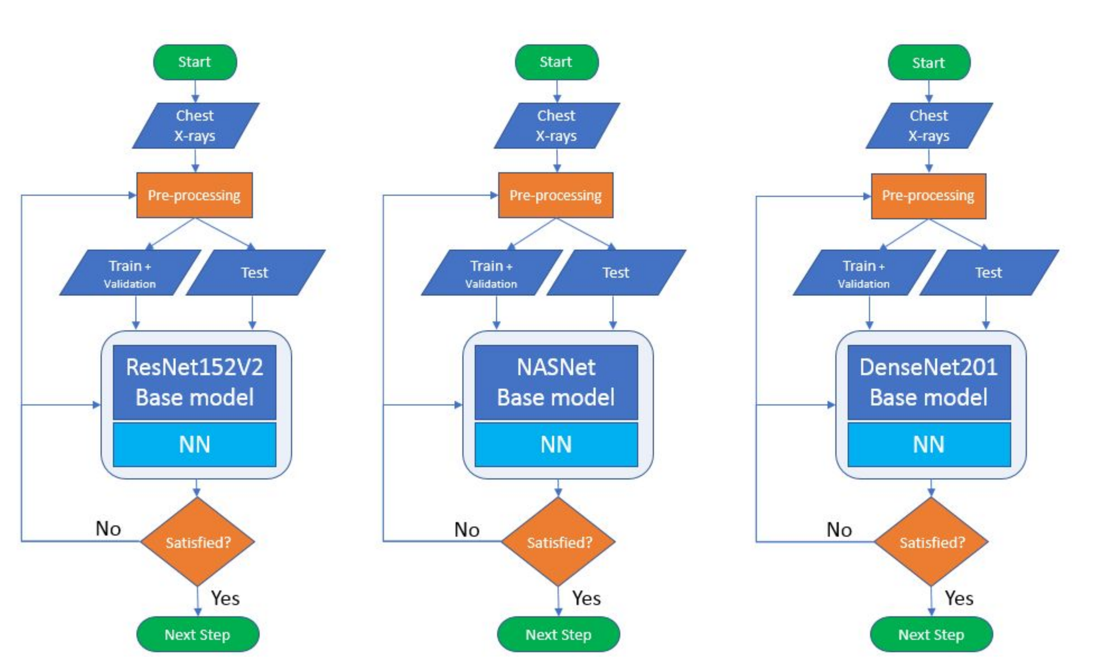
- I experimented with the following popular pre-trained CNNs and my custom CNN. Based accuracy, AUC, and F1-score, I select the highlighted 3 CNN models for this study: **NASNet(hub), ResNet152V2 and DenseNet201_fc1(256)**.

CNNs Tested	Total Parameters	Trainable Parameters	Training Time per epoch (mins)	Accuracy	Loss	Test Accuracy	AUC	F1-score
My custom CNN	6,446,369	6,446,369	2	0.8377	0.4958	0.909091	0.47931	0
VGG16	14,739,777	25,089	7	0.833	0.4666	0.714734	0.45755	0.06186
VGG19	20,049,473	25,089	13	0.8522	0.3589	0.902821	0.53365	0
NASNet(hub)	4,270,773	4,234,035	10	0.9800	0.4422	0.875	0.64	0.13333
NASNetMobile	4,321,461	51,745	9.0	0.7923	0.4932	0.871473	0.59168	0.25455
NASNetLarge	85,404,691	487,873	44.0	0.7645	1.8637	0.862069	0.63139	0.2069
ResNet50	23,688,065	100,353	7	0.753	0.6162	0.902821	0.59655	0
ResNet101	42,758,529	100,353	8	0.7933	0.7505	0.909091	0.62319	0
ResNet152V2	58,432,001	100,353	10	0.8708	0.5534	0.8724	0.65	0.19231
DenseNet201	18,416,065	94,081	6	0.8837	0.2928	0.874608	0.56492	
DenseNet201_fc1(256)	42,406,977	24,084,993	6	0.807	0.4283	0.877743	0.60143	0.09
DenseNet201_GlobalAvgPooling	18,323,905	1,921	6	0.8044	0.4809	0.909091	0.54804	0
EfficientNetB0	4,112,292	62,721	9.0	0.8164	0.5167	0.909091	0.58	0
EfficientNetB3	10,858,800	75,265	21.0	0.8707	0.5345	0.909091	0.50285	0

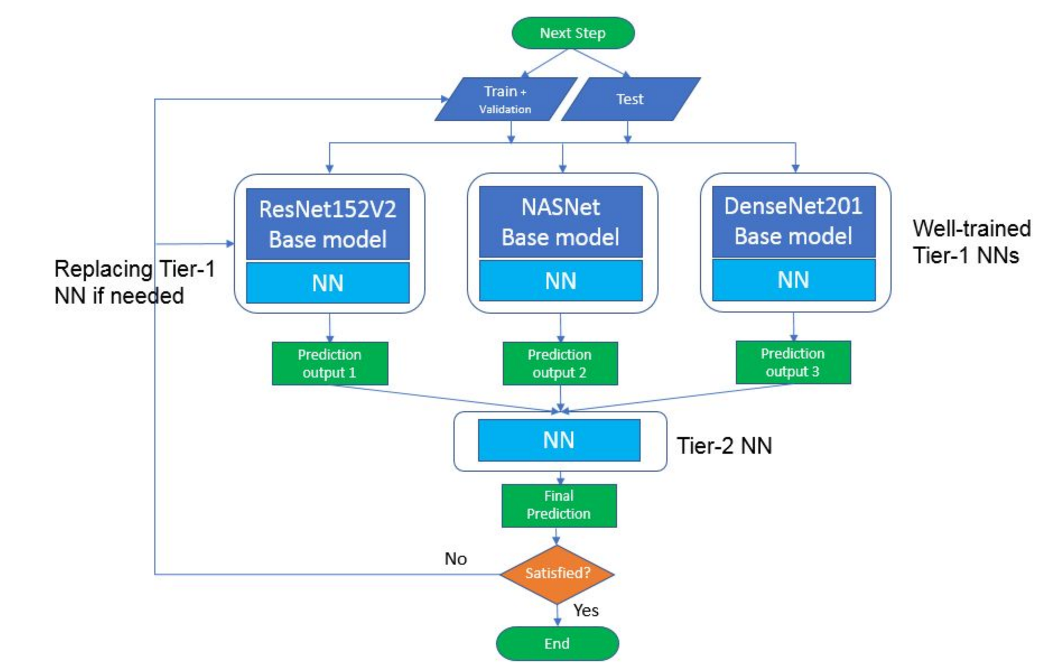
My Approach - Multi-tier NNs



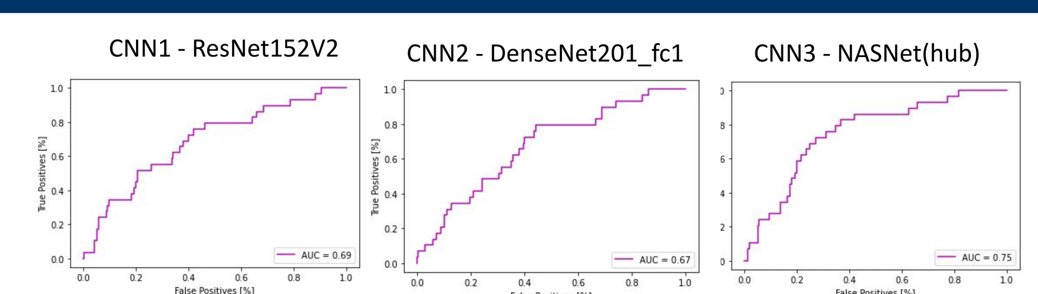
Training Tier-1 Individual NN Models Separately



Training Tier-2 NN with Well-trained Tier-1 NNs



Prediction Results of Tier-1 NNs



CNN1 - ResNet152V2

Accuracy	0.805643
Precision	0.188679
Recall	0.344828
F1_score	0.243902
Cohen's kappa	0.143216
ROC AUC score	0.686326

Confusion_matrix:

TrueP	FalseN	10	10
FalseP	TrueN	43	247

CNN2 - DenseNet201_fc1

Accuracy	0.909091
Precision	0.500000
Recall	0.034483
F1_score	0.064516
Cohen's kappa	0.053412
ROC AUC score	0.674316

Confusion_matrix:

TrueP	FalseN	1	28
FalseP	TrueN	1	289

CNN3 - NASNet(hub)

Accuracy	0.862069
Precision	0.241379
Recall	0.241379
F1_score	0.241379
Cohen's kappa	0.165517
ROC AUC score	0.752081

Confusion_matrix:

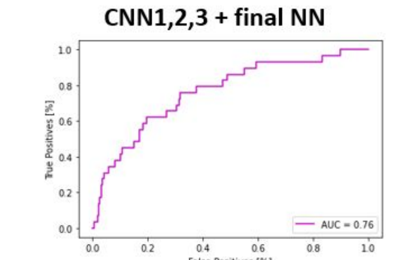
TrueP	FalseN	7	22
FalseP	TrueN	22	268

- Individual CNN predictions have high accuracy but low recall, F1 score, and Cohen's Kappa

Final Prediction Results from Tier-2 NN

- For the final prediction, ROC AUC score, recall, F1 score, and Cohen's Kappa are improved

CNN1,2,3 + final NN



Accuracy	0.830721
Precision	0.254902
Recall	0.448276
F1_score	0.325
Cohens kappa	0.236504
ROC AUC score	0.764685

Confusion_matrix:

TrueP	FalseN	13	16
FalseP	TrueN	38	252

Performance Metrics

Accuracy: $\frac{TP+TN}{TP+FN+FP+TN}$

Precision: $\frac{TP}{TP+FP}$

Recall: $\frac{TP}{TP+FN}$

F1 Score: $2/(\frac{1}{precision}+\frac{1}{recall})$

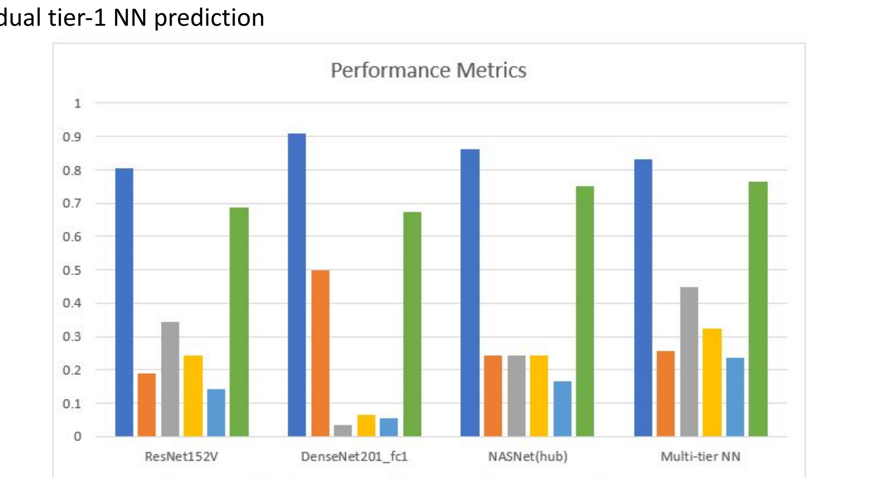
Cohen's Kappa: $\frac{p_o - p_e}{1 - p_e}$

ROC AUC: area under curve of TP vs. FP

TP = true positive
FN = false negative
FP = false positive
TN = true negative
p_o = observed probability
p_e = expected probability

Final Prediction Results vs. Individual Tier-1 NN Results

- Final prediction gives the best ROC AUC score of 0.765
- Recall, F1 score, and Cohen's Kappa are improved substantially (over 30%) for the final prediction
- Only precision and accuracy slightly decreased for the final prediction when compared to the best individual tier-1 NN prediction



preformance metrics	ResNet152V	DenseNet201_fc1	NASNet(hub)	Multi-tier NN	Improvement
	CNN1	CNN2	CNN3	Tier-1 and Tier-2 NN	
Accuracy	0.805643	0.909091	0.862069	0.830721	-9%
Precision	0.188679	0.5	0.241379	0.254902	-49%
Recall	0.344828	0.034483	0.241379	0.448276	30%
F1_score	0.243902	0.064516	0.241379	0.325	33%
Cohen's kappa	0.143216	0.053412	0.165517	0.236504	43%
ROC AUC score	0.686326	0.674316	0.752081	0.764685	2%

My Results vs. Others' Results

AUC score Comparison

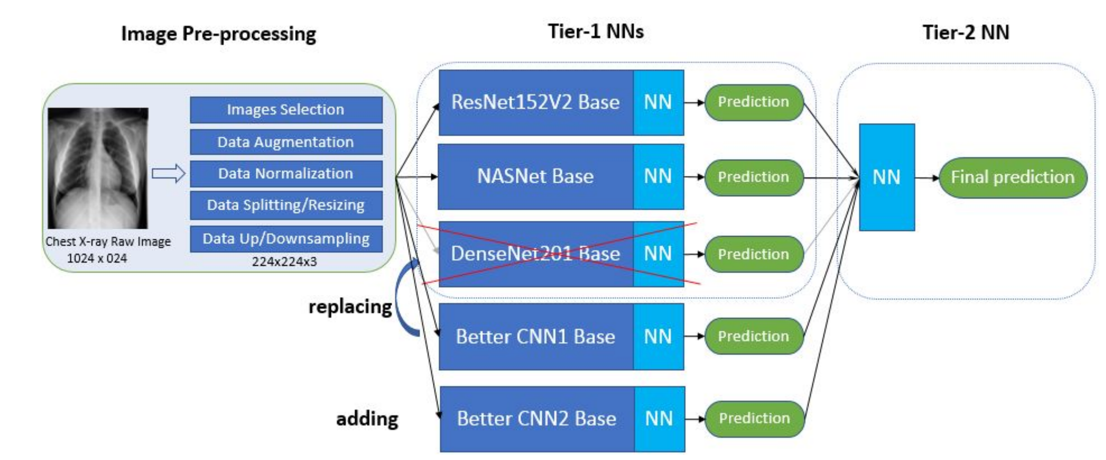
Researchers	Wang et al. 2017 [2]	Yao et al. 2017 [3]	Rajpurkar et al. 2017 [7]	Li et al. 2018 [8]	Li et al. 2018 [8]	My Research
CNN archetectures	ResNet-50	DenseNet-based	DenseNet-121	DenseNet-121	DenseNet-RNN	Multi-tier CNN
pneumonia	0.633	0.713	0.768	0.745	0.751	0.765

*All of these studies use the same NIHCC dataset

- My AUC score of 0.765 is on par with Rajpurkar et al. 2017 (0.768) and better than all other four which are 0.633, 0.713, 0.745, 0.751.

My Model Supports Incremental Learning

- My multi-tier NN model employs incremental learning
- It can be continuously improved over time by adding more accurate tier-1 NNs or replacing the existing tier-1 NNs with better performed CNNs and re-training the Tier-2 NN after replacement



Conclusions

- With my multi-tier deep learning approach, I achieved an AUC of 76.5%, on par or better than other past and current research results
- Compared to my individual tier-1 NNs' results, my final prediction gives an overall better performance
- My multi-tier NN model employs incremental implementation. It can be continuously improved over time.

Further Research

- Extend my model from binary classification to multiclass classification
- Apply the new model for detection of lung cancer and other lung diseases

Acknowledgement

- Mr. Lester Leung as the mentor from Lynbrook High School who helped me with logistics and mentored me for this project
- NIH Clinical Center for providing the raw chest X-ray images*

*NIH download site https://nihcc.app.box.com/v/ChestXray-NIHCC