Introduction:

Machine Learning is a process by which computer can make prediction through analysing the input data and it is either curve fitting or classification tasks.[Domingos] In last few years, the use of machine learning has been increased tremendously due to increase of computational power. Nowadays, it is widely used in Web search, object detection, recommender systems, drug design and many other applications. A report published by McKinsey Global Institute claims that ML will revolutionize the future innovation [15].

In ML, data plays an important role and based on that the input features, algorithm, and accuracy metric are selected. In order to train the algorithm, the data is divided into training and testing datasets. Therefore, the first question that arise while implementing an algorithm is that how much data is required to train the model effectively. As per our knowledge, there is no definite answer to this, but in most scenario, it depends on various factors like complexity of the algorithm, input features, correlation between data etc. For example, non-linear algorithms will need more training data compared to linear models.

Related Work:

In [Claudia], it was proposed that training size should be defined by specifying confidence interval widths for classification algorithm in bio spectroscopy field. As mentioned in [ESL], increasing the training dataset will overfit the model. Hence, the model will adjust closely with training data and it will not generalize well. It was found in [Hajian-Tilaki] that how the performance of models vary with the training dataset size in biomedical applications. The investigation in [Cho] describe about how much training data is required to have an accurate model in medical image deep learning systems. In [Sun], it was found that it is possible to have a better accuracy in machine translation systems even with large training datasets.