Human Protein Atlas Analysis using Image Classification

Pragatheswari D[1], Manikanda Prabhu V[2], Sibinandhan S[3], Dhivya P[4]

[1] *U.G Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India*

[2] *U.G Student, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India*

[3] *U.G Student ,Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India*

[4] *Assistant Professor, Computer Science and Engineering, Bannari Amman Institute of Technology, Tamil Nadu, India*

ABSTRACT

Proteins play a crucial function in an organism's stability. Protein subcellular localization can provide vital insights about their purposes and arrangements in cells. The trained eye can clearly locate subcellular protein localizations from microscopy images, but streamlining the process is difficult. Training on widely imbalanced groups and forecasting multiple labels per picture is among the challenges. Many popular real-world applications, such as image recognition have been powered by the emerging field of deep learning. Deep neural networks, especially Convolutional Neural Networks (CNNs) usually take raw images as inputs, and learn hierarchical feature representations in an end-to-end process .In this paper, an effective CNN prediction architecture resnet50 is used to locate the subcellular protein from microscopy images in human protein atlas dataset. This method brings greater accuracy compared to other previous models.

Keywords—CNN, Resent50, Subcellular protein localization, Human protein atlas (HPA), Deep neural networks

1. INTRODUCTION

Proteins are the core components of life. Any cell in the human body contains protein. Protein is required in your diet to assist in the repair and formation of new cells in your body. We need a thorough overview of the cell's molecular topography to consider protein activity. Genetically encoded fluorescent proteins or immunofluorescence used to show proteins subcellular localization. However, owing to the diffraction limit of light, exact localization of proteins cannot be determined. Several microscopy methods have recently been used to break the diffraction barrier. Proteins can be pinpointed with a resolution of 20 nm or less using super resolution fluorescence microscopy.

Recognizing protein subcellular localization is critical for understanding both the role of individual proteins and the overall organization of the cell. From microscopy photographs, a skilled eye can easily locate subcellular protein localizations, but automating the procedure is complicated. Due to the recent improved techniques in deep learning helps identifying the patterns in images. CNN (Convolutional Neural Networks) is a multilayer deep learning neural network that learns patterns that automatically appear in images. In this work, the goal is to streamline the function of locating the subcellular proteins in the microscopic images obtained in human protein atlas dataset. As an outcome, this paper suggests a CNN architecture resnet50 in identifying the protein's subcellular locations in those images. Also with higher accuracy value, it would be easier for the health system to detect the proteins locations in those microscopy pictures.

2.MACHINE LEARNING

Machine learning is one of the widely used methods for image recognition. However, there are still regions of machine learning which need improvements can be developed. There are algorithm which has to train in massive amount in order to get prediction and decision. As an outcome, Deep learning systems are involved in image recognition.

2.1. Deep Learning Neural Networks

Deep learning, also known as neural networks, is a branch of machine learning that employs a computing paradigm that is heavily influenced by brain structure. To complete an operation, the machine must process layers of data between the raw input and result. The more layers it must process to obtain the result, the deeper the network is regarded.

2.2 Convolutional Neural Networks

A well-known approach in computer vision systems is the convolutional neural network, also denoted as convents or CNN. It's a type of deep neural network that's used to analyses visual data. This architecture is broadly used to identify objects in a photograph or film. Photo or video recognition, neural language analysis, and other applications use it. CNN is multilayer neural network that identifies reoccurring patterns in the images. This network has four layers that is a convolutional layer, Relu Activation layer, pooling layer and densely connected layer. Some famous CNN models for image classification is VGG-16, Inceptionv3, and EfficentNet. Among those ample models resnet50 a version of resnet architecture is used in implementing this work.

2.3. ResNet Model

ResNet, short for Residual Networks, it is the special type of neutral networks. They introduced to solve complex problems. We have pile of layer in Deep Learning, these gives us improved result and performance. The reason behind adding more layer is that to learn the more complex features. While we are recognizing image, initially we have detect the edges, next we have identity the textures and next is to detect the objects and so on. But we find that there is the maximum threshold for depth with CNN. While adding more layer it leads to decrement of performance. It makes the result to overfitting too. It is the main problem in Deep learning so this could be solved using ResNet or residual networks and these resnets are made up from residual Blocks. The concept of skip connection was first introduced in ResNet.

Fig - 1: ResNet Model Connection

The figure in the left is stacking convolution layers together and after the other. On the right we still stack convolution layers as before but we now also add the original input to output of the convolution block. This what is known as skip connection. The residual is the mapping is amount of error which can be added to input so as to reach the final destination i.e. to approximate the final function. The Residual Mapping is acting as a bridge between the input and the output of the block. Note that the weight layers and activation function are not shown in the diagram but they are actually present in the network. ***F(x) = H(x) – x.***"

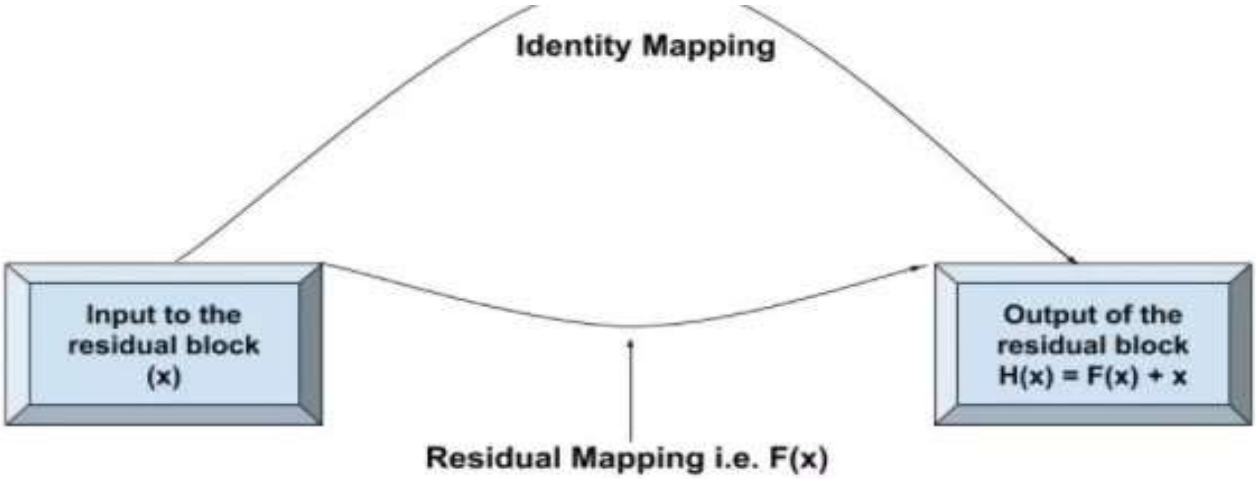


Fig – 2 : Mapping

The above statement is explaining that during training the deep residual network, the main focus is to learn the residual function i.e. ***F(x).*** So, if the network will somehow learn the difference (***F(x)***) between the input and output, then the overall accuracy can be increased. In other words, the residual value should be learned in a way such that it approaches zero, therefore making the identity mapping optimal. In this way, all the layers in the network will always produce the optimal feature maps i.e. the best case feature map after the convolution, pooling and activation operations. The optimal feature map contains all the pertinent features which can perfectly classify the image to its ground-truth class. ResNet is a powerful backbone model that is used very frequently in many computer vision tasks. ResNet uses skip connection to add the output from an earlier layer to a later
layer. This helps it mitigate the vanishing gradient problem

III. RELATED WORKS

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, Ole Winther[1] developed DeepLoc a prediction algorithm with deep neural networks to detect the eukaryotic protein subcellular locations by the reoccurring sequence information in the dataset they collected from UniProt. This model was able to achieve a better accuracy of 78% for ten categories of proteins and 92 % accuracy in identifying the membrane-bound / soluble. It outperformed the algorithms that relied on homology information.

Majid Ghorbani Eftekhar [2] tried to identify subcellular localization of protein by understanding the molecular function of protein. In this study, he found an inexpensive approach using a proposed deep learning analytical engine tool known as PSORTb v3.0. A new machine learning architecture was developed to replace the traditional SVM model using BiLSTM (State of the art deep learning method) and SeqGAN (Data Augmentation measure). This work was able to get a precision of 57.4% to 75 % on different datasets.

Long Pang, Junjie Wang, Lingling Zhao, Chunyu Wang and Hui Zhan [3] worked on the allocation of protein in the organelle or compartment that leads to various human diseases such as Alzheimer's disease. This paper suggests a novel method of incorporating the XGBoost (eXtreme Gradient Boosting) to automatically gain the features from dataset and recognize protein based on the outcome of CNN. The results surpassed the common Machine learning tools. This model works best on datasets related to Alzheimer's disease.

Rahul Semwal, Pritish Kumar Varadwaj [4] aimed to develop a tool that helps in annotating human protein subcellular locations. In this proposed model HumDLoc, they combined two widely used DL techniques the CELLO and the DeepLoc and some other ML techniques to give better accuracy. The output of this predictive model gives an 86 % precision and performs better than several other alternative tools in predicting human protein subcellular positions.

Zhen Cao, Xiaoyong Pan, Yang Yang, Yan Huang, HongBin Shen [5] report an classifier base prediction method called IncLocator for detecting IncRNA subcellular localizations. They implied four classifiers feeding with k-mer features and abstraction features respectively. The overall accuracy is 0.59 in predicting five subcellular localizations of IncRNA's along nucleus, exosome and cytoplasm.

Yu-hua Yao, Ya-ping Lv, Ling Li, Hui-min Xu, Bin-bin Ji, Jing Chen, Chun Li, and Bo Liao & Xu-ying Nan [6] has introduced two types of protein patterns encoding methods: Gapped k-mer and Dipeptide information with space. In this study, they predicted the protein component important for drug design and other biomedical applications. And achieved a method that reduces the dimension and also improved the precision.

4.PROPOSED WORK

According to the research, there have been a wide range of CNN models conceived, this work makes use of resent50 to get greater accuracy in detecting the protein's subcellular positions in microscopy images. Resnet50 is a one of CNN's architecture, a resnet version. The benefit of using this architecture is high accuracy is obtain using forty eight convolutional layers, one Max pool layer and one Average pool layer. It works best on high volume datasets like the Human protein atlas.

5. METHODOLOGY

The dataset for this project has been collected from kaggle and it has to undergo preprocessing to enhance the dataset for additional processes. The dataset will contain duplicate images and imbalanced resolutions in the images which must be rectified while pre-processing the dataset. The training data is used for working with resnet50 model and tested using the test dataset to obtain the higher accuracy rate.

5.1 Pre-Processing

Firstly, the data is collect from kaggle's human protein atlas dataset. Then the data has to preprocessed for improving the quality of the data for further procedures. The primary processing of dataset to make it suitable for further analysis is regarded as Data Preprocessing. Before preprocessing phase, the data will contain duplicate images due to the high volume of images present in dataset. After the preprocessing has been done, such images will be removed from the dataset for the forthcoming processes.

5.2 Data cleaning

In this phase, the processed dataset is converted to a format that is useful for working with this model. The removal and correction of mixed values is considered as Data cleaning. In here, the inconsistency in data distribution is cleaned, as the dataset contains an inconsistent distribution data due to huge collection of images. The images are resized to 512 x 521 resolution and made into batches for effective processing. In summary, the dataset is cleansed to suitable format for processing using CNN architecture.

6. EXPERIMENTAL RESULTS

We should analysis the data before stepping into results. The PNG data set was converted to 8 bit picture which was reduced to pixel rate of 512x512, meanwhile the TIFF data set was consider the original file which was in resolution of 2048x2048 pixel. This contains the protein pattern of infected person of different organelles which include 27 different sets of data. These all data were subjected to four filters(single file) this has protein with three cell markers(green) nuclei(blue),microtubules(red), and endoplasmic reticulum(yellow). The green filter is applied to the predictive label and the other filters are used as references. Among them, this study uses the matplotlib module in python to analyze the distribution of data sets.

From this we can say that common protein occupies the coarse cells, such as cell membranes, cytoplasm and nuclei. In our dataset we have some few components that looks similar as lipid droplets, peroxisomes, endosomes, lysosomes, microtubule ends, and rods and rings. So making classification on such data is different and is it quite difficult too because the information is based on certain category. There is chance of getting confusion in our main model, this makes the accuracy less in secondary categories. Therefore, in the data set, accuracy is not an appropriate measure of the overall performance, and there should be a better verification strategy.

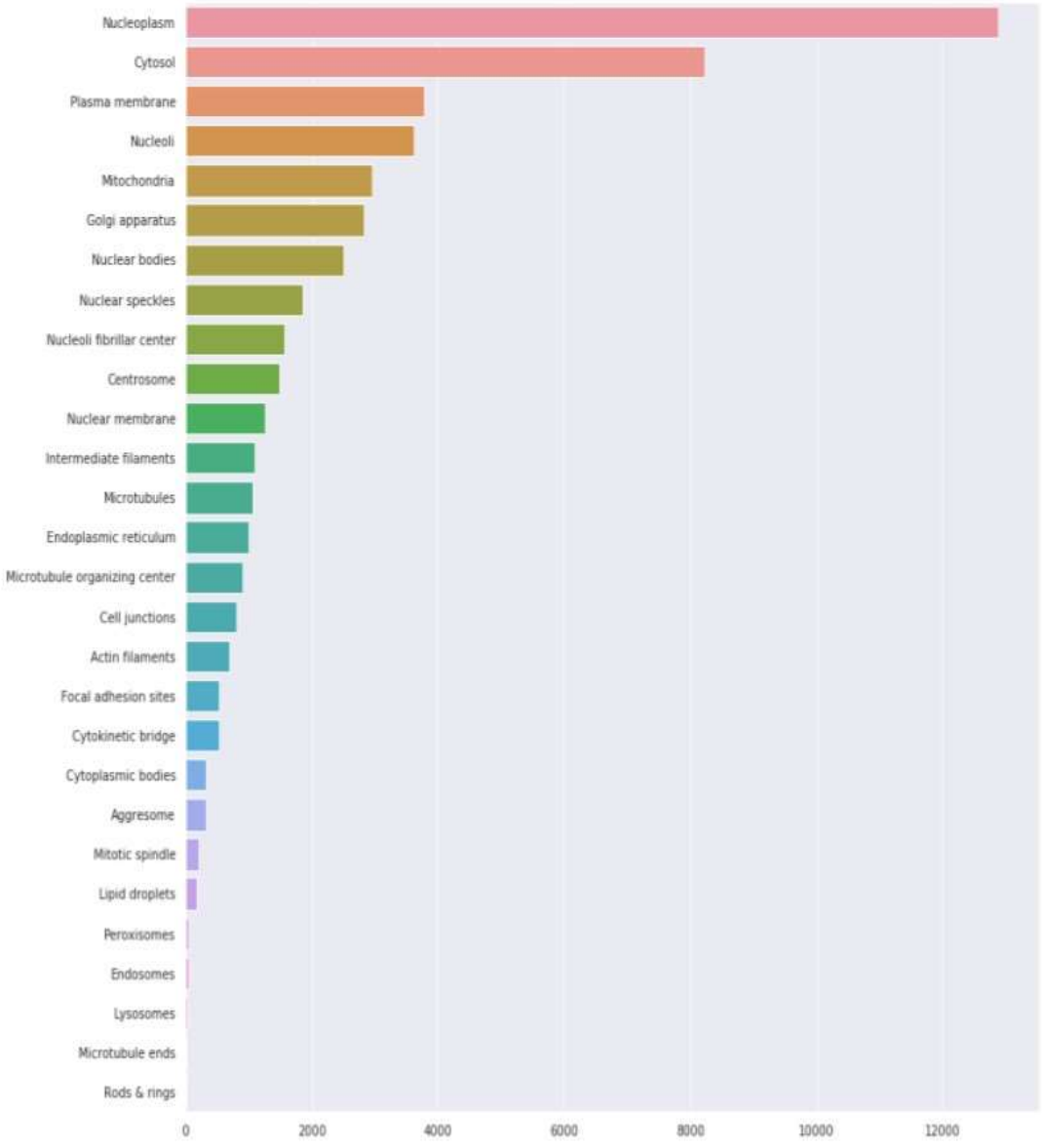
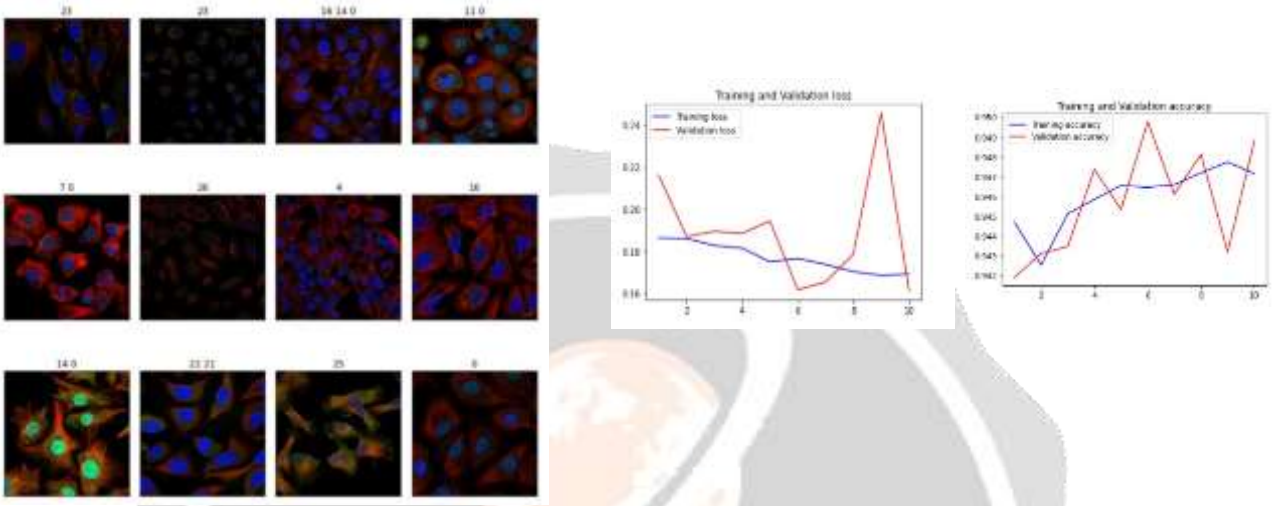


Fig-3: Filter protein

7.CONCLUSION

In this study, we had subject those material to preliminary analysis. We used ResNet mentioned above as the basline to do the training. The following are the result followed, In the training process, we can obtain a comparison of training loss and verification loss, which can be used to judge the degree of model coupling. In the training result 0.947 and validation accuracy 0.949 which is a very good training result.



8.ACKNOWLEDGMENT

It is an occasion for us to share our sincere gratitude, sincere regards, admiration, and commitments to our mentor Ms.Dhivya P, Assistant professor, Bannari Amman Institute of Technology, for her valuable encouragement, deeply seated interest, inspiration, and constant guidance during the project's period. She gave us the chance to try our hardest on such a fascinating subject, and her guidance and assistance were invaluable to this project. We had the lab space and applications for this project, thanks to Bannari Amman Institute of Technology.

9.REFERENCES

[1] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, Ole Winther, ”DeepLoc: prediction of protein subcellular localization using deep learning “, Bioinformatics, Volume 33, Issue 21, 01 November 2017, pp 3387–3395.

[2] Majid Ghorbani Eftekhar, ”Prediction of protein subcellular localization using deep learning and data augmentation”, bioRxiv, Version 3, June 15, 2020.

[3] Long Pang, Junjie Wang, Lingling Zhao, Chunyu Wang and Hui Zhan,
“A Novel Protein Subcellular Localization Method With CNNXGBoost Model for Alzheimer's Disease”, Front. Genet., 18 January 2019.

[4] Rahul Semwal , Pritish Kumar Varadwaj, “HumDLoc: Human Protein Subcellular Localization Prediction Using Deep Neural Network”, Current Genomics, Volume 21 , Issue 7 , 2020.

[5] Zhen Cao, Xiaoyong Pan, Yang Yang, Yan Huang, Hong-Bin Shen, “The lncLocator: a subcellular localization predictor for long noncoding RNAs based on a stacked ensemble classifier “,Bioinformatics, Volume 34, Issue 13, 01 July 2018, pp 2185–2194.

[6] Yu-hua Yao, Ya-ping Lv, Ling Li, Hui-min Xu, Bin-bin Ji, Jing Chen, Chun Li, Bo Liao & Xu-ying Nan, “Protein sequence information extraction and subcellular localization prediction with gapped k-Mer method”, BMC Bioinformatics volume 20, Article number: 719 (2019).