CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Bio-Image Based Prediction of Protein Subcellular Localization Using Adaptive Threshold

by

Hiba Khurshid

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2022

Copyright © 2022 by Hiba Khurshid

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

I dedicate this work to my parents, teachers and my family

CERTIFICATE OF APPROVAL

Bio-Image Based Prediction of Protein Subcellular Localization Using Adaptive Threshold

by

Hiba Khurshid

(MCS173052)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Muhammad Nazir	HITEC, Taxila
(b)	Internal Examiner	Dr. Mohammad Masroor Ahmed	CUST, Islamabad
(c)	Supervisor	Dr. Abdul Basit Siddiqui	CUST, Islamabad

Dr. Abdul Basit Siddiqui
Thesis Supervisor
January, 2022

Dr. Nayyer Masood
Head
Dept. of Computer Science
January, 2022

Dr. M. Abdul Qadir
Dean
Faculty of Computing
January, 2022

Author's Declaration

I, **Hiba Khurshid** hereby state that my MS thesis titled "**Bio-Image Based Prediction of Protein Subcellular Localization Using Adaptive Threshold** " is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Hiba Khurshid)

Registration No: MCS173052

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Bio-Image Based Prediction of Protein Subcellular Localization Using Adaptive Threshold**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Hiba Khurshid)

Registration No: MCS173052

Acknowledgement

All praise and worship belongs to Almighty Allah, who gave me the strength to complete this work.

I want to thank my family specially my parents who encouraged me in every decision of my life. Without them I would not be able to achieve anything. I also want to thank my teachers who supported me and shared their knowledge with me.

Last but not the least, a special thanks to my supervisor Dr. Abdul Basit Siddiqui who helped me throughout this work and guided me in every way possible.

(Hiba Khurshid)

Abstract

Over the past few decades, with the increase in microscopic imaging a very rapid progress has been made in predicting Protein subcellular localization. Knowledge of Protein subcellular localization is very important in understanding the function of protein. During the drug discovery, it can significantly improve the target identification process. Protein subcellular localization is also very important in disease discovery. It has been prove that abnormal protein subcellular localization causes diseases and can even involve cancer. Many researcher has come up with different model to predict Protein Subcellular Localization. With the advancement in deep learning models different architecture of CNN has been vastly used for classification of protein. But CNN comes with the computational cost and other drawbacks. In order to predict protein subcellular localization with high accuracy this study propose a methodology that uses Otsu's adaptive thresholding technique which calculate threshold value for each image by using image histogram. With this threshold value this methodology generates three binary images and for each binary image it extracts 9 feature vectors by counting the number of white pixel in neighboring pixel. In training phase Multi Label Random Forest classifier is applied on the extracted features to predict Protein subcellular localization. In order to evaluate this proposed technique this study used recently publish data HPA (version 18) and outperformed the state-of-the-art technique (macro f1 –score 0.59) by achieving macro f1-score of 0.63. This study also evaluated this technique against the fixed threshold and achieve macro f1-score of 0.44 which proves that adaptive threshold achieve more accuracy then fixed threshold technique.

Contents

List of Figures

List of Tables

Abbreviations

IHC	Immunohistochemistry
PSL	Protein Subcellular localization
SVM	Support Vector Machine
TAS	Threshold Adjacency Statistics

Chapter 1

Introduction

1.1 Biological Background

With the improvement in microscope, many scientist contributed in cell theory. The doctrine of cell states that cells are the smallest independent unit of life and all eukaryotic organisms are made of these cells. Cell theory played a very vital role in shaping the biological science. [1].

Cell is the basic building block of human body and a fundamental component of modern definition on life and living things. There are more than 10 trillion cells in human body that are highly diverse in function and their structures and play a very vital role in the development of human body. Together they work and perform function. According to many molecular biology and histology textbooks, in an adult human body there are around 200 types of cells [2]. Cells can be classified into different categories based on different factors like shape, size, complexity and their numbers. Based on the complexity cells are divided into two categories, Eukaryotic cell and Prokaryotic cell. The main difference between eukaryotic cell and prokaryotic cell is that all Eukaryotic cells have a membrane –bounded nucleus, cytoskeleton and a complex endomembrane system whereas Prokaryotic cells lack a cell nucleus or any membrane-encased organelles [3].

1.1.1 Eukaryotic Cell

Any organism or cell that contains a clear defined nucleus is Eukaryotic cell. The Eukaryotic cell's nucleus is surrounded by the nuclear membrane where chromosomes (bodies containing the hereditary material) are residing. Eukaryotic cells also contain organelles, including mitochondria (cellular energy exchangers), a Golgi apparatus (secretory device), an endoplasmic reticulum (a canal-like system of membranes within the cell), lysosomes (digestive apparatus within many cell types) and ribosomes. Figure 1.1 shows the anatomy of and Eukaryotic cell.

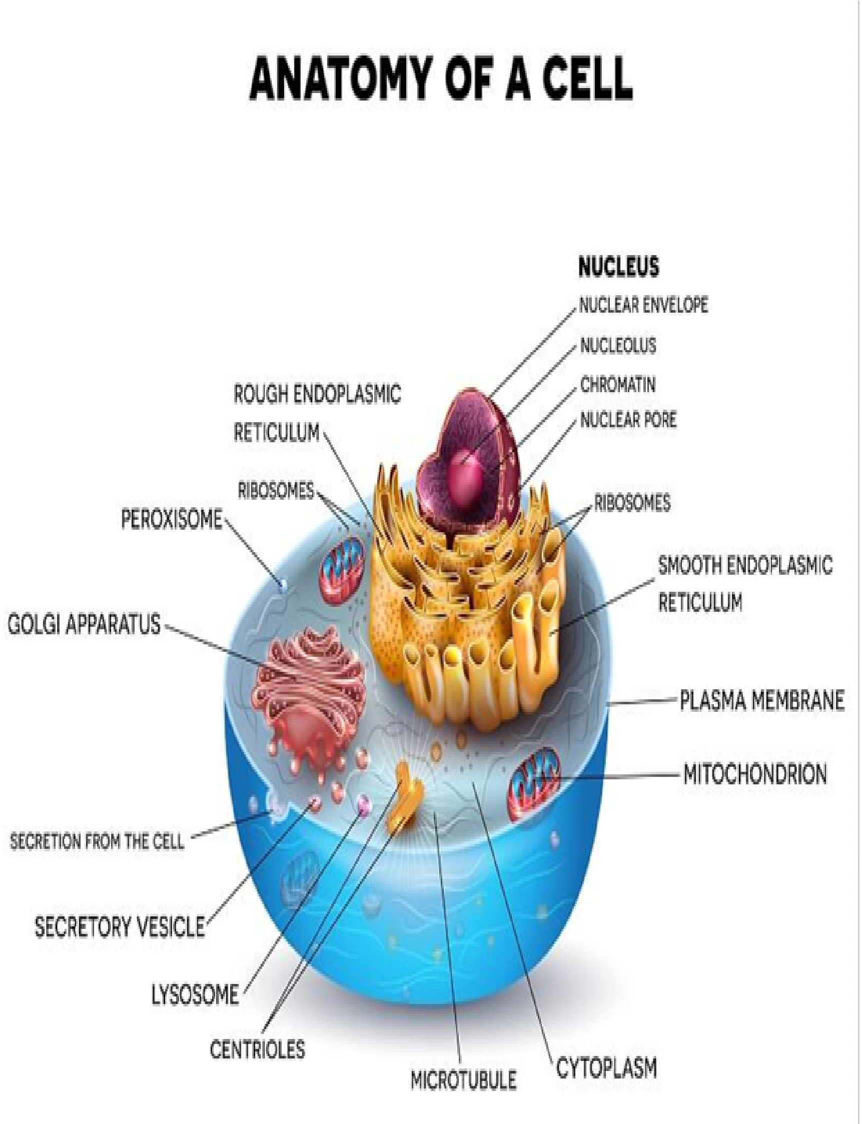


FIGURE 1.1: Anatomy of Eukaryotic cell.

1.1.2 Organelles and their Functions

As mentioned Eukaryotic cell contains different organelles which has at least one explicit tasks to act in the cell, similar as an organ does in the body. Table 1.1 shows some functionality of different organelles. For instance the purpose of Nucleus is to controls the cell activities. Mitochondria is considered as the power house of cell, Lysosome Digests food, bacteria, worn out organelle, Golgi Apparatus Sorts and packs protein into vesicle and transports them and Ribosomes are responsible for making of Proteins.

Table 1.1: Organelles and their Functions.

Name	**Prokaryotic / Eukaryotic**	**Function**
Nucleus	E, P	Controls the cell activities.
Nucleolus	E, P	Assembly of ribosomes take place here.
Cell Membrane	E, P	• Separates the cell from outside environment • Controls what goes in and out of cell.
Lysosome	E	Digests food, bacteria, worn out organelle.
Mitochondria	E, P	Power house of cell –produces energy for growth, development, and movement
Chloroplast	E, P	• Captures light & converts it into chemical energy. • Pigment chlorophyll (photosynthesis).
Golgi Apparatus	E, P	Sorts & packs protein into vesicle & transports them.
Cytoplasm	E, P	Gel-like substance that keeps organelles in place.
Vacuole	E, P	Stores food, water and other material.
Ribosome	E, P	Makes Proteins
Endoplasmic Reticulum	E, P	• Connects membrane • Moves material • Process protein

1.1.3 Ribosomes

Ribosomes, in eukaryotes, uses a process called translocation for protein synthesis by following the order from the nucleus. As shown in figure 1.2 In Translocation process nucleus has some parts of DNA (genes) that are transcribed to form mRNAs, messenger RNAs. These mRNA transfer to the ribosome, where ribosome uses this information to make a protein which has a specific amino acid sequence. As protein synthesis is an important function in any cell, therefore ribosomes are present in every cell type not only in multicellular organisms but also in prokaryotic cells as well such as bacteria. However the ribosome in eukaryotic cell is in large number than in prokaryotic cell.

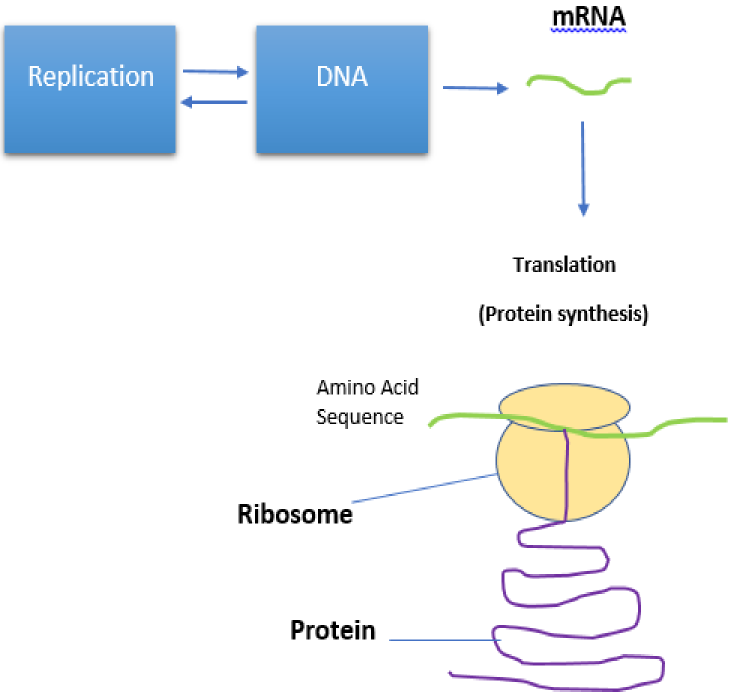


Figure 1.2: Translocation Process

1.1.4 Proteins

As mentioned above proteins are macromolecules (large size molecules) having structural unit called amino acid. There are total of 20 different amino acid that exist in protein and thousands of these amino acid attach together to form a sequence of long chains to form protein. In Eukaryotic cells ribosomes can synthesize many different sequence of amino acid proteins which then travel to their destination and according to the defined destination they perform their functions. In order to perform its function, it is very important to transport a protein to its designated organelle or subcellular location. This process of transferring a newly synthesized protein to its destination is known as Protein Sorting or Protein targeting [4]

1.2 Method of Predicting Protein Subcellular Localization

Method of predicting protein subcellular localization can be classified into two categories, through 1D amino acid sequencing and 2D-bio image based.

1.2.1 1D- Amino Acid Sequencing

In the early years many work has been done for this particular problem using amino acid sequence.Tthis is the process to identify the arrangements of amino acid in protein. Generally this can be done by two ways, one is similarity based classification and other is to find target signals that are buried into the sequence. But the main problem with the amino acid sequence is that they cannot detect protein miss translocation which can be a cause for cancer and plays very important role in cancer biomarker screening. [2]. Other drawbacks of amino acid sequence is that they are not very intuitive as they are text based and they carry less information as compare to the 2D protein image, as images are very intuitive, unambiguous, more concise, interpretable and carry more morphological details about protein subcellular location, they help in clear visualization of the size and shape the protein which can be very helpful as different proteins have different size and shapes [5].

1.2.2 Image Based Prediction

Images are very important for human experts as they can easily extract useful information when distinguishing protein subcellular locations manually. Inspired by this many automated algorithm predicts protein subcellular locations with the help of different distribution patterns which can be very useful in detection of translocation of cancer tissues. Due to the above mentioned advantages image based classification are becoming very popular among researchers.

In the recent years with the advancements in microscopy images, a large amount of images have been produced every day making it impossible for human expert to analyze each image. Day by day large dataset of images has been produce. Due to this reason researchers are finding an optimized automated method to classify large image set within less time. Machine Learning is becoming very popular for this problem among researchers. Given image dataset, features are extracted that helps in classify protein subcellular localization and different classifiers are implemented to achieve the highest accuracy. In the past years Machine Learning is used vastly in classification problems. Researchers are also using machine learning for this particular problem using different dataset, feature extraction methods and different classifiers. . Deep learning is also being used recently for this purpose especially when dataset is in image form and they are achieving good accuracy. Deep learning algorithms are capable of extracting images by themselves Different authors proposed different architecture and pre-trained model with some modification in order to achieve efficient accuracy.

1.3 Description of Subcellular Localization Distribution

In order to predict protein subcellular localization,one of the challanges is to extract useful features among many potential features [6]. It is essential to extract features that can view the texture pattern in image clearly. Below is the summary of different Global and Local feature extraction methods used by different researchers in order to predict Protein subcellular Localization.

1.3.1 Global Features

Below are some used global features descriptors used in protein subcellular localization.

1.3.1.1 Haralick Feature

In pattern recognition systems, Haralick feature is the renowned image descriptor. Along with the entropy, correlation and contrast the Haralick feature uses 13 different statistics, calculated by using image's grey-level co-occurrence matrix. In order to understand protein pattern Haralick features are made rotation invariant by taking averaged overall directions of co-occurrence. [7].

1.3.1.2 Zernike Feature

Zernike features are also well known feature descriptor when it comes to protein prediction. It performs much more effective when combined with different features descriptors like Haralick. In two polar variables Zernike calculates the image decomposition onto an orthogonal set of polynomials. [8].

1.3.1.3 Threshold Adjacency Statistics (TAS)

Studies has shown that only using Threshold Adjacency Statistics TAS, recognition system shows a good performance which indicates that it is a good feature descriptor in bio-images. TAS sets a fixed threshold value and using that value it extracts the features which perform very well in predicting protein subcellular localization. [9].

1.3.1.4 Histogram of Oriented Gradient (HOG)

For protein subcellular localization Histogram of Oriented Gradient HOG descriptor construct a feature vector by combining the values obtained by counting all the edge orientation of each cell in histogram. In case of object variation such as rotation, scale and translation, Histogram of Oriented Gradient HOG descriptor is very efficient. [10].

1.3.1.5 Texton Based Statistical Feature

Texton Based features as the name implies detect the textural information. A mask which has the ability to detect the textural information is skimmed over the entire image to produce texton image which then further computed to extract statistical features of this texton image which include energy contrast, homogeneity and entropy [11].

1.3.2 Local Features

1.3.2.1 Scale-Invariant Feature Transform

These features are obtained by detecting salient points and describing local features around these salient points in an image. It performs best on fluorescence object because of they are invariant to orientation and scaling and partly invariant to illumination changes. [12].

1.3.2.2 Speed-up Robust Feature (SURF)

SURF is composed of two-pass algorithm. In the first pass using an approximate Gaussian blob detector, it detects the interest points and in the second pass of the algorithm it calculates 64 statistical features at each interest point. As SURF is inspired by SWIFT it has been used in predicting protein subcellular localization. [13].

1.3.2.3 Subcellular Object Feature:

These features are obtained by number of pixels, fraction of overlapping pixels with the DNA and the object skeletons length in fluorescence object as they are developed to describe the pattern of fluorescence object. [14].

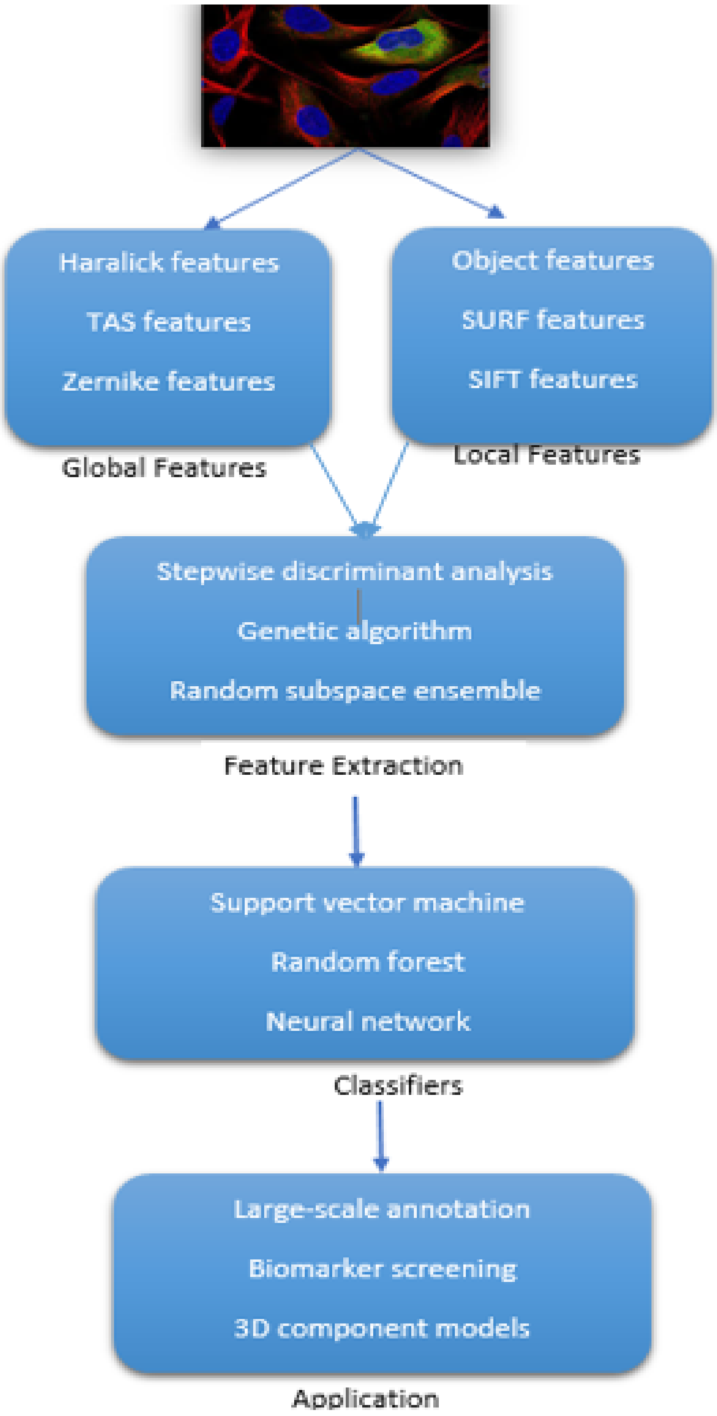


FIGURE 1.3: Process for predicting protein subcellular localization

1.4 Application of Protein subcellular localization

In molecular cell biology, proteomics and system biology, protein subcellular localization is very important. Studies shows that subcellular locations plays an important role in defining protein function. To work properly, protein has to be located in its proper place after being synthesized in the ribosome Therefore, it is very important to find the subcellular localization of a protein to understand its working. It has been prove that abnormal location of protein in subcellular compartment causes potential human diseases and even involve in cancer [15]. Protein subcellular localization can be used to make 3D model of a cell organelle that can be very helpful in medical studies and diagnosis [2].
With recent breakthrough in bioinformatics, bio images has been used to solve many complicated biological research problems, protein subcellular localization being one of the example. With the advancement in sequencing and imaging technologies, many methods has been proposed to predict accurate subcellular localization during the last decade and up till now it is one of the important task in bioinformatics and researchers are contributing in solving this problem . [3] [5]

1.5 Problem statement

To perform its function protein has to be located to its pre-determined position hence it is very important to find the subcellular localization of protein. Protein miss location has proven to be the cause of several human diseases, such as Alzhrmeir's disease and cancer [16]. Therefore, it is very important to find the subcellular localization of a protein to understand it's working. As discussed in literature review there are some drawbacks and limitations but the major problem that we identified in [3] and many other method was the use of fixed threshold. As, in large image dataset, intensities distribution are different from image to image. For instance there can be images where proteins are stained with high intensities but there can be some images where DNA channels or other cell organelles are stained with even higher intensities, therefore it is not suitable to have fixed threshold to extract protein features as it can easily confused with different component present in an image.

1.6 Purpose

The purpose of this research is to build a model that can extract features based on adaptive Threshold Adjacency statistics (TAS). Our main goal is to implement such model that can produce threshold values for each image according to their own intensity distribution and with that threshold extract features. For this purpose we will use Otsu's Adaptive thresholding, using this our model will be able to learn features that contain more information and represent proteins in a clear way and help in classifying protein subcellular localization.

1.7 Significance of the Solution

The solution that is proposed in this research is to have adaptive threshold in which threshold value is obtained by analyzing each image with the help of histogram so that every image has its own threshold value according to its intensity distribution. With this, we will be able to achieve binary images that will give maximum information regarding proteins. After that we will extract features and apply classifiers.

As our research is to make a model in such a fashion that it uses adaptive thresholding technique, by doing this we will be able to extract more information of protein location from images resulting in achieving high accuracies in classifying protein subcellular localization.

Chapter 2

Literature Review

For the past decade, with the advancement in the microscopic images, several techniques were proposed and discussed by the researchers for the prediction of protein subcellular localization based on images each discussing different problems in the literature review given below.

Wei et al [17], focused on the problem that all machine learning hand-crafted feature descriptors for protein images extract unsupervised features which do not take into the account of information of class thus resulting into non distinguishing features for the classification task. The other problem that they discussed that these descriptors learn very shallow representation of the protein images and these shallow based feature extraction methods may not be sufficient. With the advancement in machine learning, Deep learning models are capable of learning high level features to best represent the biological images. Wei et al proposed CNN model for generating the representation of protein images. They used the Human Protein Atlas (HPA) Dataset (version 13) which contains 24,028 antibodies that are related to 46 different normal human tissues. They used 7-layer CNN Alex-Net that was pertained on ImageNet and used partial parameter transfer strategy in which they used first 4 conv-layers parameters trained on ImageNet and used updated parameters using protein images for layers above conv-4. They used the Lasso model for features selection and used DECOC for solving multi-class classification problem. They did different experiments to get the best result and succeed in outperforming previous hand-crafted feature extraction methods. Using CNN model they achieve the classification accuracy of 0.579 on HPA human tissues dataset. Even though, their model outperforms the other methods, there are some drawbacks. For instance, their model classified only 20% of the human proteins that are localized in two or more than two cellular compartment, also, cellular compartments with less protein images have lower accuracies.

Jin et al [18] inspired by the performance of deep learning in image classification, proposed a method to classify protein subcellular localization using CNN. They also used dataset by Human protein Atlas (HPA) containing 563 images of 188 protein in healthy liver tissues. Before applying CNN model they preprocessed the dataset. As the original image is 3000X3000 it was improper to take the whole image as input in CNN model so they divided each image into 100 small patches with the hypothesis that small image patches with high protein expression in each image can represent the subcellular localization pattern of the whole image. They also rebalance the dataset by randomly selected image patches from the images that have less examples in the dataset and processed them through rotation and flipping the images to make the number of classes equal. The final training dataset consisted of 86400 patches. The Dataset was divided into training, testing and validation dataset by ratio 4:1:1 and there was no overlapping of protein among these three dataset. After preprocessing they applied the CNN model inspired by the existing CNN model named DeepYeast which have 11- layers as mentioned above. ReLU activation function was used as it performed better than other activation functions. Xavier initialization scheme was used for the weights of intermediate layers. Stochastic gradient descent (SGD) was used as the optimizer. To compare their model performance they implemented SVM model as SVM has been shown as the most effective model in terms of subcellular localization of protein [3]. For SVM model they extracted 57 features including Harallick texture features and overlapping features. After comparing both CNN and SVM model, the result was in the favor of CNN. Deep neural network achieved 47.31% accuracy while the SVM only managed to achieve 39.78% of accuracy. Although deep neural network outperformed the SVM but it took 17 hours to train the model, also the results shows that the rebalance performs an important role in increasing the accuracy but producing image patches from 3000x3000 images may cause loss of information.

Tanel et al [19] also used deep neural network for its capability of overcoming the feature selection problem. They trained a CNN model named Deep Yeast to classify fluorescent protein subcellular localization in yeast cells. They constructed labeled dataset based on high-throughput proteome-scale microscopy images from Chang et al. the dataset consist of 7132 images of 12 classes. DeepYeast architecture consisted of 11 layers where 8 of them were convolutional layers and 3 were fully connected. They used Glorot-normal initialization technique for weight initialization and used batch normalization. Stochastic gradient descent (SGD) was use for optimization. To compare their model they also trained random forest using features that was extracted through Cell Profiler. After several experiments results showed that the DeepYeast achieved the highest accuracy of 91% greater than random forest that only achieved 79%. The major drawback of this model is that it took 3 days to train this model.

Tahir et al. [20] proposed a hybrid model for protein subcellular localization of fluorescence microscopy images. The author addressed the problem that analysis of these images for classification is prone to human errors. They introduced threshold technique in which they used fixed interval threshold values (40, 60, 80) and based on these threshold and mean of intensity of image pixels they extracted 7 binarized images and seven features vectors.They extended their research in which they ectracted three binarized images [21]. All those seven features vectors they then applied SVM for classification using one vs all technique. At the end they predicted the final result using ensemble voting technique. The dataset that they used was LOCATE Endogenous and through this technique they achieved 99.2% accuracy. The problem with this technique was that they used fixed threshold that may work for some images but as they mentioned there are some images in which their technique was not able to identify

Mengli et al. [6] implemented various models of machine learning including different architecture of CNN and compared their performance in image analysis for protein subcellular localization. They first used the VGG-type, 11 layer visual geometry group CNN as their baseline model that was trained on Dataset of natural objects, aircraft and so forth. Their model specification was followed by Deep-Yeast model [16]. In their model every layer was followed by batch normalization and they used softmax function to generate estimated probability. For optimization, stochastic gradient descent was used with momentum 0.9. The model was trained for 195,000 iterations. They used another architecture, ResNet to reduce the number of parameters and compared it with the VGG-type CNN. In this model they used 18 layered and 50 layered model with Adam optimizer. They also discussed various CNN models and their performances. After implementing CNN they tested traditional Machine Learning methods. To compare with CNN, they implemented two tree ensemble methods, Random Forest and gradient boosting. Further they also implemented linear discriminant analysis, K-nearest neighbor, and linear SVM and lasso logistic regression. They chose VGG-19 as the feature extractor. They used dataset constructed by Tanel et al. [19] which consist of 65,000 training, 12,500 validation and 12,500 test single-cell microscopy images. The comparison of these techniques are given in table 2.1. Result shows that the accuracy of Res50 was highest (88.6%) but with the cost of 12.75hrs of training. On the other hand, 11-layer VGG-type model took almost half of the time that is 6hrs and predict with the accuracy of 87.4 % which is ony 1.2% lower. Therefore, CNN out performed all the other models.

Wei et al. [22] observed that all existing model were constructed on the independent parallel hypothesis, where he cellular compartment classes are positioned independently in a multi-class classification engine. The important structural information of cellular compartments is missed. They tried to solve this problem by proposing a cell structure-driven classifier construction approach by including the prior biological structural information in the learning model. The Dataset that they have used was generated by collecting 1636 IHC images with high validation from HPA. It consist of 21 proteins related to 46 normal human tissues. For protein images they extracted features using Haralick features with 10 different

Table 2.1: Comparision of prediction accuracy on test dataset among different methods [6].

Network	**Training Time**	**Test Accuracy**
11-layer VGG-type CNN model(Deep yeast model)	6hr	0.851
11-layer VGG-type CNN model with data augmentation(Deep yeast model)	6hr	0.874
Res-18	2.45hr	0.853
Res-50	12.75hr	0.886
Random Forest (Direct feature vectorization, 1,000 trees)	2hr	0.596
XGBoost (direct feature vectorization, 1000 trees)	10hr	0.679
Linear discrimination analysis	16min	0.289
K-nearest neighbor (K=50 selected)	18hr	0.478
Support vector machine ($c = 8$ selected)	18.3hr	0.228
Lasso logistic regression ($\lambda = 0.000796$ selected)	13hr	0.441

vanishing moments. They also applied DNA features and for local features they used LBP. Stepwise discriminant analysis method was used for feature selection. After extracting and selecting features, they used ECOC (Error correcting output coding) method to transform multiclass classification problem into a series of binary classification sub-problem according to a pre-defined codeword matrix. They constructed 10 different SC-PSorter models based on different sets of features extracted from 10 vanishing moments, and for each SC-PSorter, 14 multi-kernal based SVM classifiers were constructed. For final result they used result by combining 10 SC-PSorter- based classifiers via majority voting. This method achieved 0.89 accuracy and compared it with two methods proposed by other authors and proved that their model outperformed other two. Although it performed effectively in case of each protein corresponding to only one location. However, a new method to perform multi-label based classification problem is still an open issue.

Non image data, for instance amino acid sequence, can also be added to improve accuracy as it provides more information.

Ying et al. [7] proposed a method for solving multi-label classification of protein subcellular location. Author discussed that in most of the methods the prediction accuracy is limited by the simple linear model which lead to incorrect targets. Instead of linear statistics they proposed more flexible approach, named iLocator which can handle multi-label and single-label samples simultaneously. The dataset that they used was high quality images from HPA and UniPort. The normal image dataset contained 3240 images from 28 proteins in normal cells of which seven proteins with two or more organelle. The subcellular locations of cancer images are not annotated in HPA, so the cancer image dataset was to be predicted and compared with the data from normal cells to detect mislocalizations. This dataset contained 3696 cancer images of the same 28 proteins as in the normal dataset. Seven cancers were considered in this study. The original HPA image is the fusion of DNA (purple sections) and protein (brown sections so they tested two separation techniques, i.e., linear spectral separation (LIN) and blind spectral separation by non-negative matrix factorization (NMF). The experimental results show that LIN approach outperforms NMF by 5–10% on the testing dataset so they used LIN in further experiments. For feature extraction they used Haralick features (with 10 vanishing moments), DNA features and LBP features and used stepwise discriminant analysis for feature selection. For classification they trained SVM model using BR or CC modes. Each classifier based on BR or CC could output a seven-dimensional (7D) score vector per testing image, where score represents the confidence of belonging to the corresponding class. They used two criteria, i.e. top criterion and threshold criterion. Then they took the average of these output vector to get final result. The accuracy that they achieved with this model was 92.71%. Although is was a good classifier but in total there were 20 classifiers to be trained on each vanishing moment with both BR and CC modes resulting in high computational expense. Ying et al. [23] observed the problem with the BR framework that it separates a multi-label problem into n binary classifiers ignoring the relationship among labels. To solve this problem author proposed a solution to incorporate organelle correlation in classification underline model. For dataset they used two types of protein images, IHC and IF. Each type of images has three datasets. The ADN set contains IHC images with high-level expression o IF images with reliable annotation samples in ADN set are regarded as labeled data in semi-supervised learning. Then those protein images with medium-level expression or uncertain annotation are collected in BDN dataset, and they are taken as the candidate selective unlabeled data in semi-supervised learning. IDN is independent dataset that is used for testing. They employed five different semi-supervised learning methods. They implemented AsemiB, logistic label propagation (LLP), low density separation, cost-sensitive semi-supervised support vector machine and transductive multi-label classification. The comparison of these showed that AsemiB method outperformed other tested methods. For incorporating organelle correlation, three step method was follow. First correlation graph was constructed. Bayesian DAG was used to learn the network structure. In second step, BR predictor containing N independent SVM classifiers was built based on original features. In third stage N new binary SVM classifiers were trained, and the feature space of each binary classifiers and the order of training these classifier are determined by the correlation graph. Training chain classifier according to this order can ensure that the additional features are updated outputs, which are more accurate than rough labels. With this the accuracy achieved was 56% which is 0.75-6.25% higher than the single classifier.

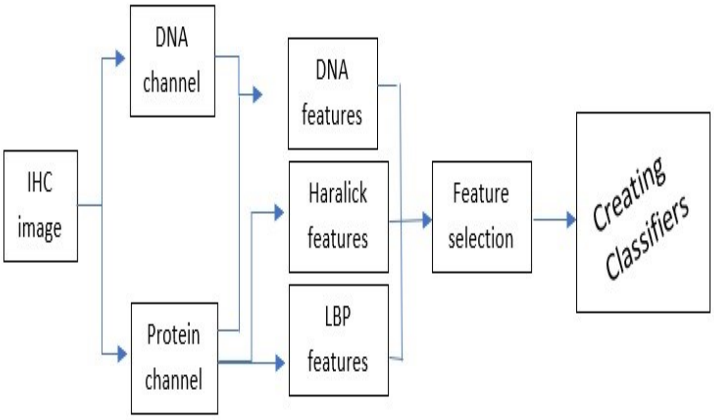


Figure 2.1: Flowchart of the procedure of iLocator creation with normal image dataset

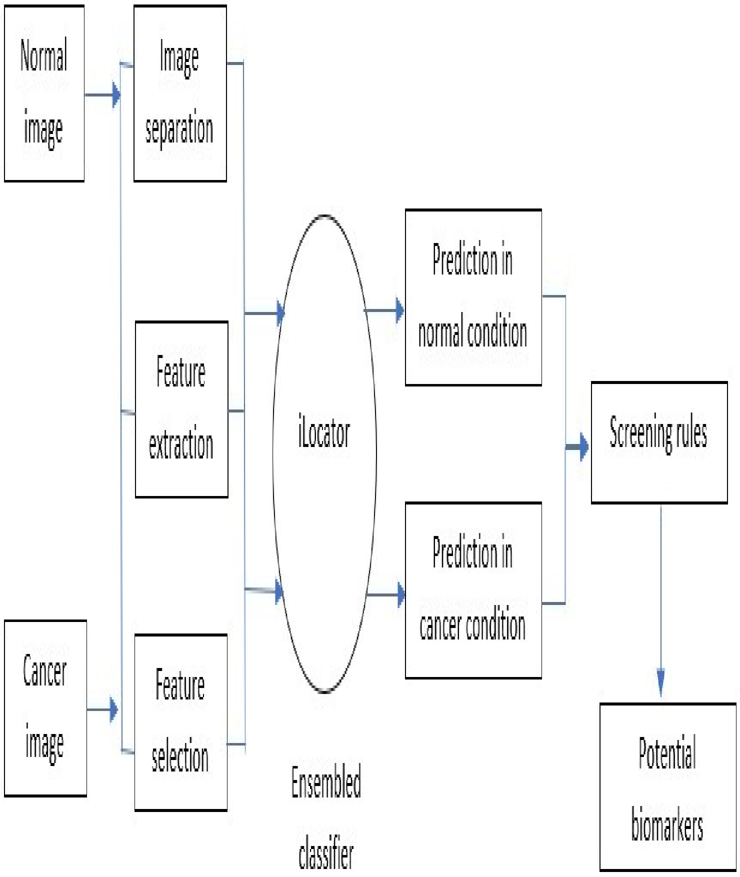


FIGURE 2.2: Flowchart of process of biomarker protein detection using iLocator.

Wei et al. [24] proposed a method to solve the problem while selecting features through SDA stepwise discriminant analysis that this method neglects to take the correlation among different compartments into consideration. For this they proposed organelle guided multi-label feature selection method CSF by employing the biological structural correlation among different cellular compartments. Two regularization items were included in the objective function. The first item was the group-sparsity regulaizer, which ensured only a small subset of common features was selected across different subcellular compartments. The second item is a cellular correlation regularized Laplacian term, which utilizes the prior biological structural information to capture the intrinsic relatedness across different cellular compartment. The dataset that they used was Protein Atlas dataset which contains a large number of immunohistochemistry (IHC) images of proteins. They used version 13 which contains 24028 antibodies, which are related to 46 different normal human tissues. First they separated DNA channel from the protein using non-negative matrix factorization then they extracted features using Haralick and LBP from protein channel and DNA features from DNA channels. For feature selection they used their proposed method, CSF, by utilizing the prior biological structural information. For prediction they used based multi-label learning mode. After certain experiments they came to the conclusion that their proposed methodology outperformed the other with which they compared. On benchmark dataset they achieved 93% accuracy.

Wei et al. [5] in 2019, Human Protein Atlas organization collaborated with kaggle and held a competition to identify the deep learning solutions which perform the best in classifying protein subcellular localization patterns [5]. Over 3 months, many teams participated using vast variety of different networks and pre-trained models. For this competition they prepared a dataset HPA(version 18) of confocal microscopic images of protein in different cells. Each image in the dataset consisted of 4 channels making the protein of interest. Green channel indicates the protein of interest, yellow channel contains ER, Red channels indicates the microtubule and Blue channels indicate nucleus. The dataset is highly imbalanced multilabel dataset. Because of imbalance dataset the measure of evaluation used is macro-f1 score. The team that ranked first in this competition achieved macro f-score of 0.593 and they used neural network with loss function Lovasz loss.

Nicholas et al. [25] proposed a methodology that can overcome the high computational cost exclude cropping of image. In the proposed technology TAS- Threshold Adjacency statistics they used a threshold value the mean of the image and binarize the image using this threshold after binarization they extracted feature by TAS by using the binarize image. TAS extract 9 features, 1st feature is the sum of all white pixel whose neighboring pixels are not white. The second feature is the sum of all white pixels whose neighboring pixels has only one white pixel.

Table 2.2: Models and their performance for top ranking teams [5].

Rank	Team Name	Member(s)	Score
1	Team 1: bestfitting	D.Shubin	0.593
2	Team 2: WAIR	J.Lan	0.571
3	Team 3: pudae	P.Jinmo	0.570
4	Team 4: Wienerschnitzelgemeinschaft	S.Mahmood Galib et al.	0.567
5	Team 5: vpp	Y.Gu, C.Li	0.566
8	Team 8: One more layer	D.Buslov et al.	0.563
10	Team 10: conv is all u need	X.Cao et al.	0.557
16	Team 16: NTU_MiRA	K.-L. Tseng	0.553
39	Team 39: Random Walk	Z. Gao et al.	0.540

Table 2.3: comparison of accuracies of different feature extraction methods on Endogenous and Transfected dataset [25]

Feature Extraction methods	Accuracy on Endogenous Dataset	Accuracy on Transfected Dataset
TAS	94.4%	90.3%
Haralick	94.2%	86.0%
Zernike	75.8%	68.6%
TAS+Haralick	98.2%	93.2%

Similarly the third feature is the sum of all pixels whose neighboring pixel has two white pixels. In the same manner the 9th feature is the sum of all white pixels whose neighboring pixels are all white. All these 9 features are normalized. After feature Extraction they used SVM .The dataset they used are endogenous and transfected and they achieved accuracies 94.4% and 86.6%. They also tested with TAS+Haralick features and achieved 98.2% and 93.2% accuracy.

TABLE 2.4: Critical analysis of Literature Review

Research paper	**Dataset**	**Methodology**	**Results**
[24]	1- Human Protein Atlas (HPA version13) 2- Dataset containing 1040 IHC images	CSF feature selection model.	Accuracy 93%
[20]	1-LOCATE-Transfected 553 images 2-Endogenous Dataset 502 images	modified TAS	Accuracy 99.2%
[17]	Human Protein Dataset (HPA version 13) containing 24,028 Related to 46 different human tissues	7-layer CNN Alex-Net that is pre-trained on ImageNet	Accuracy 57.9%
[18]	Human Protein Atlas (HPA) containing 563 images related to 188 protein in healthy liver tissues.	11-layerd CNN model	Accuracy 47.31%
[19]	Constructed dataset based on high-throughput proteome-scale microscopy images consisted of 7132 images and 12 classes	Deep Yeast 11 layers in which 8 are convolutional layers and 3 are fully connected.	Accuracy 91%
[6]	Dataset consisted constructed by Tanel et al [19]	VGG-type CNN 11 layer visual geometry group	Accuracy 87.4%
[7]	1-Human Protein Atlas (HPA) 2- UniPort Dataset consisted of 3696 cancer images	1-Haralick Features 2-DNA Features 3-LBP features	Accuracy 89.4%
[23]	ADN Dataset contains IHC images IDN Independent Dataset for testing	LLP Cost-sensitive semi supervised support vector machine	Accuracy 56%

Table 2.4 shows the critical analysis of different approaches used in order to predict protein subcellular localization. Different model performed well on different datasets. Along with others modified TAS performed very well with the achieved accuracy of 99.2%.

Chapter 3

Methodology

As discussed earlier with the advancement in microscopic images of proteins, researchers have come up with multiple techniques to predict the protein subcellular localization. These techniques mainly comprised of Machine Learning models to classify protein subcellular localization. Various Algorithms has been designed which include feature extraction methods such as Heralick Features [7] TAS and ETAS techniques [9] along with different classifiers such as svm, pretrained models etc. With the advancement in Machine Learning over the past few years Deep neural networks are widely used in image classification [26] [27]. Most of these techniques also include different model of convolutional neural network. Convolutional Neural Networks (CNN) is consider as the very powerful image classification models. CNN-based models are capable of achieving state-of-the-art results in classification, localization, semantic segmentation and action recognition tasks, amongst others [28]. Nonetheless, they have their limits and they have fundamental drawbacks. A Convolutional neural network can be slow depending on various operations such as maxpool. In order to perform well for CNN with several layers requires a good GPU because of the computational constraint otherwise the training process can take a lot of time. Along with the good computational power, large amount of Dataset is require to train the neural network. Due to these limitations and drawback we proposed the algorithm which extract features by consuming less computational power, less time and is capable of producing state-of-the-art result

in predicting Protein Subcellular localization. Our proposed methodology uses adaptive threshold to obtain 27 statistical feature vector of an image and apply classifier to classify protein subcellular localization.
This approach consists of three phases:

1. Feature Extraction phase.
2. Training phase.
3. Evaluation phase.

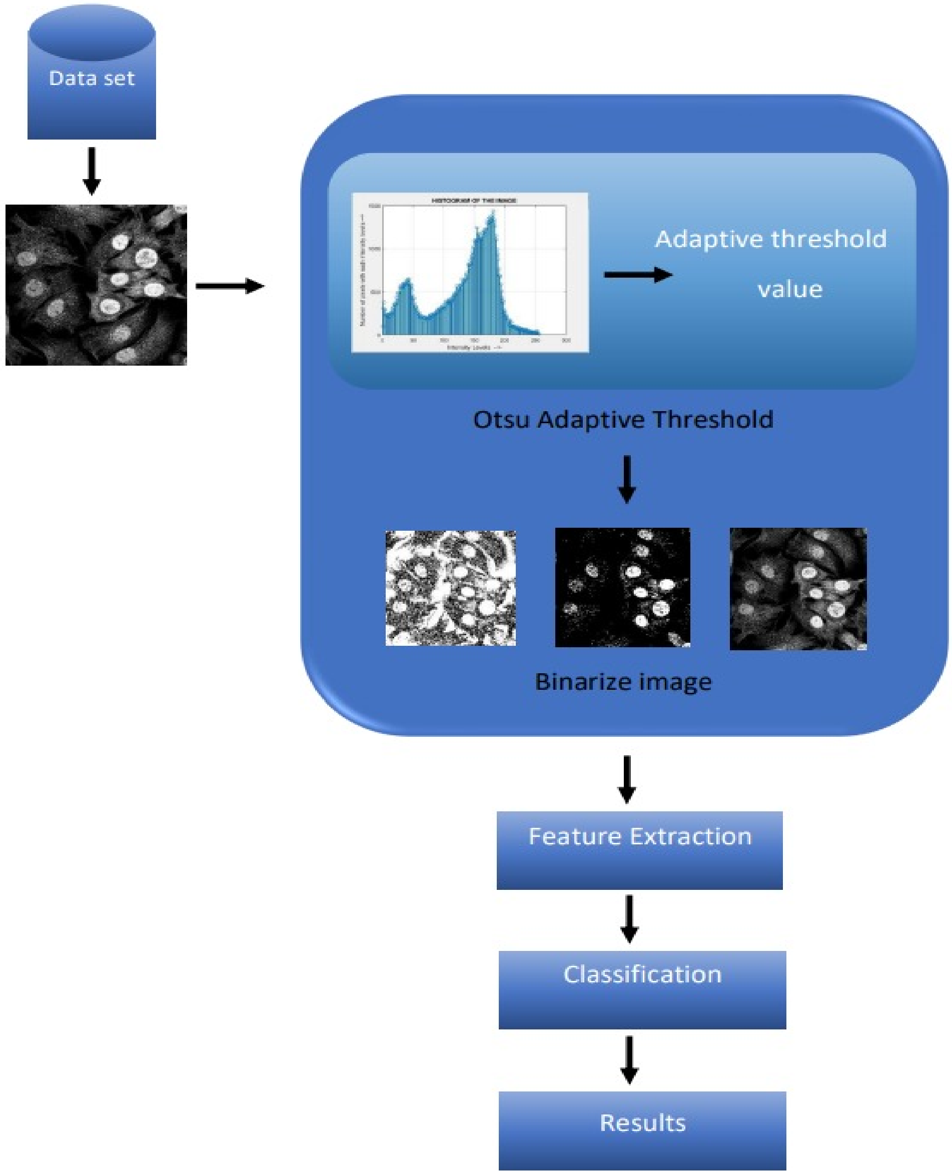


FIGURE 3.1: Proposed Methodology

In feature extraction phase features vector are extracted using adaptive threshold and train these feature vectors using classification model. In the evaluation phase this study test the model on test dataset and evaluate the performance of this proposed approach and compare the result with different approaches.

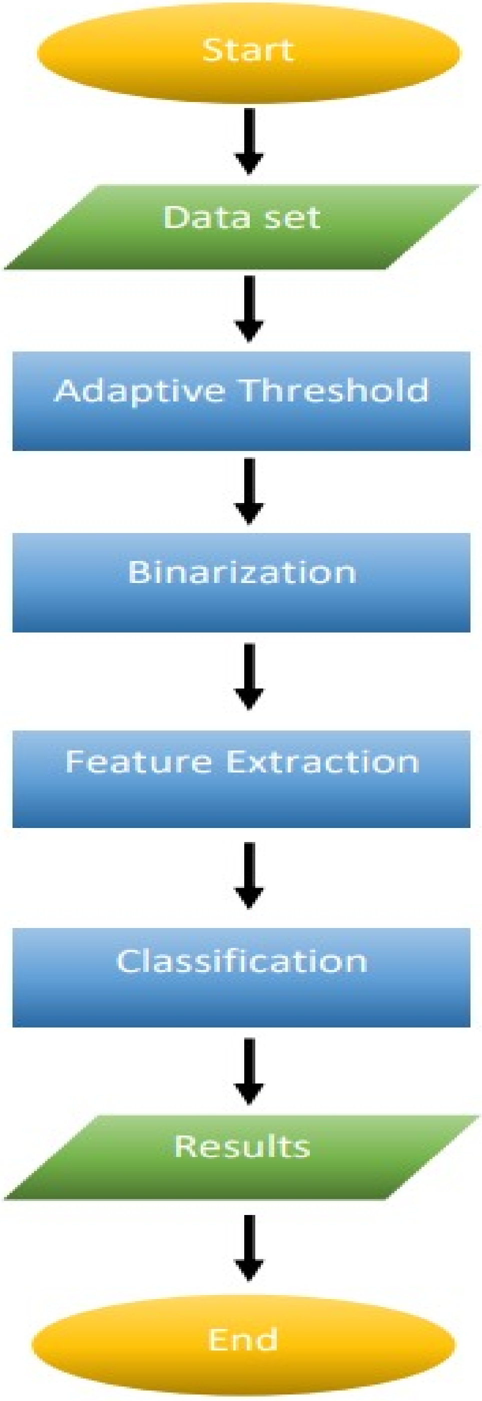
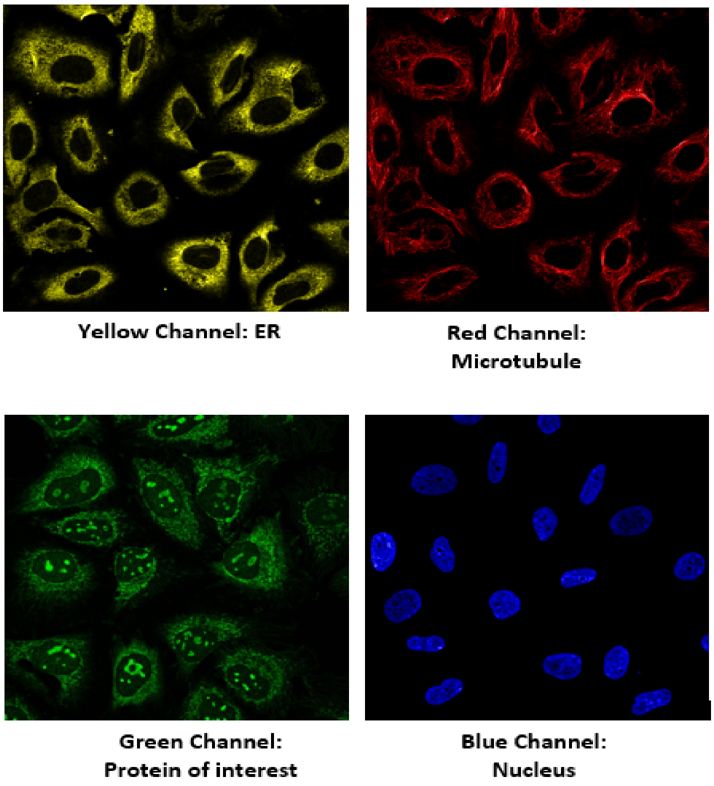


Figure 3.2: Flowchart of Proposed Methodology

3.1 Dataset

In 2019, Human Protein Atlas organization collaborated with kaggle and held a competition to identify the deep learning solutions which perform the best in classifying protein subcellular localization patterns [5]. Over 3 months, many teams participated using vast variety of different networks and pre-trained models. For this competition they prepared a dataset HPA(version 18) of confocal microscopic images of protein in different cells. Each image in the dataset consisted of 4 channels making the protein of interest.

1. Green channel indicates the protein of interest.
2. Yellow channel contains ER.
3. Red channels indicates the microtubule.
4. Blue channels indicate nucleus.



Yellow Channel: ER

Red Channel: Microtubule

Green Channel: Protein of interest

Blue Channel: Nucleus

FIGURE 3.3: 4 Different Channels

These channels can be very helpful in classifying protein localization in cell images but green channel is the one which has protein that we need to classify. Since protein of interest resides in the green channel, this study used all images of green channel and applied the proposed technique and successfully outperformed state-of-the-art technique.

3.1.1 Challenges

Challenges included in this dataset are follows:

1. First challenge is to train model on high imbalanced classes. There are 28 different protein that we need to predict in 27 different cell types. The most common label in the dataset was ‘nucleoplasm’ having 12,885 images and the rarest label was ’rods and rings’ with only 11 images in dataset.
2. Second challenge is predicting multi labels per images. There are some images that belongs to more than 1 label.

3.2 Feature Extraction

The second phase of proposed methodology is to extract features from the images in the dataset. For this purpose this study used modified technique that is used in ETAS-Subloc [20].

3.2.1 ETAS-SubLoc

In the ETAS-SubLoc [20] researchers extracted global features using modified Threshold Adjacency Statistics. Tahir et al [20] proposed this technique to increase the efficiency and enhance the discrimination power. In the modified Threshold Adjacency seven threshold ranges are obtained by using a fixed threshold value 30. With these seven threshold ranges, seven binary images are obtained. With every binary image nine feature vectors are obtained and then these feature vectors are used to train seven SVM's. Majority voting Scheme is used for the final prediction (figure 3.3). In this research methodology, modified technique EATAS-SubLoc is used. Compare to ETAS-SubLoc [20] this methodology use 3 distinct feature spaces for each image and instead of using fixed threshold it use adaptive Otsu based thresholding.

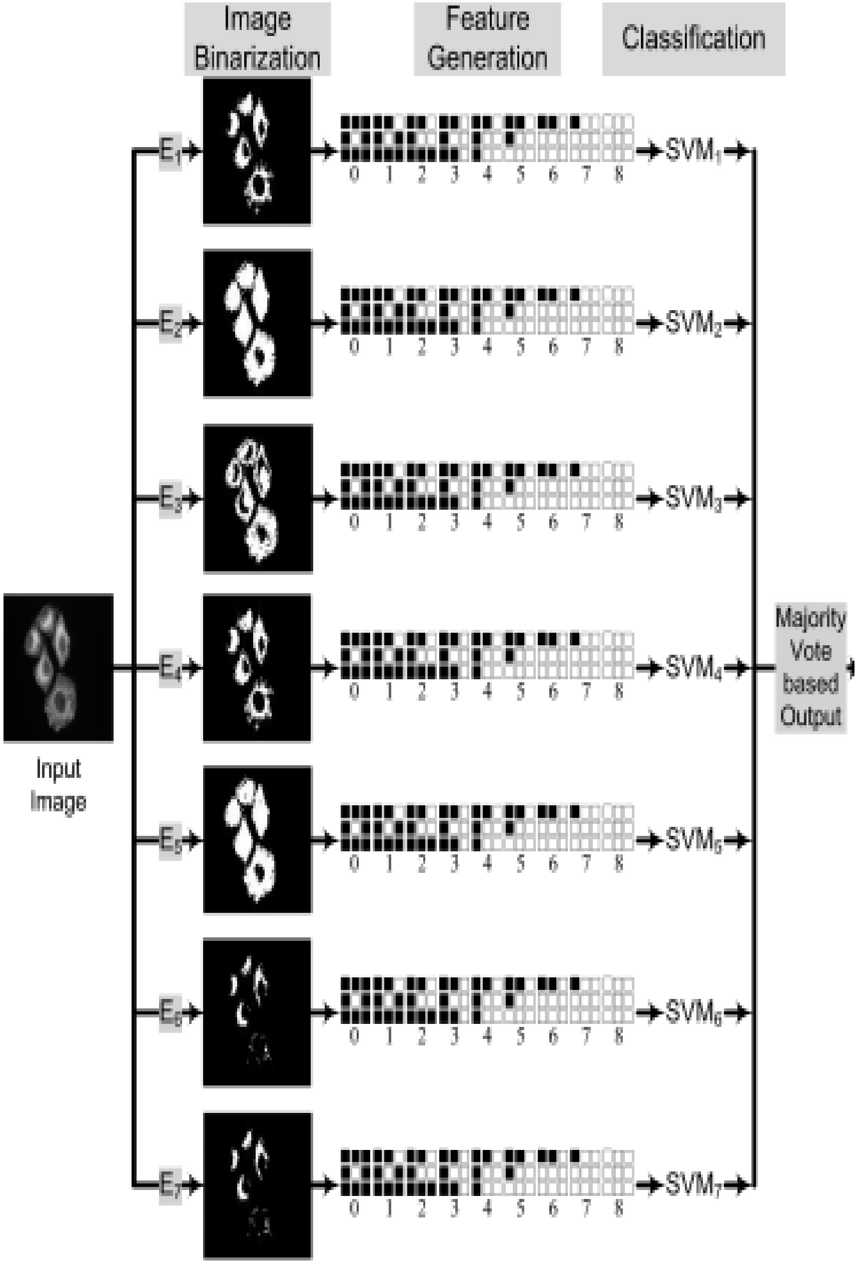


FIGURE 3.4: Working of ETAS-SubLoc [20]

3.2.2 Adaptive-TAS-SubLoc

Adaptive Threshold Adjacency Statistics are produced by first applying Otsu adaptive threshold by calculating the threshold value using image histogram so that we don't miss on the intensities in the foreground. With this threshold value and the average intensity of the pixels value we generate binary images and then extract features out of these binary images. For this research methodology we used random forest classifier for these feature vectors to train on it.

3.2.2.1 Otsu's Adaptive Thresholding

In order to predict protein subcellular localization, segmentation of an image is very important so that we can identify foreground information with the background information.

In protein images, fixed threshold will not give you good results because in some images the pixels value are so close to the background value that it completely disappear or in some cases it cannot differentiate between background and foreground. In this scenario we lose very valuable information which is required in protein subcellular localization. To overcome this problem, this study used Otsus's thresholding technique which generate threshold value by computing histogram of an image. Histogram is created by using the 8-bit grayscale [0-255] values of an image. It calculates the number of occurrences of each pixel value in an image. Otsu threshold use the histogram of an image to find the threshold value that is optimal value to separates the foreground from background.

Otsu Algorithm first establish the histogram H of an image and iterate through each threshold value t [0-255] to separate the pixels into two classes, C1 and C2; foreground and background. [29]. The formula of variance where μ is the mean, N is the total number of pixels and P_i is the value of ith pixel.

$$\sigma = \sum_{i=0}^{N} (P_i - \mu)/N \tag{3.1}$$

In order to find the with-in class variance at any threshold t is given by

$$\sigma^2(t) = W_{bg}(t)\sigma^2_{bg}(t) + W_{fg}(t)\sigma^2_{fg}(t) \tag{3.2}$$

Where W_{bg} (t) and W_{fg} (t) is the probability of pixels for each class at threshold t which is given by.

$$W_{bg}(t) = P_{bg}(t)/P_{all} \tag{3.3}$$

$$W_{fg}(t) = P_{fg}(t)/P_{all} \tag{3.4}$$

Where P_{bg} (t) and P_{fg} (t) are the total count of pixels in background and foreground classes respectively at threshold t.

Algorithm : Otsu's Adaptive Thresholding

Input: Img (Image)

```
Hist $\leftarrow calculateHistogram(Img)$
$maxintensity \leftarrow getMaxIntensityOfImage(Hist)$
$Fn\_min \leftarrow infinity$
$iinrange(1, maxintensity)do :$
$C1 \leftarrow calculateBelowThresholdPixels(Img, i)$
$C2 \leftarrow calculateAboveThresholdPixels(Img, i)$
$P1, P2 \leftarrow calculateProbabilities(C1, C2)$
$W1, W2 \leftarrow calculateweights(C1, C2, i)$
$M1, M2 \leftarrow calculateMean(P1, P2, W1, W2)$
$V1, V2 \leftarrow calculateVariance(P1, P2, W1, W2, M1, M2)$
$fn \leftarrow V1 * C1 + V2 * C2$
$iffn¡fn_minthen :$
$fn\_min \leftarrow fn$
```

```
        $thresholdvalue \leftarrow i$
    $endif$
$endfor$
```

3.2.2.2 Image Binarization

After obtaining the threshold value, EATAS-SubLoc generates three binarize images of an input image using three different ranges. These binarize images have the intensities in the range µ to255, µ-t to 255 and µ-t to µ+t as shown in equations below. Here t is the threshold value that is obtained through Otsu's algorithm and it is the optimal value that performs efficient segmentation.

$$E_1 = \mu to 255 \tag{3.5}$$

$$E_2 = \mu - \tau to 255 \tag{3.6}$$

$$E_3 = \mu - \tau to \mu + \tau \tag{3.7}$$

3.2.2.3 Feature Vectors

The optimal value of threshold is obtained and using that value images are binarized. After binarization, each binarize image produces 9 statistical feature vector to exploit the dissimilarity seen in the threshold images. These nine statistical are obtained by counting the number of the adjacent white pixel for each white pixel.[fig]. Thus the first statistic is the total number of white pixel with no adjacent white pixel; the second statistic is the total number of white pixel with only one white pixel in neighbor. By finding all the count of the neighboring white pixel up to the maximum of eight, we will get our 9 statistic feature vector. Then these

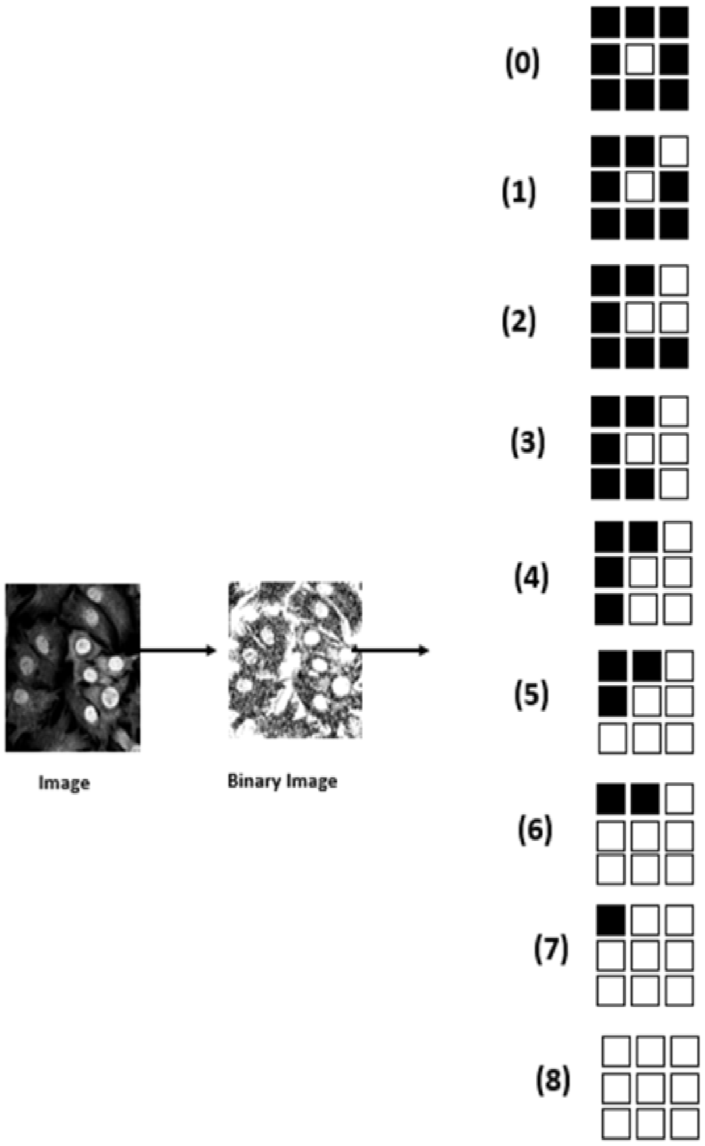


Figure 3.5: Feature Extraction from binarize image

nine statistics are normalized by dividing every feature with the total number of white pixel in binarize image.

3.3 Training Phase

After feature extraction the next phase is to build a classifier on these feature vectors. As discussed previously, dataset used in this study is imbalance and multilabel. To overcome this problem of imbalance multilable classification this research methodology use Multi-label Smote to tackle the imbalance of the dataset. For classification of feature vectors this study use Random Forest Classifier.

3.3.1 MlSmote

In some classification problem the number of instances which belong to one class is very low as compare with the other class instances which generates a problem of data imbalance and it highly effects the performance of our machine learning algorithms. Similar problem also occur in the case of multi-label classes where the class distribution is uneven. To solve this imbalance dataset problem, a very effective approach of data augmentation for imbalance multi-label data is used which is MLSMOTE-multi-label synthetic minority over-sampling.
MLSMOTE is every effective and most popular data augmentation technique that is used for imbalance multi label classification. MLSMOTE is the extension of the technique SMOTE – Synthetic Minority over Sampling Technique. SMOTE worked on the following principle.

1. select the minority class label.
2. Select an instance of the data belonging to that class.
3. Finding the K-nearest neighbor of the selected instance.
4. Select a random data point from the K-nearest neighbor of the selected instance.
5. Take data intense anywhere on the line that joins the random data point and selected instance.
6. Repeat these steps until the data is balanced.

In SMOTE we deal with samples of the same class but in multi-label classification we have multiple classes and SMOTE fail in this setting as there are more then one class associated with every instance of the dataset. In multi-label classification we have the possibility that the instance of the data belongs to majority label and also have another label that belong to minority. The majority labels are called the head labels and the minority as the tail label in Multi-label setting.

Steps include in MLSMOTE is to first select data to augment with proper criteria as to which labels are considered as minority. Once the data is selected for the minority or tail label instances. We generate new instances according to selected instances.

In order to generate synthetic instances we need to choose instance from the dataset from which data is to be created. For this purpose we need select tail labels so that we can generate instances belonging to the tail labels. In order to select tail labels Imbalance ratio per label (IRPL) is calculated individually for each label along with Mean Imbalance ratio (MIR) which is defined as the average of IRPL of all the labels.

Every label whose IRPL(l) > MIR is considered as a tail label and all the instance of the data which contain that label is considered as minority instance data.

It then generates the sysnthetic instances based on the selected tail label and random selected KNN data point and clones the label of the selected instance and assign it to newly generated instances. [30]

Algorithm

Inputs:

D (Dataset) K (Number of nearest neighbors)

```
L ← $DatasetLabels(D)$
$MIR \leftarrow calculateMeanImbalanceRatio(D, L)$
$foreachlabelinLdo$
$LIR \leftarrow calculateMeanImbalanceRatioPerLabel(D, label)$
$ifLIR¿MIRthen$
$MinoritylabelsinstancesBag$
$minorityBag \leftarrow getAllInstencesOfLabel(label)$
$foreachinstanceinminorityBagdo:$
```

```
$distances \leftarrow calculateDistance(instance, minorityBag)$
$AsscendingSort(distances)$
$SelectionofNeighborset$
$neighbors \leftarrow getHeaditems(distances, k)$
$RNeigh \leftarrow getRandomNeighbor(neighbors)$
$FeaturesandLabelssetgeneration$
$syntheticInstance \leftarrow newInstance(instance, RNeigh, neighbors)$
$D = D + syntheticInstance$
$endfor$
$endif$
$endfor$
```

3.3.2 Multilabel Random Forest Classifier

After feature Extraction Phase, next step is to train a classifier on the extracted feature vectors. For this purpose this study use Multilable Random Forset classifier. Random forest is a supervised learning algorithm that can be used for both regression and classification but mainly it is used for classification. Random Forest generate decision trees on data instances and get the prediction from each tree and based on voting scheme select the best solution. Rndom forest has almost the same hyperparameters as a decision tree but Random forest is an ensamble method which makes it better than the single decision tree because by averaging the result, random forest reduces the over-fitting. While growing the trees, Random Forest searches for the best feature among the set of features which result in better model. In Random Forest algorithm it is very convenient to measure the relative importance of each feature on the prediction .

Below are the steps for working of Random Forest Algorithm.

Following are the steps that describes the algorithm of Random Forest.

Step1: First select the random sample from the given dataset.

Step 2: Construct the forest by making multiple decision trees for the selected samples and then get the prediction from every Decision tree.

Step 3: For every predicted result voting will be performed.

Step 4: Finally select the majority voted result as your final prediction.

3.3.3 Evaluation

After Building a classifier on training dataset next phase is to evaluate the performance of the classifier on testing dataset. To evaluate the proposed methodology, this study use F-score measure because of highly imbalance dataset.

3.3.4 Macro F1-Score

For a balance between precision and recall Macro F1- score is used because it gives equal importance to every label.

The Macro F1-score is defined as the mean of label-wise F1 scores. Greater value of Macro F1-score indicates the good performance. [31]

$$MacroF1 - score = 1/N\sum_{i=0}^{N}F1 - score_i \tag{3.8}$$

Chapter 4

Results and Experimentation

The proposed methodology in explained in detail in chapter 3. In this chapter experiments and results are discussed. By applying the proposed methodology these results are obtained.

4.1 Dataset

The evaluation of the proposed methodology is dependent on the dataset. As in chapter 3, details about the dataset is already been mentioned. Dataset consisted of 31,072 public confocal microscopic images of protein. Each image has four channels that are used in protein subcellular localization.

With this dataset two main challenges arise. First is the highly imbalance dataset. There are 28 different classes, Table 4.1 shows the number of instances of each categories in this imbalance dataset. The most common label in dataset is nucleoplasm with more than 12,000 images, after nucleoplasm cytosol has less than 9,000 images then plasma membrane has 3777 images, Nucleoli has 3621 images, Mitochondria has 2965 and some rare labels in dataset are'Mitotic spindle' with 210,'Lipid droplets' with 172, 'Proxisomes' with 53 , 'Endosomes'with 45 'Lysosomes' with 28, 'Microtubule ends' with 21 and 'rods and rings' with just 11 images.

TABLE 4.1: Number of instances of each category

Categories	**Number of instances**
Nucleoplasm	12884
Cytosol	8228
Plasma membrane	3777
Nucleoli	3621
Mitochondria	2965
Golgi Apparatus	2822
Nuclear Bodies	2513
Nuclear speckles	1858
Nucleoli fibrillar center	1561
Centrosome	1482
Nuclear Membrane	1258
Intermediate Filaments	1093
Microtuble	1066
Endoplasmic Reticulum	1008
Microtubule organizing center	902
Cell junctions	802
Actin Filaments	688
Focal Adhesion site	537
Cytokinetic Bridge	529
Cytoplasmic bodies	328
Aggresome	322
Mitotic spindle	210
Lipid droplets	172
Proxisomes	53
Endosomes	45
Lysosomes	28
Microtuble ends	21
Rods and rings	11

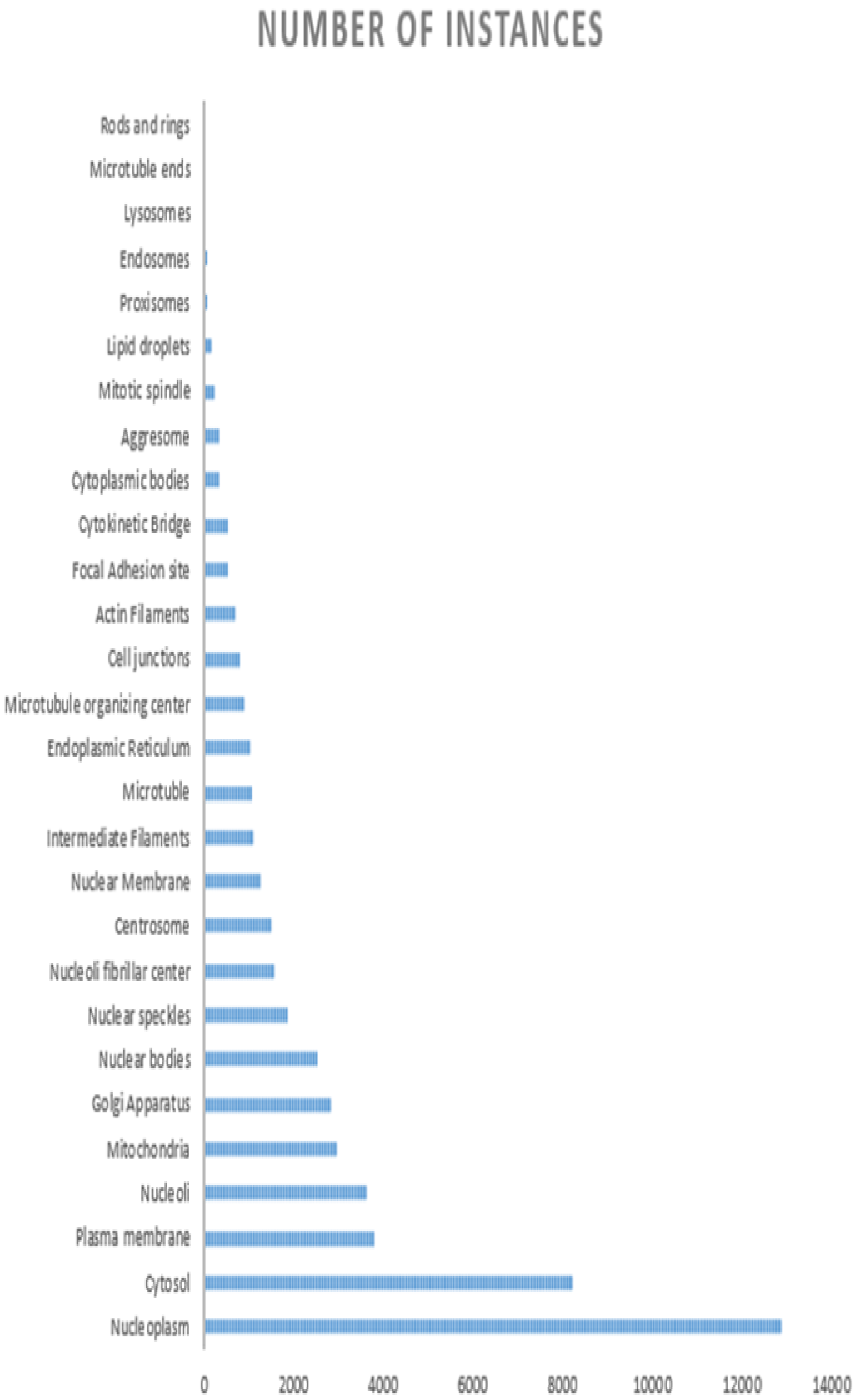
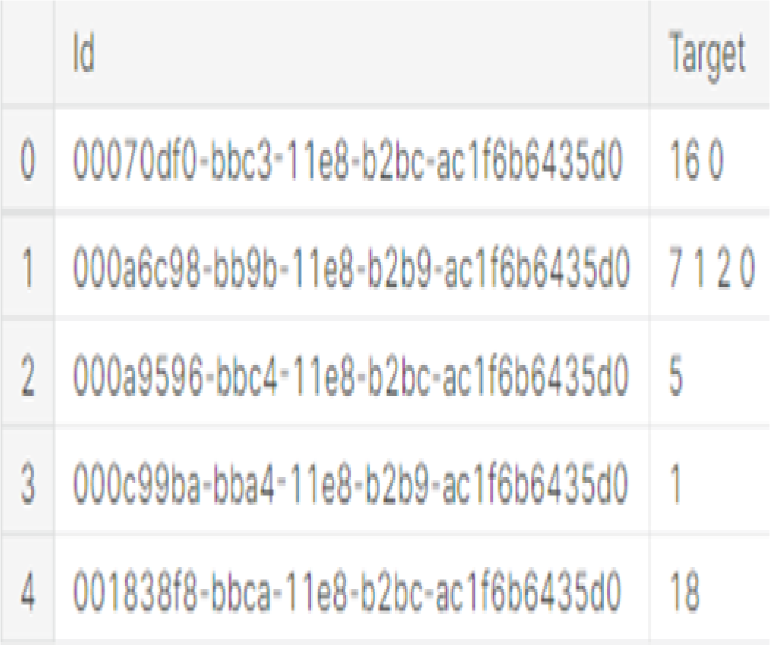


Figure 4.1: Number of Instances

The second challenge that comes with this dataset is the multi label classification in which some images contains more than one label. Figure 4.2 shows that 1st instance of the image belongs to two categories 16 and 0 which are “Nucleoplasm” and “Cytokinetic bridge” respectively. Similarly 2nd instance of the image belongs to 0, 1, 2 and 7 which are “Nucleoplasm”, ”Nuclear membrane”, ”Nucleoli” and



	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

FIGURE 4.2: Sample Dataset with multiple classes

Golgi apparatus respectively and instances 2, 3 and 4 belongs to only one category which is “Nuclear bodies”, “Nucleoplasm” and “Microtubule organizing center” respectively.

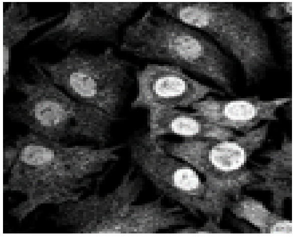
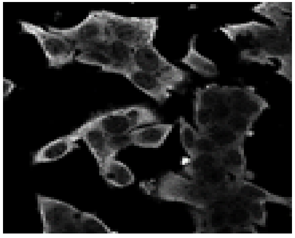
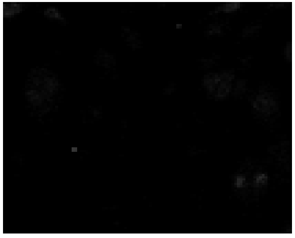
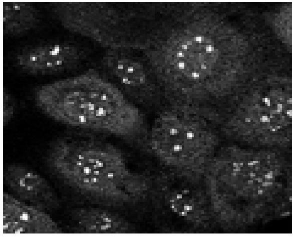
4.2 Feature Extraction

In order to predict Protein Subcellular Localization, segmentation plays a very important role in microscopic images. In order to differentiate between background and foreground images we used threshold value and using that threshold value we binarize image. In proposed methodology instead of using fixed threshold we used adaptive threshold.

4.2.1 Adaptive Threshold

To generate threshold value of each image according to the different intensities in background and foreground we used Otsu’s adaptive threshold method. By applying adaptive threshold we are not losing any valuable information that we might have lose using fixed threshold. Table 4.2 shows the fixed threshold values and threshold values obtained by Otsu's adaptive thresholding of five images with different ranges of intensities in foreground and background. Figure 4.3 shows clearly that using fixed threshold values most of the information is lost.

TABLE 4.2: Fixed v.s Adaptive Threshold Values

Image ID	**Fixed Threshold value**	**Adaptive threshold value**
Image A 	40	89
Image B 	40	54
Image C 	40	8
Image D 	40	40

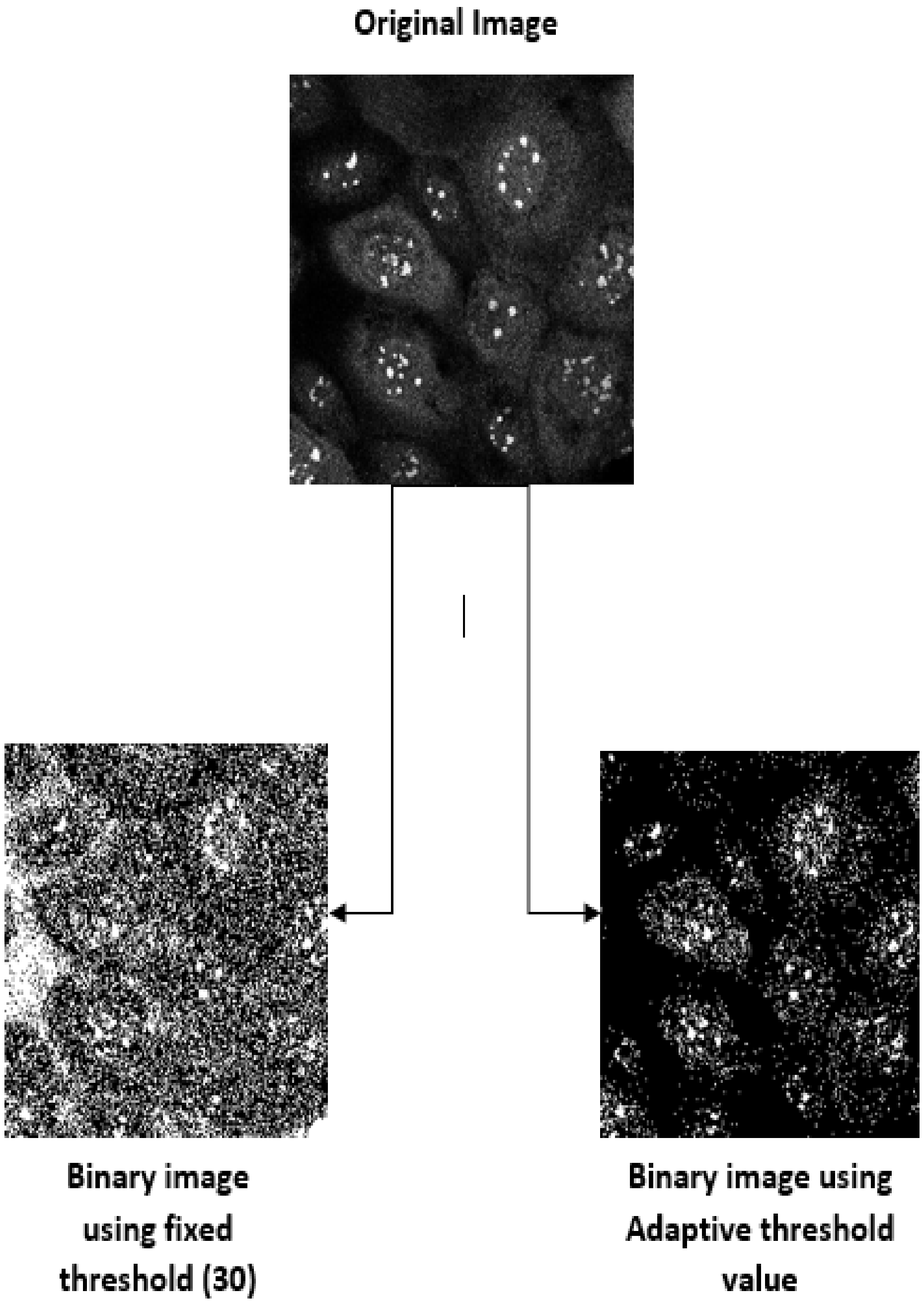

Original Image
Binary image using fixed threshold (30)
Binary image using Adaptive threshold value

FIGURE 4.3: Difference between binary image using fixed and adaptive threshold values

Image obtained by fixed threshold value has a lot of noise because the intensities of foreground and background are not separable using 40 as fixed threshold value on the other hand threshold value obtained by Otsu's adaptive thresholding technique is 51(table 4.2) and hence the segmentation of foreground and background image is much clear and visible in order to predict accurate protein subcellular localization.

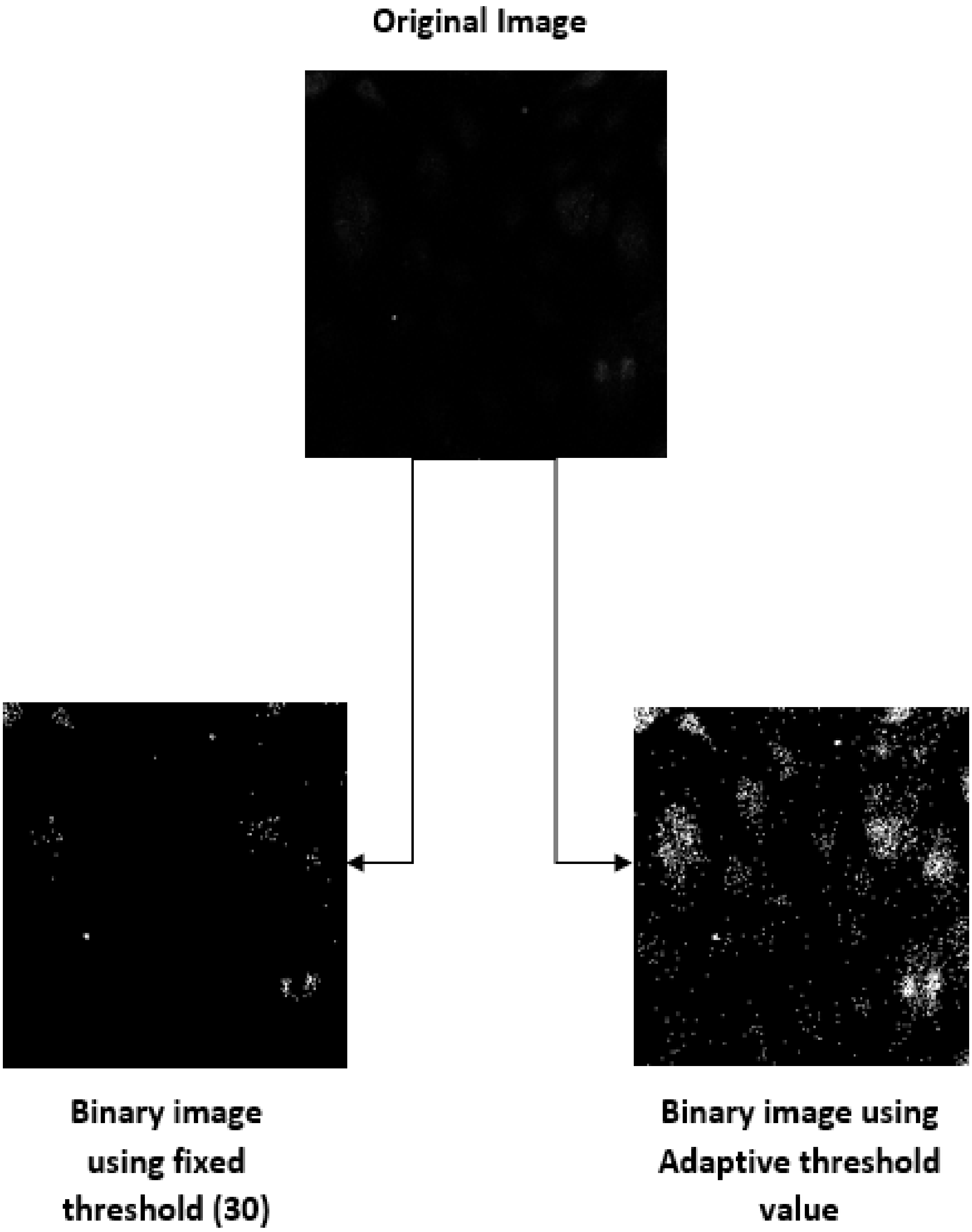

Original Image
Binary image using fixed threshold (30)
Binary image using Adaptive threshold value

Figure 4.4: Difference between binary image using fixed and adaptive threshold values

In an another example (figure 4.4) it is clearly visible that when the variant between classes variance between classes is low the fixed threshold technique fail to separate image foreground to background resulting in loss of valuable information whereas figure 4.4 shows that adaptive threshold technique is capable of differentiating between foreground and background even in such low variance between classes.

Table 4.3: Similarity percentage of image with fixed threshold binary image v/s adaptive threshold binary image

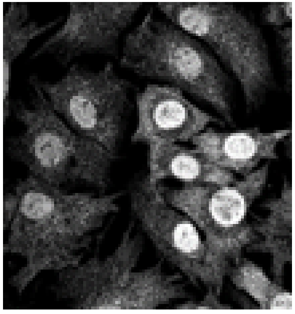
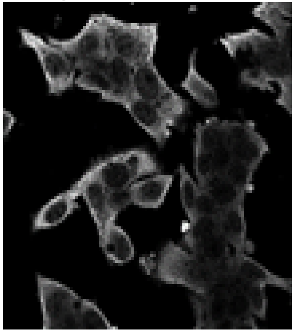
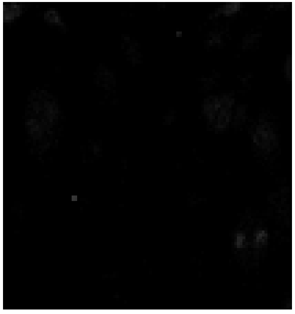
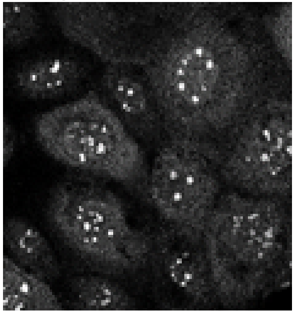
Image ID	**Fixed Threshold value**	**Adaptive threshold value**
Image A 	40.79%	73.86%
Image B 	76.47%	78.51%
Image C 	28.23%	47.77%
Image D 	31.44%	47.7%

Table 4.3 shows the similarity percentage of original images with the images obtained by fixed threshold and images obtained by adaptive threshold. It is clearly visible that images obtained by adaptive threshold is more similar to original image then the using fixed threshold.

Table 4.4: Difference of original image with binary image using fixed threshold and binary image using adaptive threshold.

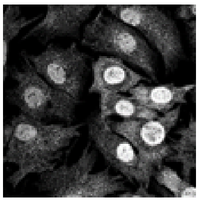
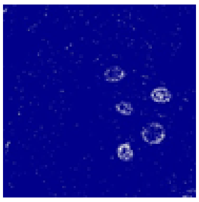
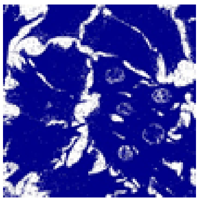
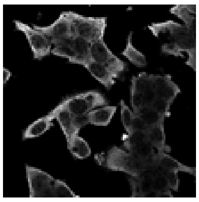
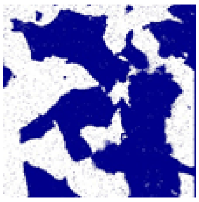
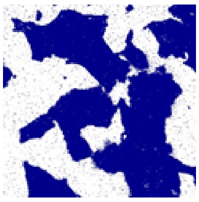
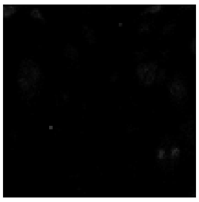
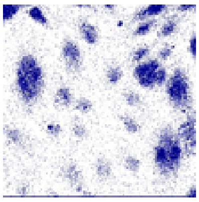
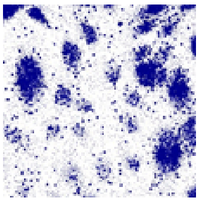
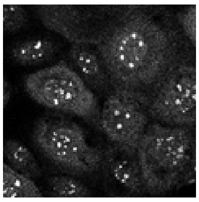
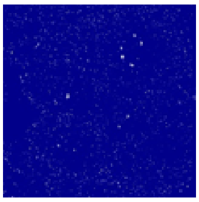
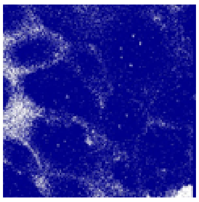
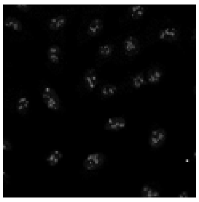
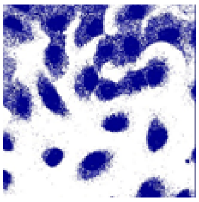
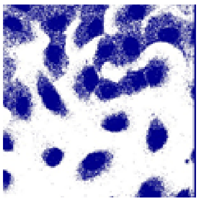
ID	**Original Images**	**Difference of Binary Image using fixed threshold with original image**	**Difference of Binary Image using Adaptive threshold with original image**
Image A			
Image B			
Image C			
Image D			
Image E			

Table 4.5: Standard Deviation of images using fixed and adaptive threshold values

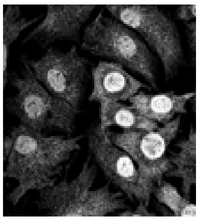
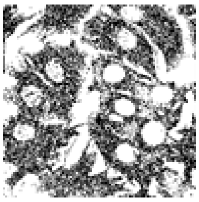
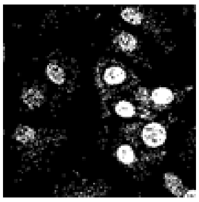
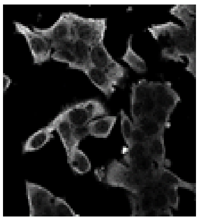
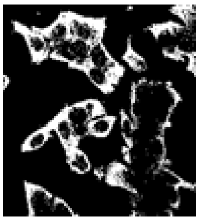
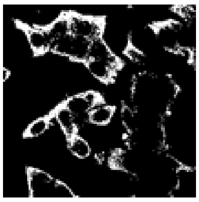
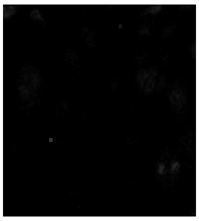
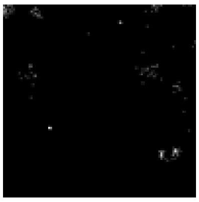
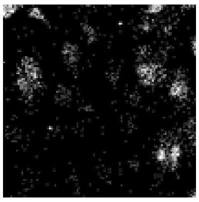
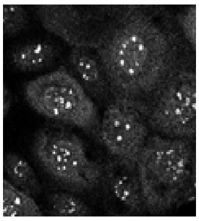
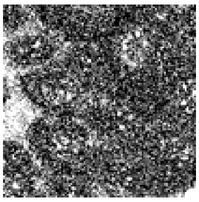
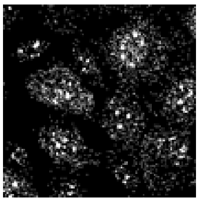
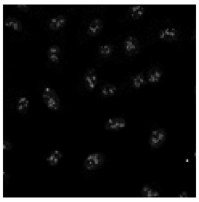
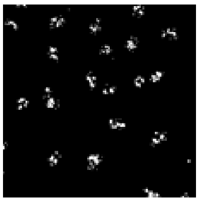
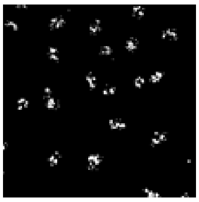
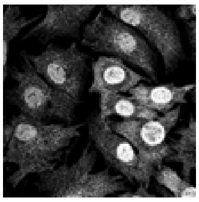
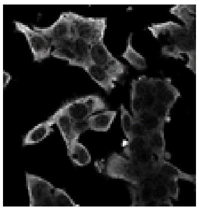
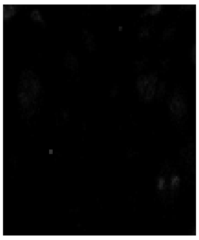
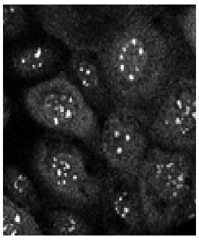
ID	Original Images	Binary Image using fixed threshold	Binary Image using Adaptive threshold
Image A	 59.73	 123.49	 74.83
Image B	 42.74	 96.94	 80.92
Image C	 4.79	 15.84	 52.80
Image D	 40.87	 123.33	 72.55
Image E	 12.57	 37.97	 37.89

Table 4.6: PSNR of binary image using fixed threshold v/s binary image using adaptive threshold

Image ID	**Fixed Threshold value**	**Adaptive threshold value**
 Image A	4.18	13.80
 Image B	12.28	14.35
 Image C	27.38	16.24
 Image D	7.3	13.56

In table 4.4 visual difference of original images with the images obtained by fixed and adaptive threshold has been shown.Images obtained from fixed threshold are more visibly different then adaptive threshold.Table 4.5 shows the standard deviation and table 4.6 shows the PSNR value to determine the signal to noise ratio.

4.2.2 Image Binarization

Using the threshold value we generated 3 binary images for each image using the equations mentioned in chapter 3. Below are the examples of 5 images which gives the clear comparison between the binary images using fixed threshold and binary images using adaptive threshold.

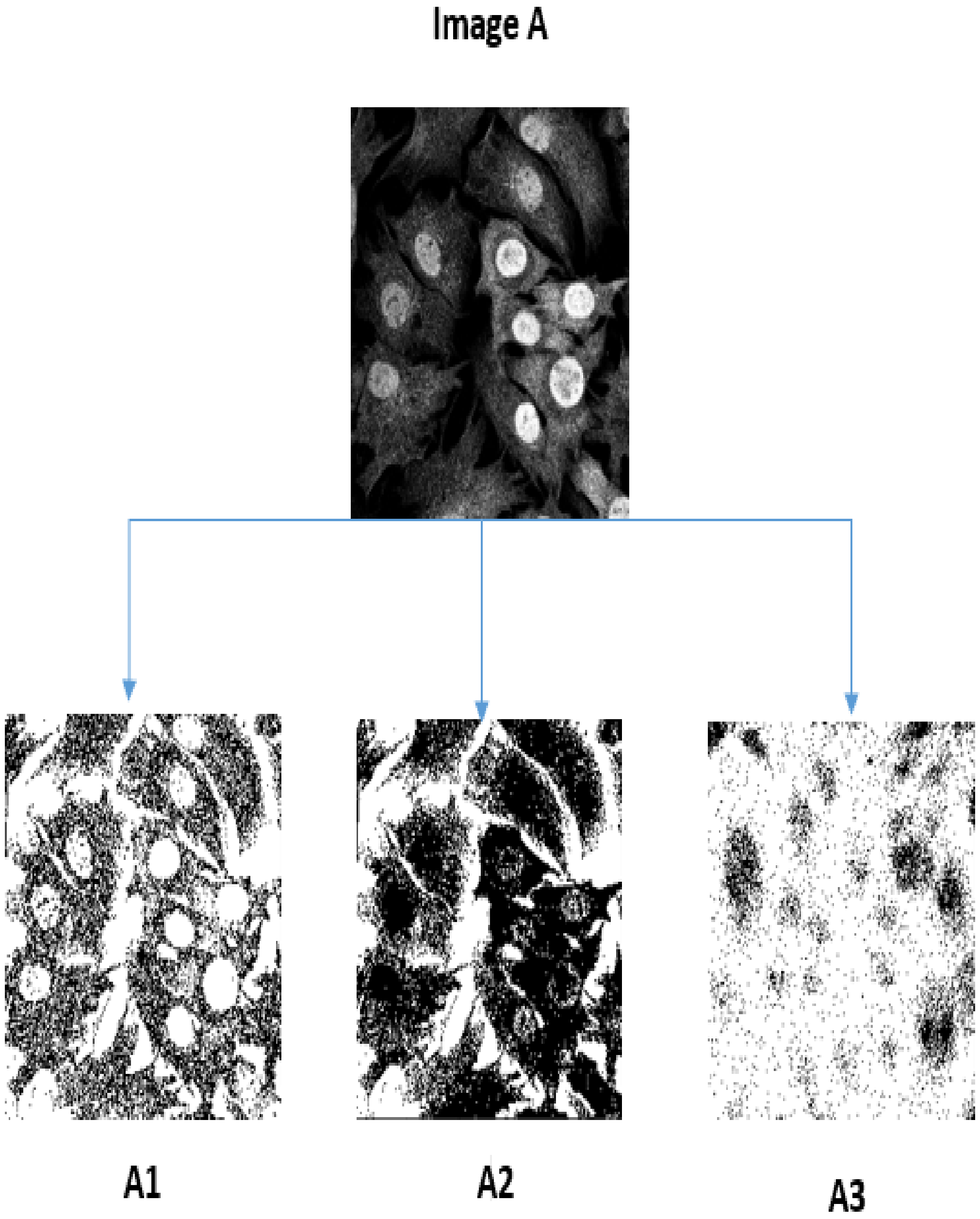


Figure 4.5: 3 Binary images of image A obtained by using equations with fixed threshold value 40.

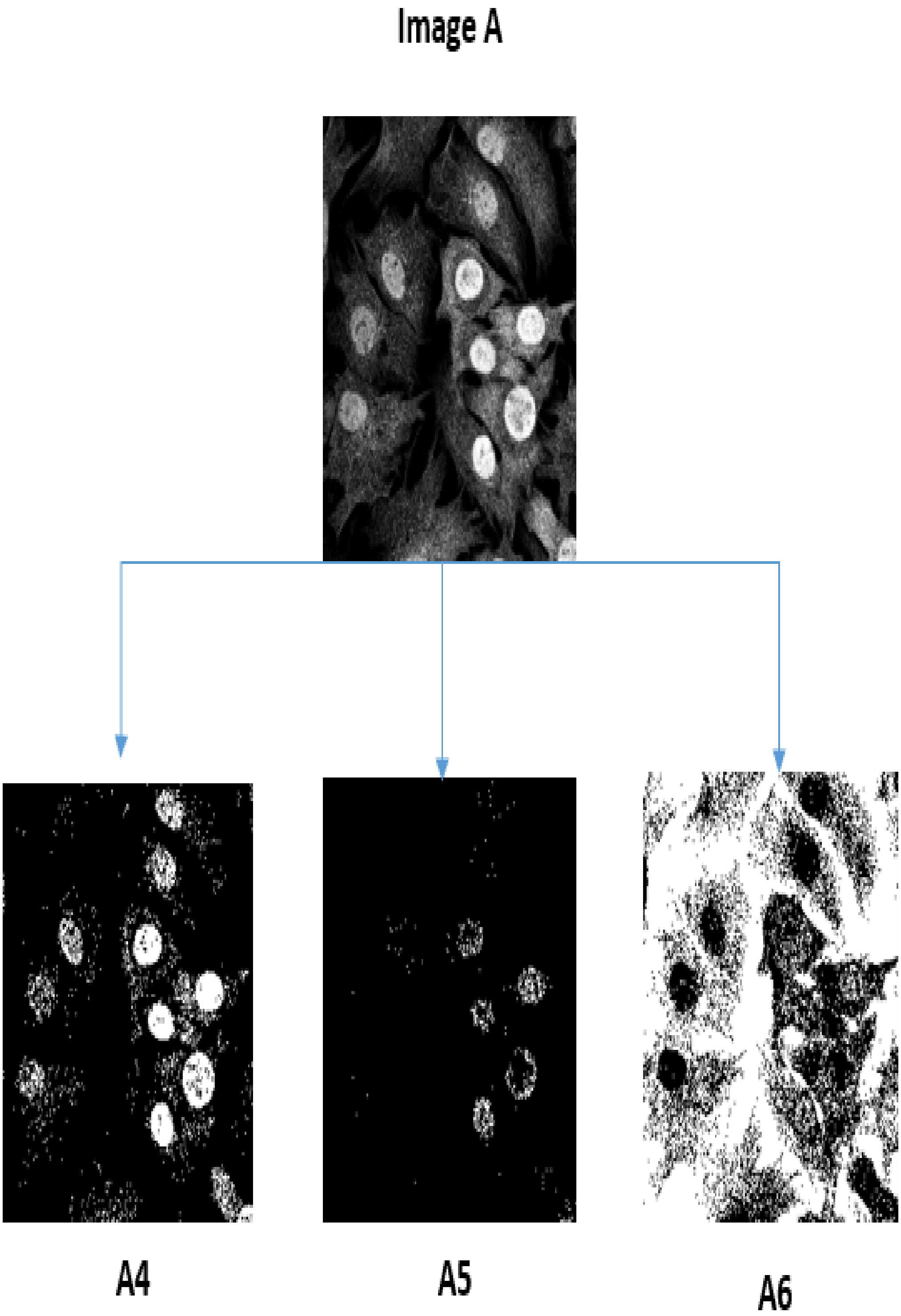


Figure 4.6: 3 Binary images of image A obtained by using equations with adaptive threshold value 89.

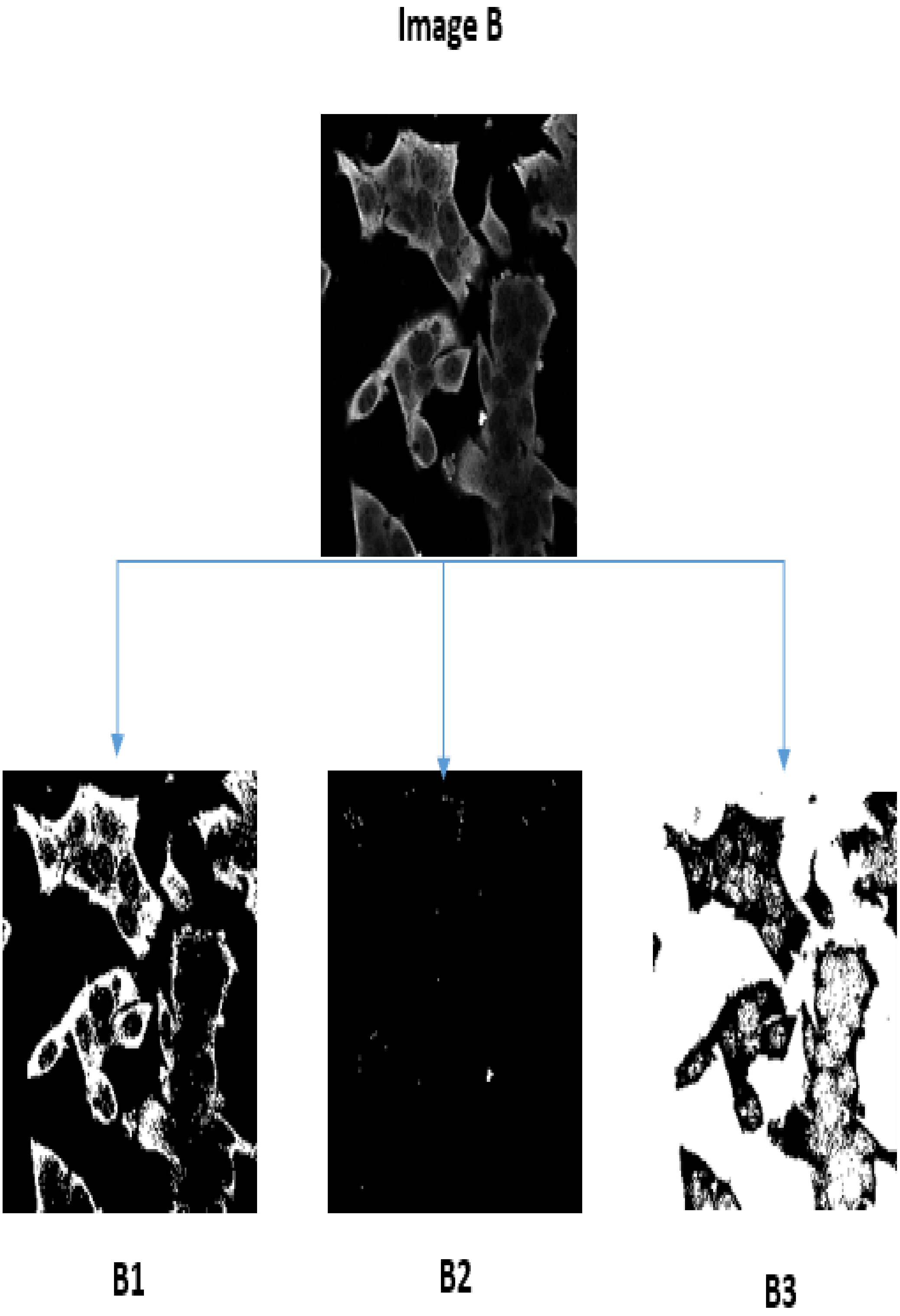


FIGURE 4.7: 3 Binary images of image B obtained by using equations with fixed threshold value 40.

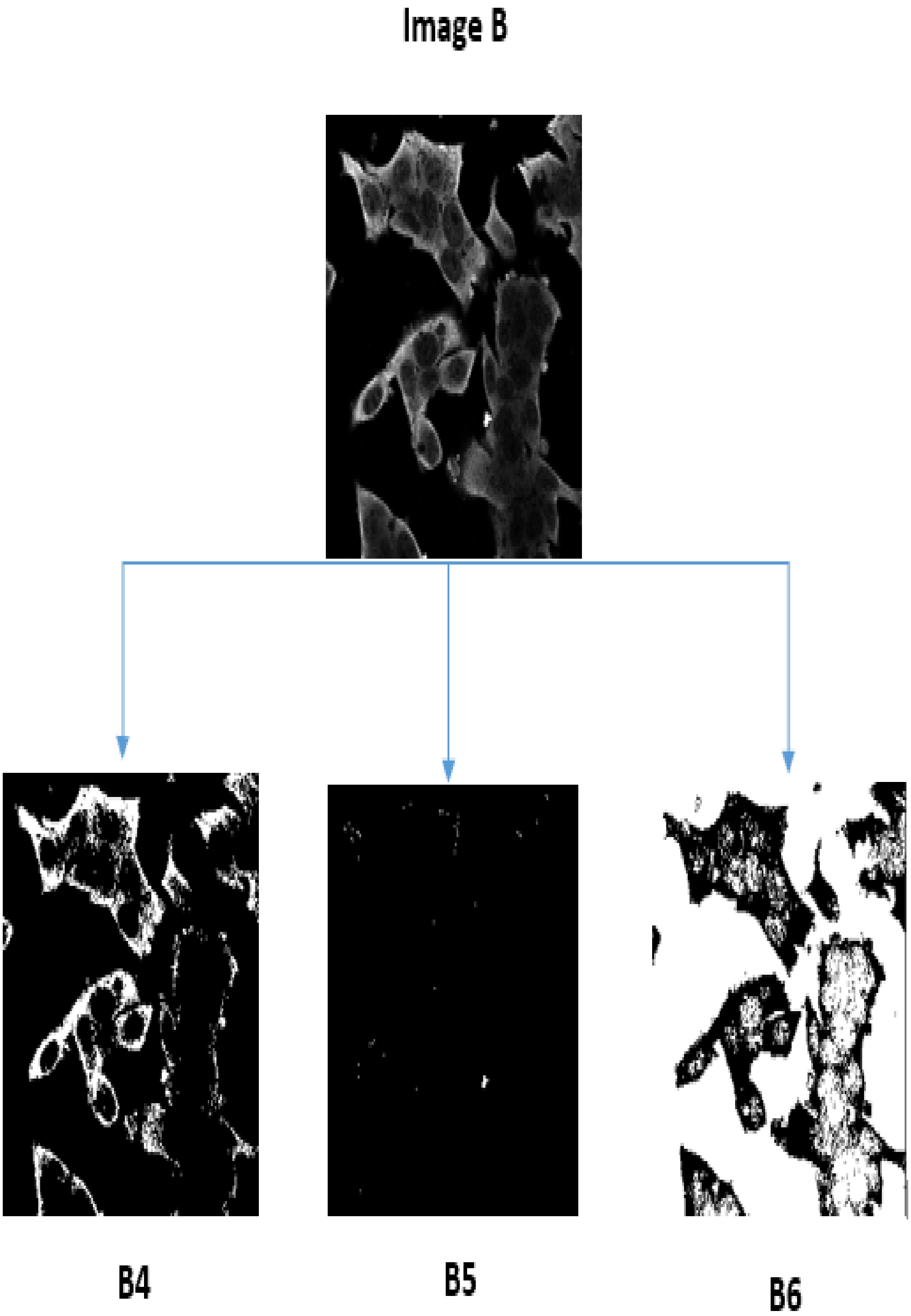


Figure 4.8: 3 Binary images of image B obtained by using equations with adaptive threshold value 54.

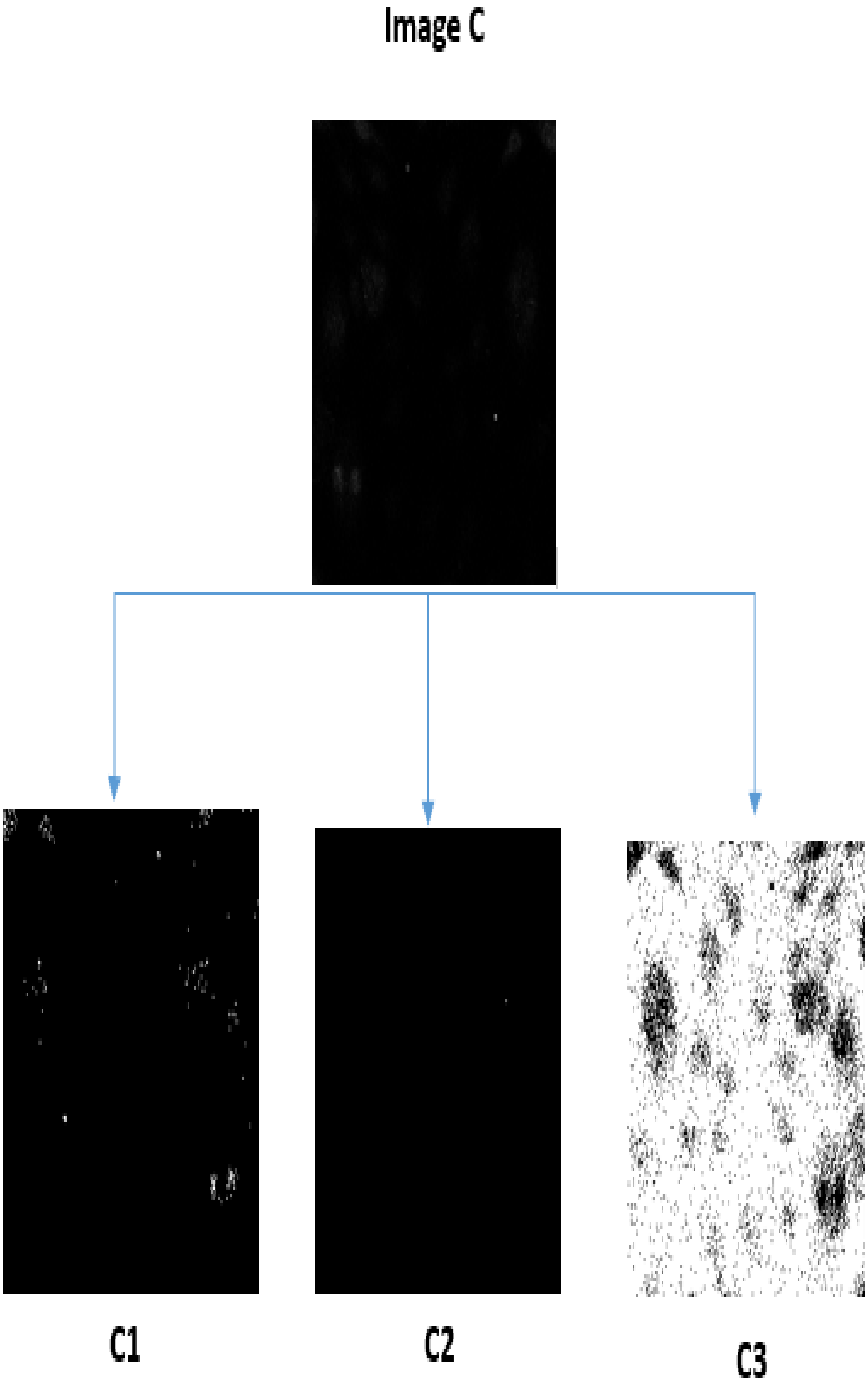


Figure 4.9: 3 Binary images of image C obtained by using equations with fixed threshold value 40.

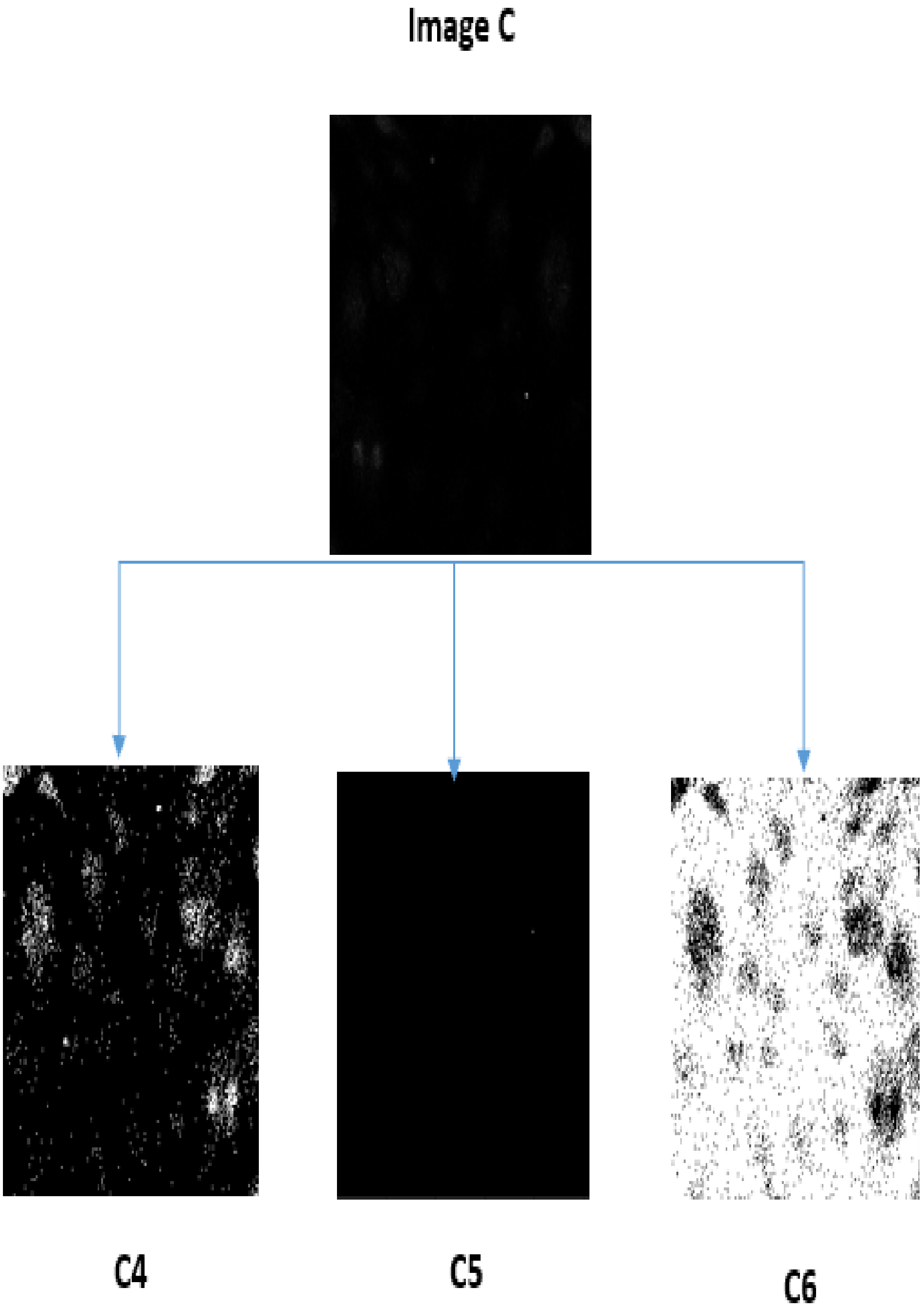


Figure 4.10: 3 Binary images of image C obtained by using equations with adaptive threshold value 8.

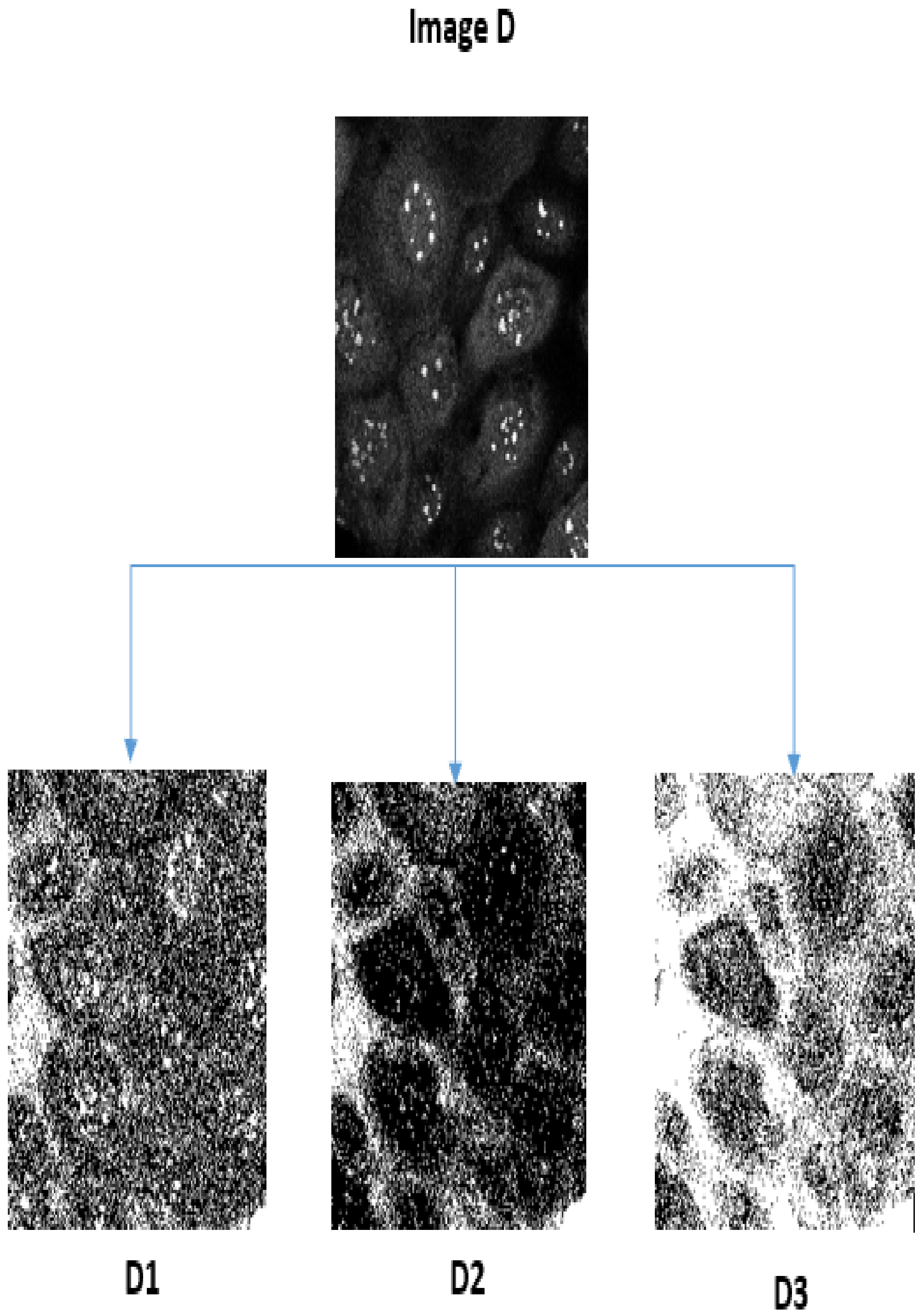


Figure 4.11: 3 Binary images of image D obtained by using equations with fixed threshold value 40.

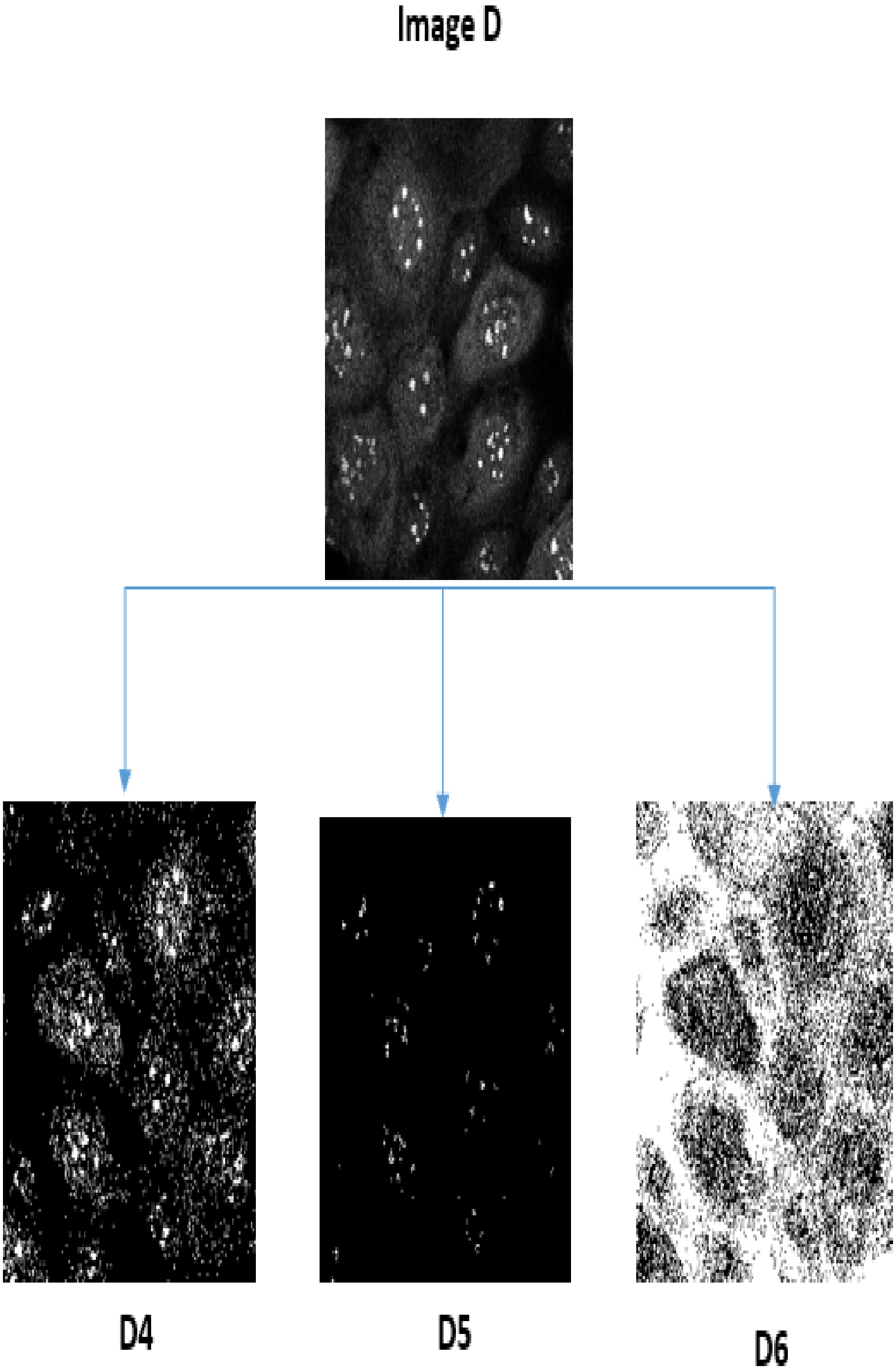


Figure 4.12: 3 Binary images of image D obtained by using equations with adaptive threshold value 51.

Table 4.7: F-Score of different techniques

Sr.	Techniques	F-Score
1	Team1- D.Shubin	0.593
2	Team2-J.Lan	0.571
3	Team3- P.Jinmo	0.570
4	Team4- C.Enkal et al.	0.567
5	Fixed Threshold	0.44
6	Proposed Methodology	**0.63**

After generating binary images we extracted features using these binary images. With the proposed methodology we extracted 7 features out of every binary image. In this manner we get 9 x 3 = 27 feature vector for each image.

4.3 Classification

After building a dataset based on feature extraction of images and performing balancing through MLSMOTE to overcome the imbalance dataset, comes the training phase. In order to predict Protein Subcellular Localization, we used ensemble classifier Random Forest for multilabel classification. We train the classifier on training dataset. After training phase the classifier predicts test dataset.. To evaluate we measure the f-score because of the imbalance dataset and achieved 0.633 which outperformed the f-score of state-of-the-art model on this dataset which is 0.593(figure 4.13).

4.4 Comparison

To evaluate our performance of our classifier we calculated the F-socre on test data using the Otsu's adaptive thresholding technique and compared it with f-score

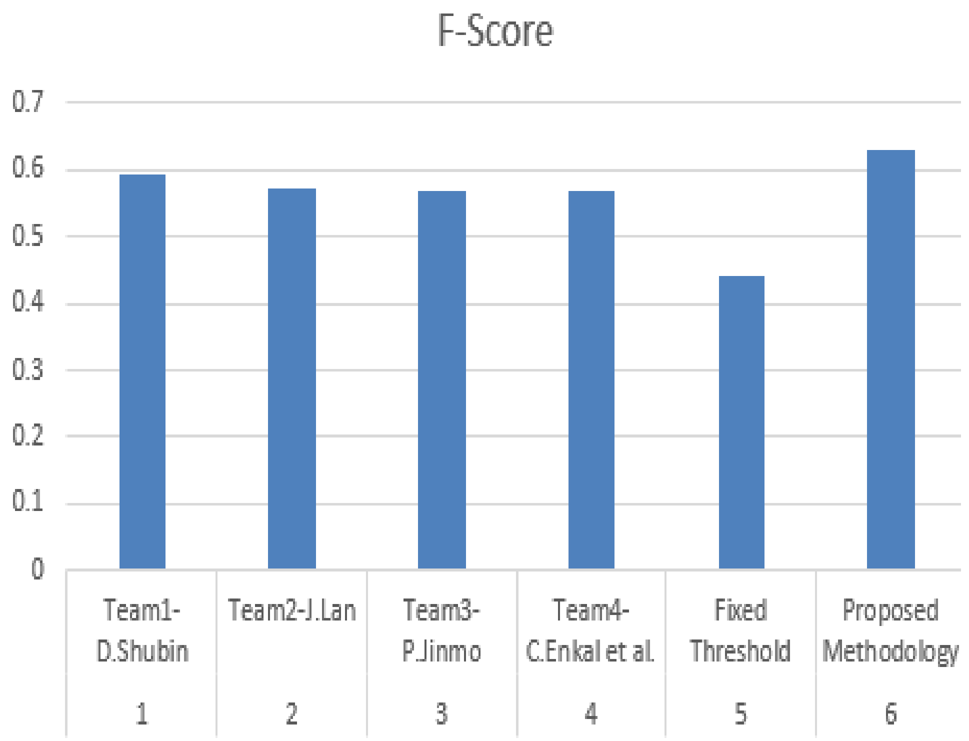


Figure 4.13: F-Score of different Techniques

using the fixed threshold. Along with this we also compared the performance of our technique with the top 4 ranked teams. Team 1 technique was to use optimized single neural network with the combination of loss function with a Lovasz loss term. Team 2 focus was on data preprocessing, Team 3 used automatic data augmentation and Teams 4 hybrid of different models. As shown in table 4.3 the highest f-score was produced by our technique and it out performed all the other techniques. With Otsu's adaptive threshold technique we achieved f-score of 0.63 whereas f-score on fixed threshold is 0.44 which is less than f-score through Otsu's adaptive thresholding technique.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

With the advancement of microscopic images, significant work has been done to predict protein subcellular localization. Up to 10,000 different kinds of proteins are synthesize by Eukaryotic cells which are destined for different organelles. It is crucial to understand protein subcellular localization for functional annotation of protein [32]. To perform its function protein has to be located to its pre-determined position hence it is very important to find the subcellular localization of protein. Protein miss location has proven to be the cause of several human diseases, such as Alzhrmeir's disease and cancer [16]. Different researchers have produced different methods to predict protein subcellular localization. Significant work has been done to predict Protein subcellular localization through amino acid sequencing. With the advancement in bio images, researchers are more focused on image based Protein subcellular localization and has produced different methodologies for this purpose. This research study also contribute towards the development of methodologies for predicting protein subcellular localization.

To predict protein subcellular localization, this study has proposed a methodology that uses adaptive threshold value to segment protein image and produce state-of-the-art result on Human Protein Atlas Dataset (HPA-Version 18). The methodology that this study offers is based on the Global feature method which is TAS to extract features from the image. Before Feature Extraction images are binarized first using the Otsu's adaptive thresholding technique which ensures to segment image according to the intensity variance of each image individually. By using adaptive threshold this study shows that it is better than using fixed threshold because while using fixed threshold, it may lose some important information or it may add noise to the segmented image. Adaptive threshold minimize this problem by performing effective segmentation and separate foreground image with background image efficiently. Hence this study use adaptive threshold value to generate three binarize image using three different ranges. In the feature extraction phase the, features are extracted from these binarize images using TAS in which pixels with 0-8 white neighboring pixels are calculated in binary images and by doing so it generates 3x9=27 feature vector of each original image. After completing the feature extraction phase this study use Multi label random forest classifier for training and then it evaluate this trained model on test dataset. Due to the imbalance dataset the performance measure that we used is macro f1-score and compare the results obtained from this proposed methodology with different techniques. This study also compare the results obtained by using fixed threshold on the same dataset. The macro f1-score obtained from the methodology proposed in this research study is 0.63 which outperformed all the other techniques.

Bibliography

[1] FrantišEk BalušKA, Dieter Volkmann, and Peter W Barlow. Eukaryotic cells and their cell bodies: cell theory revised. *Annals of Botany*, 94(1):9–32, 2004.

[2] Akiko Hatano, Hirokazu Chiba, Harry Amri Moesa, Takeaki Taniguchi, Satoshi Nagaie, Koji Yamanegi, Takako Takai-Igarashi, Hiroshi Tanaka, and Wataru Fujibuchi. Cellpedia: a repository for human cell information for cell studies and differentiation analyses. *Database*, 2011, 2011.

[3] Tibor Vellai and Gabor Vida. The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1428):1571–1577, 1999.

[4] Pierre Dönnes and Annette Höglund. Predicting protein subcellular localization: past, present, and future. *Genomics, proteomics & bioinformatics*, 2(4): 209–215, 2004.

[5] Wei Ouyang, Casper F Winsnes, Martin Hjelmare, Anthony J Cesnik, Lovisa Åkesson, Hao Xu, Devin P Sullivan, Shubin Dai, Jun Lan, Park Jinmo, et al. Analysis of the human protein atlas image classification competition. *Nature methods*, 16(12):1254–1261, 2019.

[6] Wei Shao, Mingxia Liu, Ying-Ying Xu, Hong-Bin Shen, and Daoqiang Zhang. An organelle correlation-guided feature selection approach for classifying multi-label subcellular bio-images. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(3):828–838, 2017.

[7] Detlev Bannasch, Alexander Mehrle, Karl-Heinz Glatting, Rainer Pepperkok, Annemarie Poustka, and Stefan Wiemann. Lifedb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. *Nucleic Acids Research*, 32(suppl_1): D505–D508, 2004.

[8] Muhammad Tahir, Asifullah Khan, Abdul Majid, and Alessandra Lumini. Subcellular localization using fluorescence imagery: Utilizing ensemble classification with diverse feature extraction strategies and data balancing. *Applied Soft Computing*, 13(11):4231–4243, 2013.

[9] Muhammad Tahir and Asifullah Khan. Protein subcellular localization of fluorescence microscopy images: employing new statistical and texton based image features and svm based ensemble classification. *Information Sciences*, 345:65–80, 2016.

[10] Afzal Godil, Zhouhui Lian, and Asim Wagan. Exploring local features and the bag-of-visual-words approach for bioimage classification. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 694–695, 2013.

[11] Luis Pedro Coelho, Joshua D Kangas, Armaghan W Naik, Elvira Osuna-Highley, Estelle Glory-Afshar, Margaret Fuhrman, Ramanuja Simha, Peter B Berget, Jonathan W Jarvik, and Robert F Murphy. Determining the subcellular location of new proteins from microscope images using local features. *Bioinformatics*, 29(18):2343–2349, 2013.

[12] Ting Zhao, Meel Velliste, Michael V Boland, and Robert F Murphy. Object type recognition for automated analysis of protein subcellular location. *IEEE transactions on image processing*, 14(9):1351–1359, 2005.

[13] Tanel Pärnamaa and Leopold Parts. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *G3: Genes, Genomes, Genetics*, 7(5):1385–1392, 2017.

[14] Mengli Xiao, Xiaotong Shen, and Wei Pan. Application of deep convolutional neural networks in classification of protein subcellular localization with microscopy images. *Genetic epidemiology*, 43(3):330–341, 2019.

[15] Mien-Chie Hung and Wolfgang Link. Protein localization in disease and therapy. *Journal of cell science*, 124(20):3381–3392, 2011.

[16] Hakan Wieslander, Gustav Forslid, Ewert Bengtsson, Carolina Wahlby, Jan-Michael Hirsch, Christina Runow Stark, and Sajith Kecheril Sadanandan. Deep convolutional neural networks for detecting cellular changes due to malignancy. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 82–89, 2017.

[17] Wei Shao, Yi Ding, Hong-Bin Shen, and Daoqiang Zhang. Deep model-based feature extraction for predicting protein subcellular localizations from bioimages. *Frontiers of Computer Science*, 11(2):243–252, 2017.

[18] Ying-Ying Xu, Fan Yang, Yang Zhang, and Hong-Bin Shen. An image-based multi-label human protein subcellular localization predictor (i locator) reveals protein mislocalizations in cancer tissues. *Bioinformatics*, 29(16):2032–2040, 2013.

[19] Ying-Ying Xu, Fan Yang, Yang Zhang, and Hong-Bin Shen. Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning. *Bioinformatics*, 31(7):1111–1119, 2015.

[20] Muhammad Tahir, Bismillah Jan, Maqsood Hayat, Shakir Ullah Shah, and Muhammad Amin. Efficient computational model for classification of protein localization images using extended threshold adjacency statistics and support vector machines. *Computer methods and programs in biomedicine*, 157:205–215, 2018.

[21] Muhammad Tahir, Asifullah Khan, and Abdul Majid. Protein subcellular localization of fluorescence imagery using spatial and transform domain features. *Bioinformatics*, 28(1):91–97, 2012.

[22] Sebastian Briesemeister, Jörg Rahnenführer, and Oliver Kohlbacher. Going from where to why—interpretable prediction of protein subcellular localization. *Bioinformatics*, 26(9):1232–1238, 2010.

[23] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347 (6220), 2015.

[24] Lu Zhu, Ralf Hofestädt, and Martin Ester. Tissue-specific subcellular localization prediction using multi-label markov random fields. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1471–1482, 2019.

[25] Nicholas A Hamilton, Radosav S Pantelic, Kelly Hanson, and Rohan D Teasdale. Fast automated cell phenotype image classification. *BMC bioinformatics*, 8(1):1–8, 2007.

[26] Devin P Sullivan, Casper F Winsnes, Lovisa Åkesson, Martin Hjelmare, Mikaela Wiking, Rutger Schutten, Linzi Campbell, Hjalti Leifsson, Scott Rhodes, Andie Nordgren, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nature biotechnology*, 36(9):820–828, 2018.

[27] Armaghan W Naik, Joshua D Kangas, Devin P Sullivan, and Robert F Murphy. Active machine learning-driven experimentation to determine compound effects on protein patterns. *Elife*, 5:e10047, 2016.

[28] Jin-Xian Hu, Ying-Ying Xu, Hong-Bin Shen, et al. Deep learning-based classification of protein subcellular localization from immunohistochemistry images. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 599–604. IEEE, 2017.

[29] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[30] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015.

[31] Wei Shao, Mingxia Liu, and Daoqiang Zhang. Human cell structure-driven model construction for predicting protein subcellular location from biological images. *Bioinformatics*, 32(1):114–121, 2016.

[32] Yuki Shimahara, Ko Sugawara, Kei H Kojo, Hiroki Kawai, Yuya Yoshida, Seiichiro Hasezawa, and Natsumaro Kutsuna. Imacel: A cloud-based bioimage analysis platform for morphological analysis and image classification. *PloS one*, 14(2):e0212619, 2019.