Improvised CNN model to Predict SARS by Detecting the Localisation of Proteins

M. Nirmala Devi
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
nirmaladevi2004@gmail.com

S. Mahima
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
mahimasoosairaj@gmail.com

R. Ramupriya
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
ramupriya069@gmail.com

Sumaya Abdul Sathar
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
sumayaasathar@gmail.com

2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-3524-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICECA52323.2021.9676045

Abstract— **Having extensively plenty of humans on earth, each one of us is made up of enormous number of cells. Every human being on this earth is different and unique, even in genetically identical twins, scientists are able to identify differences between the genetically identical cells in their bodies. Therefore, discrepancies in the localization of proteins can also lead to such cellular heterogeneity. Therefore, the identification of mis localized proteins may hint at cellular dysfunctions, advancing our knowledge about diseases. Proteins plays a highly important role in almost all the cellular processes in our body. Most frequently, many different proteins come together at a specific location to deploy a function, and the exact result of the task is based on the available proteins. It is important to know why and how proteins occur to completely understand how cells functions, how diseases develop and ultimately how to develop better treatment for those diseases. To comprehend the complexity of human cells, it is essential to segregate the mixed patterns across a wide range of human cells. Our primary aim is to predict localization of the protein organelle among 28 different labels, where each sample is being imparted by "The Human Protein Atlas" using high throughput microscopy. The "Human Protein Atlas" was initiated by Sweden to locate all the proteins in cells, tissues, etc. All the data in the knowledge resource has open access which allows anyone to explore human proteome. High throughput images from Human Protein Atlas provide a rich source of information on the protein location which can be utilized by computational methods. The proposed paper focuses on deep learning image analysis methods and, in particular, on Convolutional Neural Network which is effective at determining sub cellular locations. CNN outperforms all the computing methods and reaches close to perfect localization with accuracy of 96.6% significantly outperforming the best human expert with an accuracy of 72% and shows an improvement of 3% over the other existing models.**

Keywords— Protein Localization, Machine Learning, Deep learning, Image Classification, Convolutional Neural Networks, Feature Selection

I. INTRODUCTION

Proteins are "the doers" in human cells, performing diverse range of functions. The localization of proteins in a cell talks about the function it does to give rise to functional heterogeneity among cells, the identification of mis localized proteins may hint at cellular dysfunctions, advancing our knowledge about diseases [28]. Advancement in the high throughput microscopy has improved the creation of enormous amounts of biological images to form a dataset. The images are produced at a higher pace compared to what can be manually evaluated [9]. To further understand the functions of cell, the best way is to determine the distribution pattern of protein [13]. Therefore, the need would be, to automate biological image analysis that would improve the understanding of human protein cells and disease better. The Human Protein Atlas Image Classification focuses on training deep learning models to determine the different patterns of protein expression within the microscopic images of human cells given in the high throughput images [20], [22]. Image classification algorithms, powered by Deep Learning such as Convolutional Neural Networks and fully convolutional network fuels many industries scaling from plantation to healthcare [3]. The proposed paper represents Convolutional Neural Network model. Convolutional Neural Networks are state of the art for image analysis in a various field. Recently, CNNs have been tremendously successfully applied to analyze bio-logical and medical images, in tasks such as detection of melanoma, performing on par with dermatologists on bright field microscopy images. The Convolutional Neural Network model is based on the idea that the model function is based on recognizing of each image [1], [2]. CNN uses fewer parameters multiple times whereas fully convolution network uses more parameters [23]. While a fully connected network produces more weights than CNN as it produces weight for each image whereas in CNN, it produces enough weight for small area at the given time. However, data pre-processing and cell segmentation algorithms still plays a large role in high-throughput image dataset analysis [21].

II. LITERATURE SURVEY

Automated analysis of Human Protein Atlas immunofluorescence image

In this work, the authors provided an approach to automate the process of classifying sub cellular patterns in the images. Support vector classification framework has been used. They are effective at determining sub cellular locations. They completed their analysis by using around 3600 images [15]. They were able to obtain an accuracy of 87.5% for all samples and around 98% accuracy only for samples with high confidence. Long training time for datasets is the major limitation.

A Gene centric Human Protein Atlas for Expression Profiles based on Antibodies

With time, the human research on genetics is constantly evolving with the help of technology. It has been realised that different protein patterns may provide information on pivotal diagnostic which would be a breakthrough in the field of medicines. Thus, this paper provided a way in which proteins can be localised in cells, tissues and organs using a method based on immunohistochemistry and random forests. The search algorithm helps in resolving complicated queries on protein labels and localisation of chromosomes [16]. Since this methodology uses random forests, the larger the dataset becomes, greater would the space that it takes up. And to avoid the problem of overfitting hyper parameters must be tuned [17].

Predicting protein condensate formation using machine learning

Advancements in the machine learning classifier (PSAP) to forecast the candidate PPS proteins. PSAP led to the findings of new PPS proteins, including DAZAP1 and CPEB3. Immunofluorescence has constraints to fixed (i.e., dead) cells when structures present in the cell are to be visualised, as antibodies generally don't enter the cell membrane when reacting with fluorescent labels [19]. Antigenic material must be placed tightly on the site of its natural localisation inside the cell. As the entire prediction relies on the immunofluorescence, there may be few differences in the outcomes [18].

III. MATERIALS AND METHODS

Dataset

The "Human Protein Atlas" was initiated by Sweden to locate all the proteins in cells, tissues, etc. The data provided by them is freely accessible for everyone around the world. The dataset for our project was provided by Kaggle which consisted of around 31072 samples for training and 11,702 samples for testing. Each sample has four confocal images – green, red, blue and yellow. The green filter represents protein of interest which is used for localisation of proteins. The other red, blue and yellow filters represent microtubules, nucleus and endoplasmic reticulum respectively which are used for reference. The dataset provided is a multi-label dataset which is further converted into binary data frame, [16], [20].

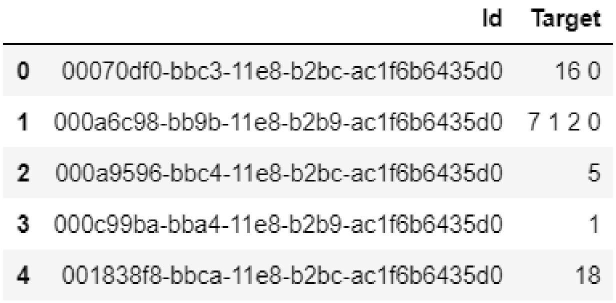
The dataset used for our project has 28 labels which are different organelles present in the cell. The labels are mapped to integers which is given in Table 1.

Table 1. Labels and their mapped integers.

Integer Mapped	**Label**
0	Nucleoplasm
1	Nuclear membrane
2	Nucleoli
3	Nucleoli fibrillar center
4	Nuclear speckles
5	Nuclear bodies
6	Endoplasmic reticulum
7	Golgi apparatus
8	Peroxisomes
9	Endosomes
10	Lysosomes
11	Intermediate filaments
12	Actin filaments
13	Focal adhesion sites
14	Microtubules
15	Microtubule ends
16	Cytokinetic bridge
17	Mitotic spindle
18	Microtubule organising center
19	Centrosome
20	Lipid droplets
21	Plasma membrane
22	Cell junctions
23	Mitochondria
24	Aggresome
25	Cytosol
26	Cytoplasmic bodies
27	Rods & rings

The CVS file for training consists of two columns:

1. Id – Represents id of the image
2. Target – Represents label(s) assigned to each sample



	Id	Target
0	00070df0-bbc3-11e8-b2bc-ac1f6b6435d0	16 0
1	000a6c98-bb9b-11e8-b2b9-ac1f6b6435d0	7 1 2 0
2	000a9596-bbc4-11e8-b2bc-ac1f6b6435d0	5
3	000c99ba-bba4-11e8-b2b9-ac1f6b6435d0	1
4	001838f8-bbca-11e8-b2bc-ac1f6b6435d0	18

Fig. 1. The top 5 rows of the training dataset which has two columns – Id and Target.

It can be seen from Fig. 2. that our methodology starts by encoding the multi-label list into binary target labels. This makes it easier to study the dataset and hence an exploratory data analysis is conducted to thoroughly understand the data [4]. After which a simple image preprocessing is done which includes resizing, reshaping and normalizing. It is always better to feed our data into a simple baseline model to know how it works [21]. Hence it is fed into a baseline CNN model and its performance is analyzed. The baseline model is further improved by changing the architecture and the performance of model is compared with the existing ones to know how much better it works [22].

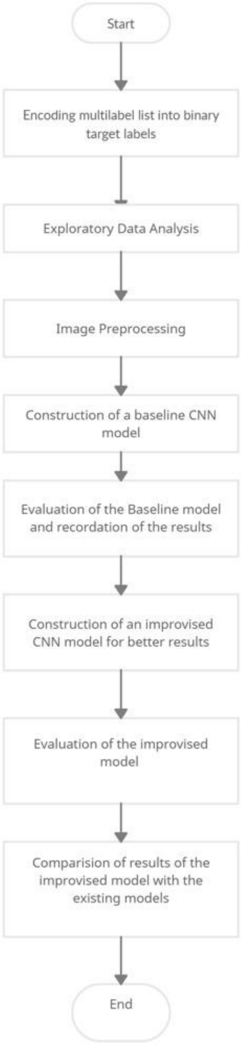

Start
Encoding multilabel list into binary target labels
Exploratory Data Analysis
Image Preprocessing
Construction of a baseline CNN model
Evaluation of the Baseline model and recordation of the results
Construction of an improvised CNN model for better results
Evaluation of the improvised model
Comparision of results of the improvised model with the existing models
End

Fig. 2. Flowchart of methodology

Exploratory Data Analysis

Exploratory Data Analysis is the first step in our methodology. It refers to the process of taking investigating measures in order to know more about the data that we are dealing with. It can be used to know about the kind of data that we are going to be dealing with, to find out if there are any missing values and how they can be replaced it with the accurate data. With EDA, we can identify the features that can be included or excluded to get the best out of the data.

Occurrence of protein in organelles

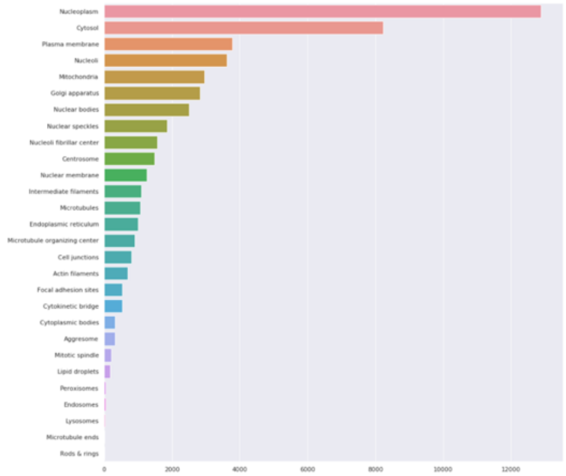


Fig. 3. Class Imbalance

From Fig. 3. we can see that most of the protein structures are present in cellular components like nucleoplasm, cytosol and plasma membrane. And components like peroxisomes, endosomes, lysosomes are quite rare in our dataset. This creates an imbalance in our dataset and prediction for these labels will be difficult due to fewer examples.

Number of targets per image

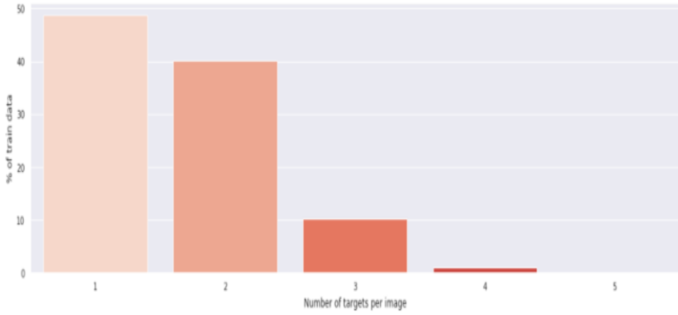


Fig. 4. Number of targets per image

From Fig. 4., it is observed that in the training data the number of targets per image is mostly one or two and above three is seldom.

Correlation between targets

It can be observed that most of the targets have only slight correlation between them. But it is noticeable that endosomes mostly occur together with lysosomes and mitotic spindle occurs with cytokinetic bridge. Hence, we can find a positive correlation among these organelles [30].

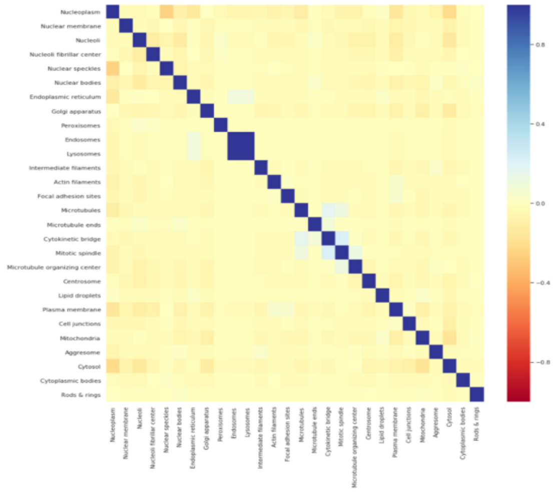


Fig. 5. Correlation Matrix

Visualization of images in dataset

Each sample has four filters – green, blue, red and yellow. Hence, we will find four images for a single sample. The different filters represent different components.

- *Green* represents target protein of interest
- *Blue* represents nucleus
- *Red* represents microtubules
- *Yellow* represents endoplasmic reticulum

In the below figure, we have taken lysosomes and endosomes as the protein of interest.

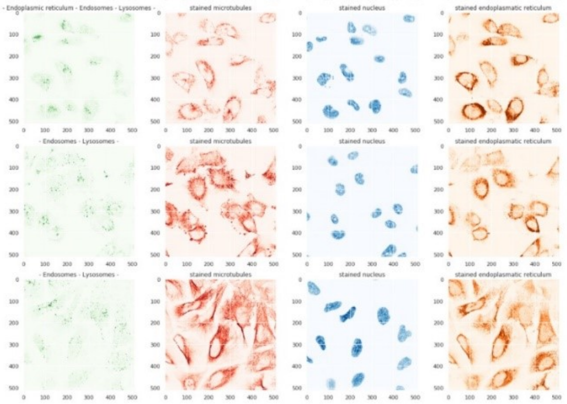

- Endoplasmic reticulum - Endosomes - Lysosomes -
stained microtubules
stained nucleus
stained endoplasmatic reticulum
- Endosomes - Lysosomes -
- Endosomes - Lysosomes -
0
100
200
300
400
500

Fig. 6. Visualization of images

We can infer that the staining in green channel is not successful as the images differ in their intensities. From the red channel, it can be noticed that there are cells of different types.

Building CNN Model

K-Fold Cross Validation

For the purpose of evaluating our model, we use k-fold cross validation. The “k” refers to the number of segments the data is split into. The procedure of k-fold Cross Validation is as follows:

1. Mix up the dataset randomly.
2. Split it into k segments.
3. For each segment:
   i. Take one segment as test dataset.
   ii. Consider the rest as train dataset.
   iii. Fit the model on training data and evaluate it on test set.
   iv. Keep the evaluation score and drop the model.
4. Sum up the performance of the model using the evaluation scores.

Since our dataset has class imbalance, we might get low evaluation score if the chunk chosen contains seldom labels or high evaluation score if the chunk contains common targets, [25]. To eradicate this issue, the k-fold needs to repeated several times.

It is found that the test dataset is 38% of the train dataset. Hence, we can use 3-fold cross validation where test dataset will be 33% of the train dataset.

Baseline CNN Model

Our first step is to setup a baseline CNN model. This model does not have to be complex or it does not have to provide a very good accuracy or score. Our main objective is to start with a simple model and then improve that model by changing the CNN architecture.

After a detailed exploratory data analysis, we could come to a conclusion that green channel images could be used for prediction and others for reference [29]. Even though we are just using green channel images, our dataset is huge which makes our task complex. Hence, we could use a data generator with the help of keras. This helps in loading the data in batches. But before that it is important to preprocess our images to remove unwanted distortions and enhance features which would help our prediction.

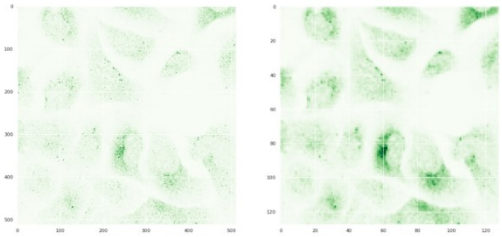


Fig. 7. Before and after image pre-processing

In image pre-processing, we have resized, reshaped and normalized which can be seen in fig. 7. With the data generator and pre-processed images ready, the next important step is to setup a CNN model and fit the training dataset. Since we are going to build a baseline CNN model, the number of layers is going to be limited.

The CNN architecture uses ReLu activation function in the first two convolutional layer and each of them is followed by a MaxPool layer [5],. The resultant output is flattened and then given as input to the fully connected layer which uses ReLu initially and then it uses Sigmoid activation function after dropout [24]. After training the dataset and testing them, we got an accuracy around 94% which seems to be a good number but in fact this model identifies the absence of proteins more than the presence of proteins. Hence this calls for an improvement in the model, for which we will design a complex CNN architecture by increasing the number of layers and continued epochs may also help us increase our accuracy and with the prediction.

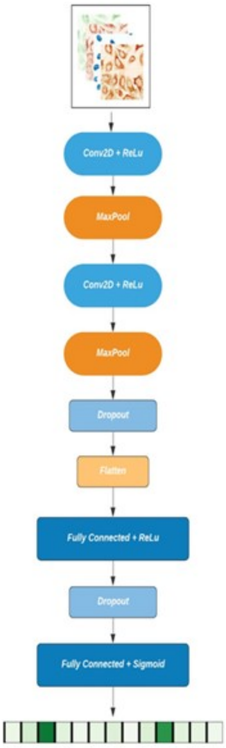


Fig. 8. CNN Architecture

Improvised CNN Model

The new improvised model has many layers for the purpose of the feature extraction. The new CNN architecture (as shown in fig. 9.) used ReLu as activation function for almost all Conv2D layer [12]. After that, batch normalization layers were also added. The final model was trained for 100 epochs. The parameters and batch size were compromised due to GPU since it needs to be handled at once in a single go. As the epochs increased the validation loss decreased and the accuracy kept increasing up-to 96.6% and even more.

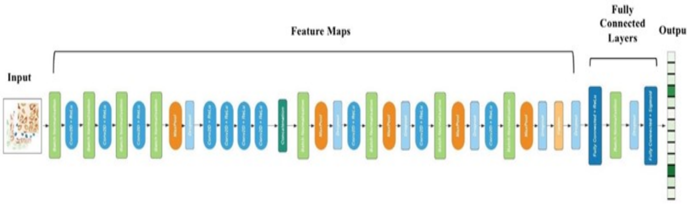


Fig. 9. Improvised CNN Model

Performance metrics for model evaluation

Once the model is trained, evaluation is done by the below mentioned metrics:

***i.* Accuracy**

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

ii. Precision

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

iii. Recall (Sensitivity)

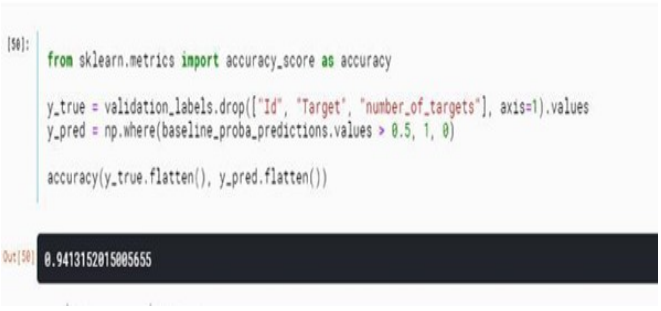
$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

iv. F1 score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

IV. RESULTS

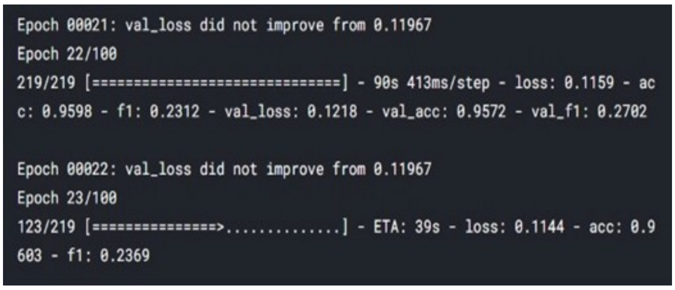
This chapter gives the performance analysis of the baseline CNN model and the improvised Model. The analysis also serves as a comparison based on the accuracy and performance between the baseline CNN model and the improvised model.



```
[50]:  from sklearn.metrics import accuracy_score as accuracy

       y_true = validation_labels.drop(["Id", "Target", "number_of_targets"], axis=1).values
       y_pred = np.where(baseline_proba_predictions.values > 0.5, 1, 0)

       accuracy(y_true.flatten(), y_pred.flatten())

Out[50]: 0.9413152015005655
```

Fig. 10. Accuracy of Baseline Model

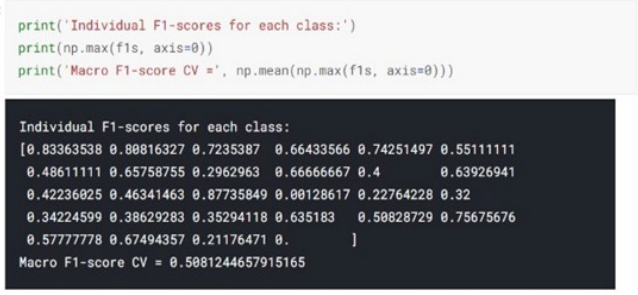
The baseline CNN model provides an accuracy of 94.13%.



```
Epoch 00021: val_loss did not improve from 0.11967
Epoch 22/100
219/219 [==============================] - 90s 413ms/step - loss: 0.1159 - ac
c: 0.9598 - f1: 0.2312 - val_loss: 0.1218 - val_acc: 0.9572 - val_f1: 0.2702

Epoch 00022: val_loss did not improve from 0.11967
Epoch 23/100
123/219 [===============>..............] - ETA: 39s - loss: 0.1144 - acc: 0.9
603 - f1: 0.2369
```

Fig. 11. Accuracy of Improvised Model

The improvised model provides an accuracy of 96.6%.



```
print('Individual F1-scores for each class:')
print(np.max(f1s, axis=0))
print('Macro F1-score CV =', np.mean(np.max(f1s, axis=0)))
```

```
Individual F1-scores for each class:
[0.83363538 0.80816327 0.7235387  0.66433566 0.74251497 0.55111111
 0.48611111 0.65758755 0.2962963  0.66666667 0.4        0.63926941
 0.42236025 0.46341463 0.87735849 0.00128617 0.22764228 0.32
 0.34224599 0.38629283 0.35294118 0.635183   0.50828729 0.75675676
 0.57777778 0.67494357 0.21176471 0.        ]
Macro F1-score CV = 0.5081244657915165
```

Fig. 12. Macro F1-score

Thus, on the comparison of the results between the baseline CNN model and the improvised CNN model, it is clearly visible that the improvised CNN model has an enhanced performance with an accuracy of 96.6%.

V. CONCLUSION

Technology plays a prime role answering to the novel coronavirus (SARS-CoV-2) and the COVID-19 pandemic. The transmissibility of the virus in the era of covid'19 has challenged medical officials and technologists, and exhibited constraints to the conventional health industry. Nevertheless, in the entire pandemic epoch, technology has acknowledged the redesigned public health which offers opportunities for better agility, scale, and responsiveness. Protein being the building blocks of tissues, is one of the important macronutrients that is required to build our immune system. A weaker immune system leads to higher risks of getting infected by COVID-19. From the existing facts it is proven that protein along with other nutrients are essential for effectively tackling the risk and severity of the disease. Therefore, we have to know the parts of the body where we have sufficient protein content, and it is also important to know the parts in which the proteins are absent so as to develop the content of proteins by the intake of suitable food that boosts the level of protein in our body. The severity of the disease and the importance of the protein in the disease led us to develop a model that helps in detecting the localization of proteins. The baseline CNN model that we constructed for the localization of protein produced an accuracy of 94%. With great efforts we tried maximizing the accuracy of the model. We built an improvised CNN model that produced enhanced results of 96.6 %. Technology cant completely keep us away from the ongoing pandemic disease, However, it can help us out in handling the catastrophe more constructively than the existing. All of us are aware that COVID-19 has severely affected our personal and professional living. In this era of unpredictable uncertainty and agitation, our preparedness to opt in technology will be the foremost option.

VI. FUTURE WORK

The rapid increase in the number of COVID-19 cases, and the victims in need of medical emergency highly pressurises the technologists and the medical professionals across the world. The blended efforts of the medical officials and the techno-developers will bring in sustainable and better results to overcome the pandemic. With all the efforts we had put in we were able to develop a model with 96.6% accuracy. But when we are dealing with human lives, even 1% lesser than 100 would cause a tremendous difference. Therefore, we are further working on provided the fullest accuracy of 100% by enhancing the features of the model that we had built. All that is mainly needed to fight a major disease like COVID is preparedness. The progress in technology is advancing like never before; and also, it will undoubtedly continue to expand more. And we humans are responsible for adapting the changes in the advancing technology quickly and also continuously invest in constructing the technology systems for improved readiness and enhanced results.

REFERENCES

[1] N. Wang, Y. Zhang and L. Zhang, "Dynamic Selection Network for Image Inpainting," in IEEE Transactions on Image Processing, vol. 30, pp. 1784-1798, 2021, doi: 10.1109/TIP.2020.3048629.

[2] J. Yan, M. Jin, Z. Xu, L. Chen, Z. Zhu and H. Zhang, "Recognition of Suspension Liquid Based on Speckle Patterns Using Deep Learning," in IEEE Photonics Journal, vol. 13, no. 1, pp. 1-7, Feb. 2021, Art no. 6800207, doi: 10.1109/JPHOT.2020.3044912.

[3] W. Wang, F. Bu, Z. Lin and S. Zhai, "Learning Methods of Convolutional Neural Network Combined With Image Feature Extraction in Brain Tumor Detection," in IEEE Access, vol. 8, pp. 152659-152668, 2020, doi: 10.1109/ACCESS.2020.3016282.

[4] P. Ribalta Lorenzo, L. Tulczyjew, M. Marcinkiewicz and J. Nalepa, "Hyperspectral Band Selection Using Attention-Based Convolutional Neural Networks," in IEEE Access, vol. 8, pp. 42384-42403, 2020, doi: 10.1109/ACCESS.2020.2977454.

[5] N. R. C. Monteiro, B. Ribeiro and J. Arrais, "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2020.2977335.

[6] G. He, J. Ji, H. Zhang, Y. Xu and J. Fan, "Feature Selection-Based Hierarchical Deep Network for Image Classification," in IEEE Access, vol. 8, pp. 15436-15447, 2020, doi: 10.1109/ACCESS.2020.2966651.

[7] Y. Wang et al., "Breast Cancer Image Classification via Multi-Network Features and Dual-Network Orthogonal Low-Rank Learning," in IEEE Access, vol. 8, pp. 27779-27792, 2020, doi: 10.1109/ACCESS.2020.2964276.

[8] Z. Zheng and J. Cao, "Fusion High-and-Low-Level Features via Ridgelet and Convolutional Neural Networks for Very High-Resolution Remote Sensing Imagery Classification," in IEEE Access, vol. 7, pp. 118472-118483, 2019, doi: 10.1109/ACCESS.2019.2936295.

[9] C. Le and X. Li, "JigsawNet: Shredded Image Reassembly Using Convolutional Neural Network and Loop-Based Composition," in IEEE Transactions on Image Processing, vol. 28, no. 8, pp. 4000-4015, Aug. 2019, doi: 10.1109/TIP.2019.2903298.

[10] H. Dong, W. Ma, Y. Wu, M. Gong and L. Jiao, "Local Descriptor Learning for Change Detection in Synthetic Aperture Radar Images via Convolutional Neural Networks," in IEEE Access, vol. 7, pp. 15389-15403, 2019, doi: 10.1109/ACCESS.2018.2889326.

[11] L. Wu, J. Cheng, S. Li, B. Lei, T. Wang and D. Ni, "FUIQA: Fetal Ultrasound Image Quality Assessment With Deep Convolutional Networks," in IEEE Transactions on Cybernetics, vol. 47, no. 5, pp. 1336-1349, May 2017, doi: 10.1109/TCYB.2017.2671898.

[12] W. Shao, M. Liu, Y. Xu, H. Shen and D. Zhang, "An Organelle Correlation-Guided Feature Selection Approach for Classifying Multi-Label Subcellular Bio-Images," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 828-838, 1 May-June 2018, doi: 10.1109/TCBB.2017.2677907.

[13] P. Wang, Y. Chen, J. Lü, Q. Wang and X. Yu, "Graphical Features of Functional Genes in Human Protein Interaction Network," in IEEE Transactions on Biomedical Circuits and Systems, vol. 10, no. 3, pp. 707-720, June 2016, doi: 10.1109/TBCAS.2015.2487299.

[14] M. Pop et al., "High-Resolution 3-D T${{\bf _1}^{\bf *}}$-Mapping and Quantitative Image Analysis of GRAY ZONE in Chronic Fibrosis," in IEEE Transactions on Biomedical Engineering, vol. 61, no. 12, pp. 2930-2938, Dec. 2014, doi: 10.1109/TBME.2014.2336593.

[15] Charvi Wadhwa; P. Prabu, "An empirical analysis of ICT tools with gamification for the Indian school education system", International Journal of Enterprise Network Management, vol .12, pp. 258-274, 2021.

[16] K. Amrutha, P. Prabu, "ML Based Sign Language Recognition System", 2021 International Conference on Innovative Trends in Information Technology (ICITIIT), pp. 1-6, 2021.

[17] Soumi De, P. Prabu, Joy Paulose, "Effective ML Techniques to Predict Customer Churn", IEEE Xplore 2021 Third International Conference on Inventive Research in Computing Applications, pp. 895-902, 2021.

[18] K.S.Bhuvansehwari, et al. "Improved Dragonfly Optimizer for Instrusion Detection Using Deep Clustering CNN-PSO Classifier.", CMC-Computers, Materials & Continua, vol.70, pp.5949-5965, 2021.

[19] Mohd Anul Haq, Abdul Khadar Jilani and P. Prabu, "Deep Learning Based Modeling of Groundwater Storage Change.", CMC-Computers, Materials & Continua, vol .70, pp. 4599-4617, 2021.

[20] Ankur Rameshbhai Khunt and P. Prabu, "An Empirical Analysis of Android Permission System Based on User Activities.", Journal of Computer Science, vol .14, pp. 324-333, 2018.

[21] Megana Ghosh and P. Prabu, "Empirical analysis of ensemble methods for the classification of robocalls in telecommunications.", International Journal of Electrical and Computer Engineering, vol .9, pp. 3108-314, 2019.

[22] Jay Kotecha and P. Prabu, "An Investigation on android background services for controlling the unauthorized accesses using android LOG system.", International Journal of Engineering & Technology(UAE), vol .7, pp. 301-305, 2018.