Severstal Steel Defect Detection

**Introduction:**

Steel is one of the most important building materials of modern times. Steel buildings are resistant to natural and man-made wear which has made the material ubiquitous around the world. Serverstal is the Russian Steel producing company, looking for machine learning approach to improve automation, increase efficiency, and maintain high quality in their production. The production process of flat sheet steel is especially delicate. From heating and rolling, to drying and cutting, several machines touch flat steel by the time it’s ready to ship. This project will help engineers improve the algorithm by localizing and classifying surface defects on a steel sheet.

**Problem Statement:**

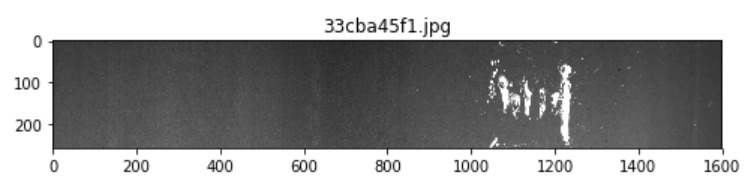
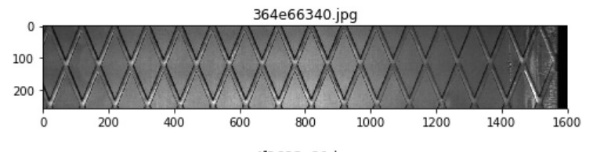
The objective of this project is to predict the location and type of defects found in steel manufacturing using the images provided. The images are named with a unique ImageId, and our task is to segment each image and classify the defects in the test set.

**Dataset:**

Data is taken from the Kaggle challenge, link: <https://www.kaggle.com/c/severstal-steel-defect-detection/data>. which is around size ~2GB and consists of 18076 images. There are 12568 images in the train set. 5506 images in the test set Of the training 12568 images, only 7095 images have defects, rest have no defects. Each image is of size 1600\*256. The location of the defects we given as encoded pixels. We have to decode them using a mask function to get the defect location.

**Data Description:**

Each Image may have no defects, a defect of a single class, or defects of multiple classes (ClassId = [1, 2, 3, 4]). Given a image, our task is to classify the defect and locate the segmentation of the defect. For each image you must segment the defects if it belongs to each of the class (ClassId = [1, 2, 3, 4]).



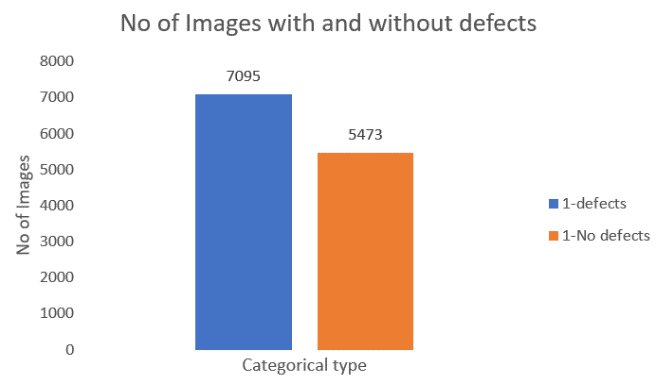
**Fig 1: left image with no defect and right images with defects**

Since all images are in the grey scale just by visually can not identified defects in them. Hence Using Machine learning and Keras Application like CNN, Resent they are classified and Where exactly defect is present is seen by Unet model.

**Task Performed:**

We have performed 4 tasks in this project for classification and object detection:

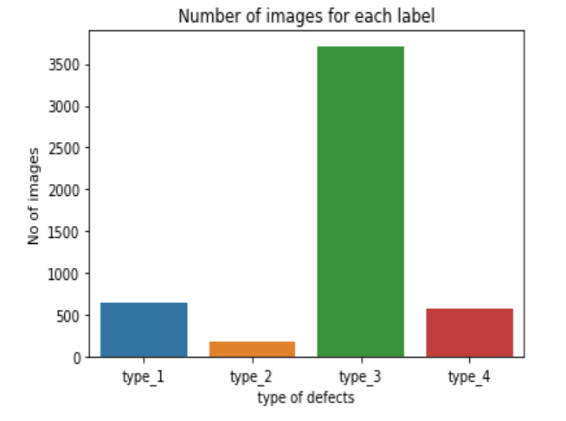
1. **Exploratory Data Analysis**: Data cleaning finding the defects types findng the
2. **Binary classification problem**: Identify whether an input images has a defect or no defects.
3. **Multi-class Classification problem**: Identify the number of class type of defects an image contains, each image can have up to 4 types of defects corresponding to classes [1,2,3,4]
4. **Object Detection**: For each defect type, identify regions of the image that has this defect. This is an image segmentation problem, where each pixel is labeled as having no defect or a defect of a particular type.

**Exploratory Data Analysis:**

We performed EDA, using bar chart we classified images which has defect and with no defects. From the whole dataset 5473 images with no defects were labeled as **0** and7095 images with defects were labeled as **1.** The number of images with and without defects are proportionate and the data seems quite balanced. From the fig 1, we concluded that data is balanced for doing binary classification.

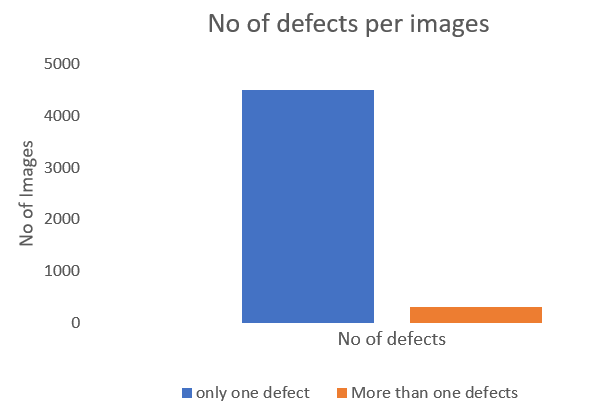
**Fig 2: Binary classification**

Now we will plot the bar graph to see class type of defects each image has, there are 4 types of defects **Type1** defect which conation small seed like scratch or crack on the steel sheet. **Type2** defect which has thin vertical crack on the images. **Type3** multiple vertical line cracks on sheet and Type4 has medium area size defects in images, In the segmentation plot we can see it more clearly the defects.

Categories of defects – Type1, Type2, Type3, Type4. From that Lot of images with defects belong to Type3. Very few samples for the type2 defects data is very imbalance here. So, to balance the data we tried data augmentation has performed for data augmentation resampling techniques also used and also tried SMOTE method to balance the data. Batch normalization had performed best compared to other methods.

**Fig 3: Multiclass classification**

EDA was also performed to check whether there are multilabel image, some of the images had more than one type of defects in them. Use bar graph these images were also classified.



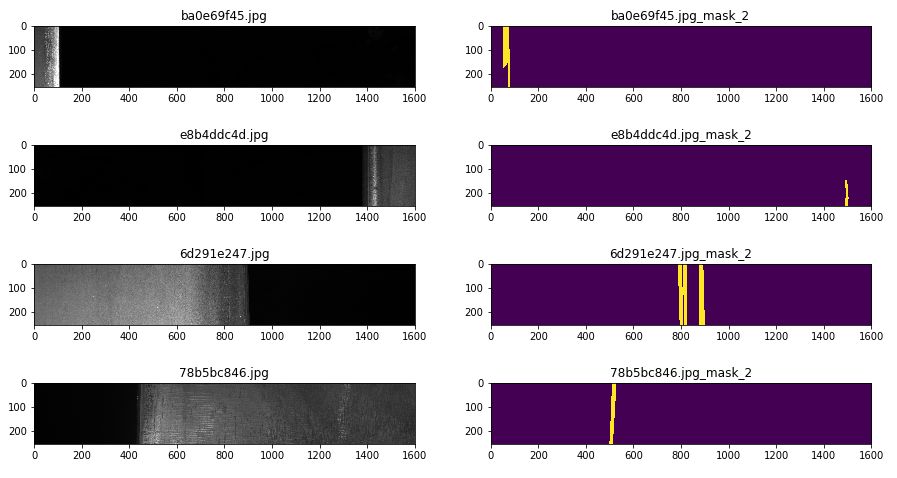
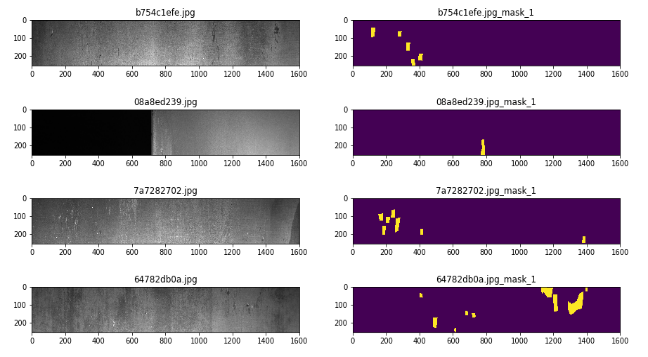
These images were classified as multilabel, since they were very few in number. We removed from the data. Now our data is resized and ready to feed into neural network for binary classification and multi class classification model.

**Fig 4: Multilabel classification**

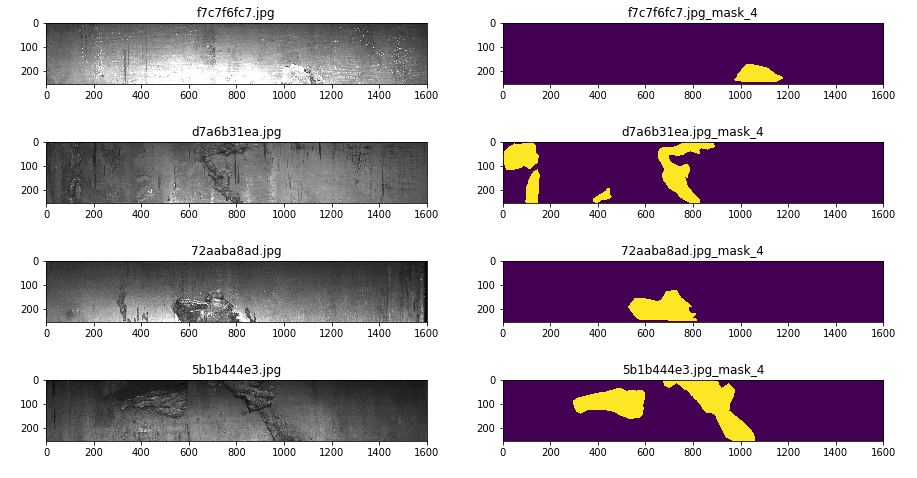
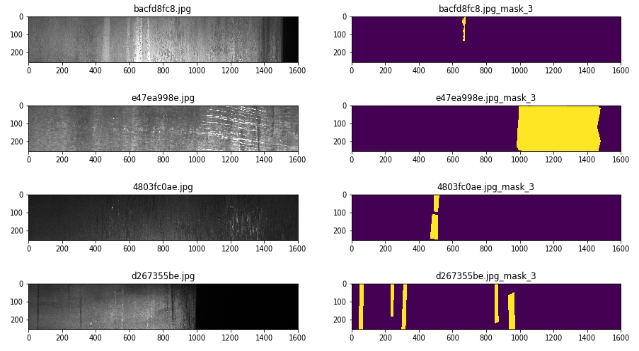
**EDA conclusion**:

The dataset is imbalanced thus we will use stratified sampling for splitting the dataset into train and validation datasets. This is a multi-label image segmentation problem. As there are around 50% of images with no defects, it is equally important to identify images with no defects. Convert masks to EncodedPixels and filter them as per classification probabilities.

**Segmentation:**



**Fig 5: Type 1 defects Fig 6: Type 2 defects**



**Fig 7: Type 3 defects Fig 8: Type 4 defects**