ARTIFICIAL INTELLIGENCE

IN

LUNGS DISEASE CLASSIFCATION

1. Abstract:

The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal. A subset of artificial intelligence is machine learning, which refers to the concept that computer programs can automatically learn from and adapt to new data without being assisted by humans.

Lung disease is common throughout the world. These include chronic obstructive pulmonary disease, pneumonia, asthma, tuberculosis, fibrosis, etc. Timely diagnosis of lung disease is essential.

If we can detect these dangerous infections early, we can reduce the mortality and morbidity associated with them.

So, main objective is to use Machine Learning, which not only gives faster results but also demonstrates higher accuracy in the prediction process.

2. Problem Statement:

Every year, Pathologists diagnose 14 million new patients with lung diseases around the world. Most of them have 96-98 % success rate in diagnosing but have only 60% accuracy in prognoses. Pathologists take 10 or more days to evaluate the prognosis process. To overcome this hazardous situation and to help more and more people fight these diseases, Machine Learning technology is used in prediction these diseases. Machine learning takes huge amount of data & gives accurate output in seconds faster than pathologists. Finally, something we are certain of is that ML is the next step of pathology.

3. Market Need:

Customers like great and fast service. Pathologists take 10 or more days to evaluate the prognosis process, which is a huge time for patient as the spread of infected cells are rapid once they are produced. With the advent of the Internet of Things technology, there is so much data out in the world that humans can’t possibly go through it all. That’s where machines help us. They can do work faster than us and make accurate computations and find patterns in data. This will benefit the customer and brings greater profits to the pathology firm by adapting this advanced technology.

4. Target Specifications and Characterization:

A. To change traditional pathology process to faster and accurate process.

B. Reducing frustration and death of patients due to delay in the prognosis process

C. Predetermined dataset of X-Ray images of infected lungs of patients and of normal lungs of patients is taken and based on that prediction is performed

Above, mentioned targets can be achieved by analysing:

1. What the patient looks for
2. How are present pathology processes are being performed
3. Problems faced by people suffering from disease
4. How to identify and provide treatment in initial stage accurately.
5. How efficiently are the pathologists performing prognosis process
6. When and where a patient likes to trust and spend on?
7. Analysing the needs of the patients
8. To help patient fight in early stages
9. To send results to the patient within minutes
10. To remind the patient about the latest changes in the prognosis process

5. External Searches:

5.1. Applications of Machine Learning in Lungs Disease Prediction and Prognosis

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to “learn” from past examples and to detect hard-todiscern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in diagnosis and detection.

5.2. Dataset :

X-ray images of pneumonia, tuberculosis, normal lungs etc. can be collected from various websites (Kaggle, GitHub, data-world).

5.3. Machine learning is the future of Chest Disease prediction:

AI is set to change the medical industry in the coming decades — it wouldn’t make sense for pathology to not be disrupted too. Currently, ML models are still in the testing and experimentation phase for prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models. Machine Learning is the next step forward for us to overcome this hurdle and create a high accuracy pathology system.

6. Benchmarking:

The benchmark model will be a model of vanilla CNN. In this proposed work, “vanilla CNN” have been used. The architecture or structure of the vanilla CNN is described below in the fig.

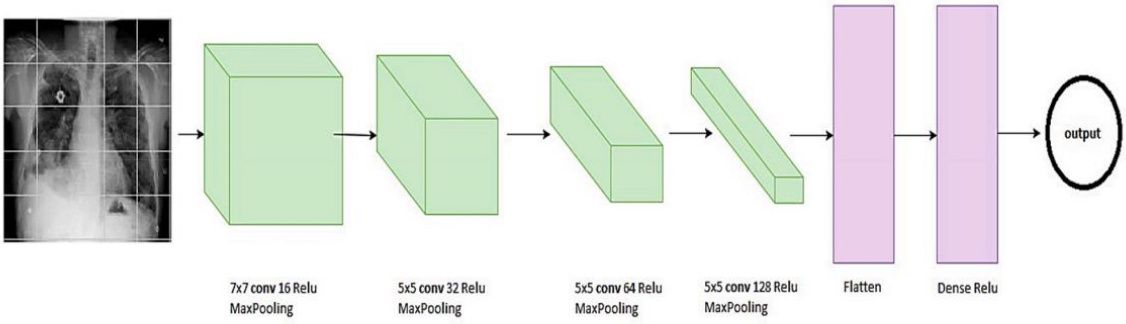


Fig. 8. Structural design for the model of vanilla CNN.

Fig. shows a model of vanilla CNN where there are four convolutional layers each followed by maximum pooling operation. The convolution layers are growing in depth. Next, is the flattening layer which is followed by a fully connected (FC) dense layer. Finally, the classification output is obtained.

7. Applicable Regulations (Government and Environmental)

a. Patents on ML algorithms developed

b. Laws related to privacy for collecting data from users

c. Protection/ownership regulations

d. Creating an e-mail service to mail the report to the patient and doctor.

e. Being responsible by design.

f. Ensuring open-source, academic and research community for an audit of Algorithms.

g. Review of existing work authority regulations

8. Applicable Constraints Expertise:

A. Requires a lot of research to obtain universal dataset of cancer patients in-order to provide more sophisticated and accurate results.

B. Establishing e-mail service in the product which have to send the report after the machine learning model is deployed in any server.

C. Confidential health data to be obtained to train the model.

D. Thorough understanding of dataset and verification of the results must be performed by the pathologist from the machine learning model to provide a great health prescription and service to the user.

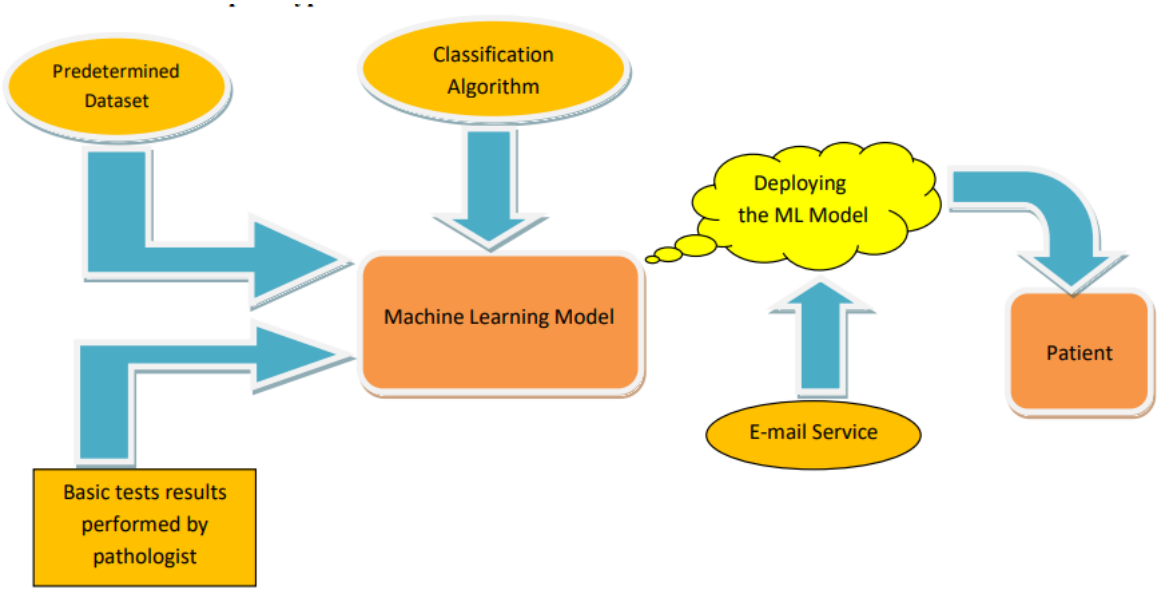
9. Business Opportunity:

Pathologists are pretty good in diagnosing while they are not so good in the prognosis. It takes more than two weeks to identify the kind of disease in an individual. To overcome this hazardous circumstance, our main objective is to use Machine Learning, which not only gives faster results but also demonstrates higher accuracy in the prediction process.

10. Concept Generation:

Machine learning not only predicts lungs diseases prognosis results faster but also gives higher accuracy which is around 70 to 80 % which is greater when compared with the pathologists. Chest diseases which when treated early would save many lives. So, machine learning is most sort after technique which is very useful in replacing present prognosis process.

11. Final Product prototype:



12. Product details:

It’s a system which takes in data, finds patterns, trains itself using the data and outputs an outcome. ML has key advantages over Pathologists. Firstly, machines can work much faster than humans. A biopsy usually takes a pathologist 10 days. A computer can do thousands of biopsies in a matter of seconds. Machines can do something which humans aren’t that good at. They can repeat themselves thousands of times without getting exhausted. After every iteration, the machine repeats the process to do it better. Humans do it too, we call it practice. While practice may make perfect, no amount of practice can put a human even close to the computational speed of a computer. Machines have greater accuracy. With the advent of the Internet of Things technology, there is so much data out in the world that humans can’t possibly go through it all. That’s where machines help us. They Predetermined Dataset Machine Learning Model Classification Algorithm Basic tests results performed by pathologist Deploying the ML Model E-mail Service Patient can do work faster than us and make accurate computations and find patterns in data. That’s why they’re called computers.

12.1. Algorithm:

CNN and CapsNet:

CNN can be considered as one of the most powerful deep learning based network that can contain multiple hidden layers. These hidden layers are very effective in performing convolution and subsampling for the purpose of extracting low to high levels of features of the input data. Capsule networks have been developed to maintain the positions of objects and their properties in the image, and to model their hierarchical relationships [39]. In the convolution neural networks, valuable information in the data comes to the fore with the pooling layer. Since the data is transmitted to the next layer by pooling, it may not possible for the network to learn small details [43]. In addition, CNN produces a scalar value in neural output. Capsule networks create vectorial output of the same size but with different routings, thanks to the capsules, which contain many neurons. The routings of a vector represent the parameters of the images [44]. CNNs use scalar input activation functions such as ReLU, Sigmoid, and Tangent. On the other hand, capsule networks use a vectorial activation function called squashing. One of the key features of this network is equivariance which keeps the spatial relationship of objects in an image without affecting the object’s orientation and size.

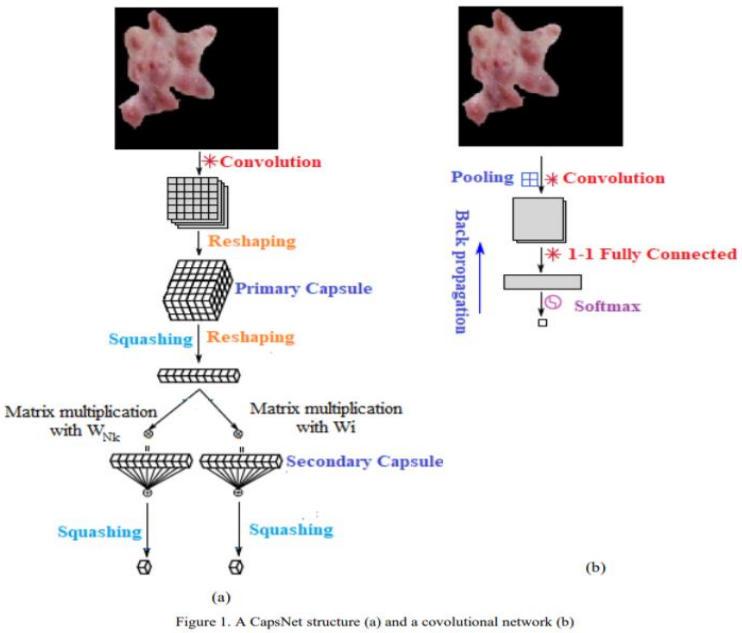


Figure 1. A CapsNet structure (a) and a covolutional network (b)

12.2. Working:

A basic CNN with three layers of ConvLayer is selected as the baseline model and the performance of CapsNet is compared with LeNet and the baseline model on four datasets. Their final result shows that CapsNet exhibits better performance than the other two networks for the case of a small and imbalanced dataset. The performance of CapsNet for the case of the large dataset is observed and compared with the other models. The performance capability of basic and modified CapsNet is also evaluated in terms of accuracy and training time calculation. So, a hybrid model is proposed in order to improve the training time and to detect the disease effectively with less number of tests.

CNN has a number of advantages for example, it can extract important features from images at low computational complexity. In this work, a number of aspects of CNN are considered. These are pre-processing parameters which can be sufficient tuning, training parameters, and data enhancement in the system not only lung X-ray images. Using the influence to discriminate several objects from various perspectives, the capsule network can be suitable for the reason that our lung X-ray image data has two categories of view positions. The capsule network is modified by tuning the training parameters.

12.3 Team required to develop:

1. Machine learning engineering
2. Business analyst
3. Software developer
4. Cloud engineer
5. Data Researcher

13. Conclusion:

AI is set to change the medical industry in the coming decades — it wouldn’t make sense for pathology to not be disrupted too. Currently, ML models are still in the testing and experimentation phase for prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models. While we might not see AI doing the job of a pathologist today, we can expect ML to replace our local pathologist in the coming decades, and it’s pretty exciting! ML models still have a long way to go, most models still lack sufficient data and suffer from bias. Machine learning can train just as well as doctor prognosis, it doesn’t require extra pay for prognosis. Manual treatment take long time to show the result, while machine learning gives output in seconds .To save people’s life and allow doctor to fully concentrate in diagnosis, Yet, something we are certain of is that ML is the next step of pathology, and it will disrupt the industry.

DATA ANALYSIS AND PREPROCESSING

Load data

In [2]:
```
import numpy as np
import pandas as pd

df = pd.read_csv("Dataset/data sample/sample_labels.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImageWidth	OriginalImageHeight	Origina
0	00000013_005.png	Emphysema|Infiltration|Pleural_Thickening|Pneu...	5	13	060Y	M	AP	3056	2544	
1	00000013_026.png	Cardiomegaly|Emphysema	26	13	057Y	M	AP	2500	2048	
2	00000017_001.png	No Finding	1	17	077Y	M	AP	2500	2048	
3	00000030_001.png	Atelectasis	1	30	079Y	M	PA	2992	2991	
4	00000032_001.png	Cardiomegaly|Edema|Effusion	1	32	055Y	F	AP	2500	2048	

PreProcessing

In [4]:
```
diseases = ['Cardiomegaly','Emphysema','Effusion','Hernia','Nodule','Pneumothorax','Atelectasis','Pleural_Thickening','Mass','Ede
#split diseases
for disease in diseases :
    df[disease] = df['Finding Labels'].apply(lambda x: 1 if disease in x else 0).astype(int)

df['No_Finding']=df['Finding Labels'].apply(lambda x: 1 if 'No Finding' in x else 0)
df['Finding']=df['Finding Labels'].apply(lambda x: 0 if 'No Finding' in x else 1)
```

In [5]:
```
df.head()
```

Out[5]:

	Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImageWidth	OriginalImageHeight	Origina
0	00000013_005.png	Emphysema|Infiltration|Pleural_Thickening|Pneu...	5	13	060Y	M	AP	3056	2544	
1	00000013_026.png	Cardiomegaly|Emphysema	26	13	057Y	M	AP	2500	2048	
2	00000017_001.png	No Finding	1	17	077Y	M	AP	2500	2048	
3	00000030_001.png	Atelectasis	1	30	079Y	M	PA	2992	2991	
4	00000032_001.png	Cardiomegaly|Edema|Effusion	1	32	055Y	F	AP	2500	2048	

5 rows × 27 columns

In [6]:
```
df['Patient Age']
```

Out[6]:
```
0       060Y
1       057Y
2       077Y
3       079Y
4       055Y
        ...
5601    058Y
5602    061Y
5603    052Y
5604    010Y
5605    024Y
Name: Patient Age, Length: 5606, dtype: object
```

In [7]:
```
#remove last character in Age
df['Age']=df['Patient Age'].apply(lambda x: x[:-1]).astype(int)
df['Age Type']=df['Patient Age'].apply(lambda x: x[-1:])

np.unique(df['Age Type'], return_counts=True)
```

Out[7]:
```
(array(['D', 'M', 'Y'], dtype=object), array([   1,    1, 5604], dtype=int64))
```

In [8]:
```
df['Age Type']
```

Out[8]:
```
0       Y
1       Y
2       Y
3       Y
4       Y
       ..
5601    Y
5602    Y
5603    Y
5604    Y
5605    Y
Name: Age Type, Length: 5606, dtype: object
```

Seeing that there is an age case is calculated in Day, and one case is in Month, so it's all in Year

In [9]:
```
df.loc[df['Age Type']=='M',['Age']] = df[df['Age Type']=='M']['Age'].apply(lambda x: round(x/12.)).astype(int)
df.loc[df['Age Type']=='D',['Age']] = df[df['Age Type']=='D']['Age'].apply(lambda x: round(x/365.)).astype(int)
```

In [10]:
```
df.describe()
```

Out[10]:

	Follow-up #	Patient ID	OriginalImageWidth	OriginalImageHeight	OriginalImagePixelSpacing_x	OriginalImagePixelSpacing_y	Cardiomegaly	Emphysem
count	5606.000000	5606.000000	5606.000000	5606.000000	5606.000000	5606.000000	5606.000000	5606.00000
mean	8.616661	14330.617017	2644.795755	2491.087406	0.155467	0.155467	0.025152	0.02265
std	15.565815	8411.477789	347.188754	399.119063	0.016201	0.016201	0.156599	0.14881
min	0.000000	13.000000	1362.000000	966.000000	0.115000	0.115000	0.000000	0.00000
25%	0.000000	7289.000000	2500.000000	2048.000000	0.143000	0.143000	0.000000	0.00000
50%	3.000000	13993.000000	2542.000000	2544.000000	0.143000	0.143000	0.000000	0.00000
75%	10.000000	20655.500000	2992.000000	2991.000000	0.168000	0.168000	0.000000	0.00000
max	177.000000	30797.000000	3266.000000	3056.000000	0.198800	0.198800	1.000000	1.00000

8 rows × 23 columns

In [11]:
```
df.iloc[[4242], :]
```

Out[11]:

	Image Index	Finding Labels	Follow-up #	Patient ID	Patient Age	Patient Gender	View Position	OriginalImageWidth	OriginalImageHeight	OriginalImagePixelSpacing_x	...	Mass	E
4242	00020900_002.png	No Finding	2	20900	411Y	M	AP	3056	2544	0.139	...	0	

1 rows × 29 columns

We see something special in the Age field, which is max 411, no one living up to 411 years is recorded, this is certainly a statistical error in the process.

In [12]:
```
df = df.drop(df['Age'].sort_values(ascending=False).head(1).index)
```

DATA VISUALISATION

Number of each diseases by patient gender

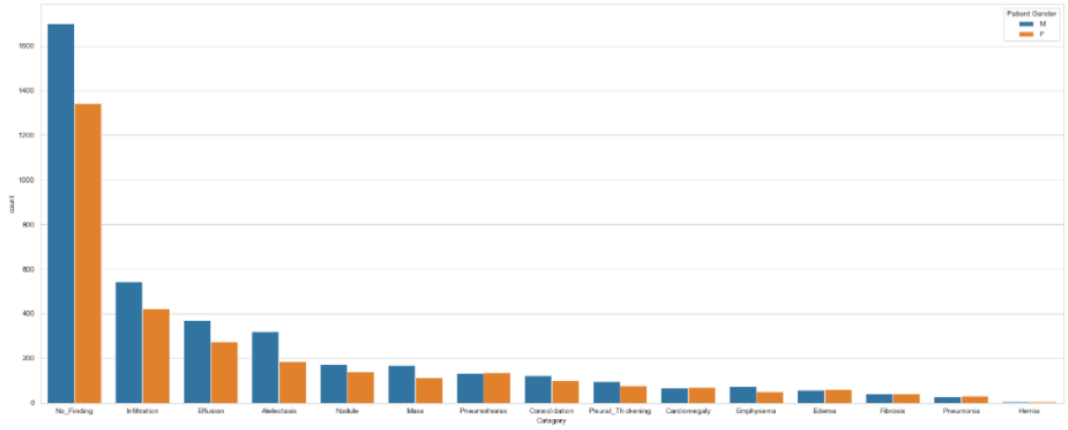
In [13]:
```
import seaborn as sns
import matplotlib.gridspec as gridspec
import matplotlib.ticker as ticker
sns.set_style('whitegrid')
%matplotlib inline
import matplotlib.pyplot as plt

plt.figure(figsize=(25,10))
ax = plt.subplot()

data = pd.melt(df, id_vars=['Patient Gender'], value_vars = list(diseases + ['No_Finding']), var_name = 'Category', value_name =
data = data.loc[data.Count>0]

sns.countplot(x='Category',hue='Patient Gender',data=data, ax=ax, order = data['Category'].value_counts().index)
```

Out[13]: <AxesSubplot:xlabel='Category', ylabel='count'>



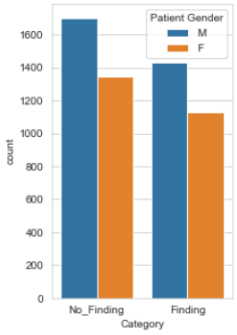
Distribution of Patient Gender and Finding diseases

In [14]:
```
plt.figure(figsize=(3,5))
ax = plt.subplot()

data = pd.melt(df, id_vars=['Patient Gender'], value_vars = list(['Finding', 'No_Finding']), var_name = 'Category', value_name =
data = data.loc[data.Count>0]

sns.countplot(x='Category',hue='Patient Gender',data=data, ax=ax, order = data['Category'].value_counts().index)
```

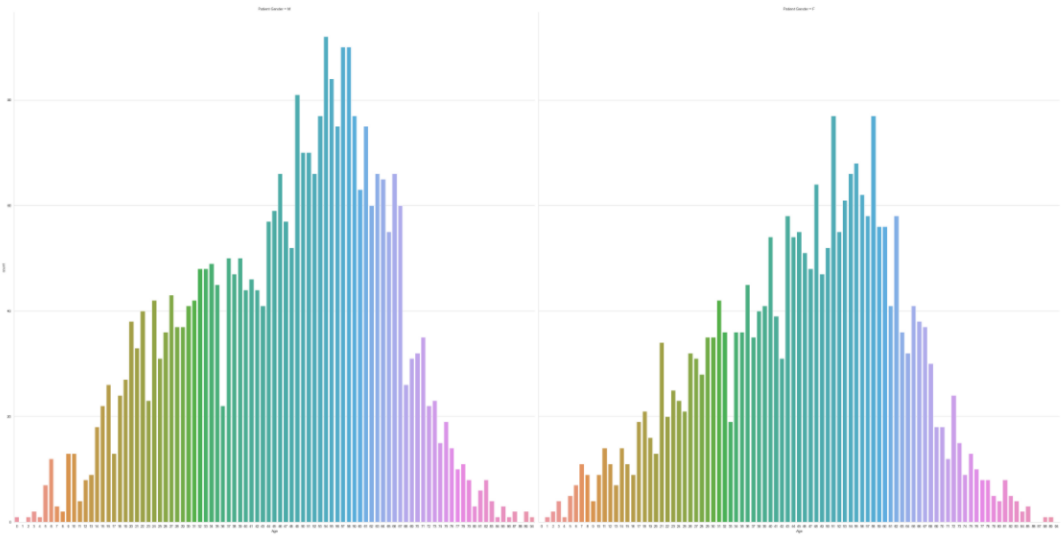
Out[14]: <AxesSubplot:xlabel='Category', ylabel='count'>



Age Distribution

In [15]:
```
sns.catplot(x="Age", col="Patient Gender", data=df, kind="count", height=20)
```

Out[15]: <seaborn.axisgrid.FacetGrid at 0x1ba91093d90>



Distribution of Age for each disease

In [16]:
```
f, axarr = plt.subplots(7, 2, sharex=True,figsize=(15, 20))

i=0
j=0
x=np.arange(0,100,10)
for pathology in diseases :
    g=sns.countplot(x='Age', hue="Patient Gender",data=df[df['Finding Labels']==pathology], ax=axarr[i, j])
    axarr[i, j].set_title(pathology)
    g.set_xlim(0,90)
    g.set_xticks(x)
    g.set_xticklabels(x)
    j=(j+1)%2
    if j==0:
        i=(i+1)%7
f.subplots_adjust(hspace=0.3)
```