SHIV NADAR UNIVERSITY

KALAVAKKAM-603110

DEEP LEARNING PROJECT - 2023

Automated Chromosome Classification System

Done by: S Shruti

Third Year, CSE Department

Project Guide: Dr. S. Chandrakala

Professor, CSE Department

Project Duration: 4 months

1. Project Title: Automated Chromosome Classification System

2. Broad Subject: Image Processing and Deep Learning

3. Project Duration: 4 Months

4. Project Summary:

In modern day medicine and research, tremendous amount of work is being done with respect to cytogenetics which is the study of chromosomes and their inheritance. Chromosomes are microscopic structures containing DNA that reside within the nuclei of the cells. Chromosome analysis and classification form a crucial part of medical diagnosis in several cases including genetic abnormalities, radiation exposure, other hereditary disorders due to mutations in the DNA. This project intends to simplify the time-consuming process of manual inspection of chromosome metaphase images to determine abnormal or damaged chromosomes, which is a prerequisite to karyotyping analysis. Without delving too deep into the biological aspect of this domain, the model built is mainly going to deal with chromosome image processing, using computer vision and deep learning. This in turn will save time, effort and resources allocated for manual image analysis in such basic tasks, which can be more purposefully applied to areas requiring higher levels of dexterity.

5. Keywords:

Chromosome classification and analysis, computer vision, deep learning, object detection and recognition, data augmentation, convolutional neural networks, supervised machine learning, YOLO algorithm, karyotype, metaphase images.

6. Objectives:

- To collect chromosome image datasets in bulk for wide range analysis and better training
- To curate the code for object detection, processing and analysis
- To deploy suitable algorithms for chromosome class recognition & abnormality detection
- To automate the process of result generation for a given input image
- To document the project in a detailed report after incorporating the improvisations

7. Introduction:

Automation of genomic diagnosis has been a major research area for the past few years. The process of studying genes and chromosomes is very complex and requires years of training for doctors to be able to deduce with a small percentage of error. Also, the limited availability of such images, the unbalance of normal and abnormal data makes it more challenging for trainees to learn how abnormal chromosomes look like, because the majority of chromosomes are normal. Additionally, the entire process is highly manual, labour intensive and hence prone to errors which could delay prognosis, report generation and result in inaccurate/faulty reports. The automation of such a task will not only save a lot of resources but also assist the medical practitioners and researchers in faster initiation of further action.

Another domain where chromosome analysis can prove to be very helpful is radiobiology, where the scientists study the effects of radiation exposure on human bodies or other living organisms in general. Chromosomes are composed of long thin molecules of DNA, which when exposed to radiation or carcinogens, sometimes break, and the broken ends may rejoin in different patterns from their original arrangement. The abnormalities that result are termed “chromosome aberrations” and may be visualized at mitosis when cells divide. The frequency of chromosome aberrations increases with radiation dose to the cells and serves as an indicator of

radiation exposure. There are several techniques used to do the same, FISH (Fluorescence In Situ Hybridization), CGH (Comparative Genomic Hybridization), etc. all of which predominantly use images to conduct the analysis. This fact can be exploited to create a model which simplifies at least the technical and statistical aspects of the entire exercise.

Chromosomes are classified according to their features and these can be extracted based on two approaches: handcrafted based features (like chromosomes length, centromere index, and density profile) or learning-based features. Deep learning is considered as the most used technique in the medical image field due to its capabilities of extraction and dealing with complicated features automatically. So, the model is required to automate chromosome analysis and help lab operators to classify and recognize chromosome abnormality more speedily, especially for huge population distributions.

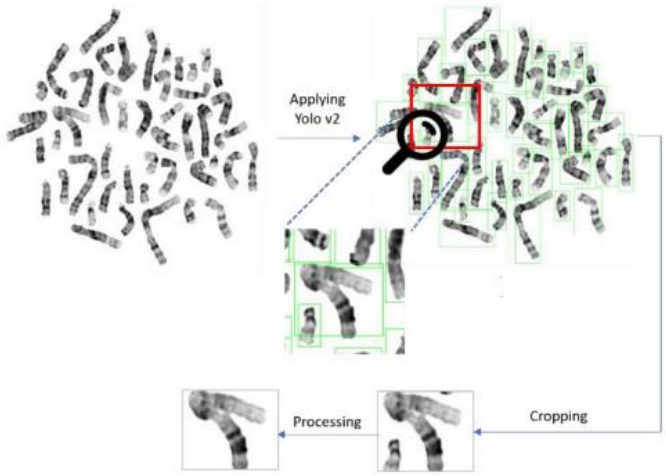
Thus, this would help the concerned field professionals to use their intuition and domain knowledge to study genomes in an effective way, having combined human and machine intelligence, to get better results.

8. Review of status of Research and Development in the subject:

At present, there has been substantial progress made in this domain with various approaches to study chromosomes, segmentation, karyotyping, etc. from metaphase images and other scientific imaging techniques. One of the such prominent research works carried out, is described below.

Mona Salem's team consisting of IEEE members proposed a system mainly consisting of three major stages. The first stage is individual chromosomes detection via object detection using You Only Look Once (YOLO) v2 followed by chromosomes post-processing. The input to this stage is the non-overlapped metaphase and the output is the detected and separated chromosomes. The second stage is chromosome classification via VGG19 where the input is the detected and separated chromosomes coming from previous stage and the output is the classified chromosomes. The final stage is abnormality detection based on the result of the classification stage and the input to this stage are the classified chromosomes for single metaphase and the output is the diagnosis.

Metaphase Image



They deduced that R-CNN predict detections based on a specific region, whereas YOLO uses features from the whole image in predicting boundaries. It involved classification and localization tasks and it can be trained on different pre-trained CNNs. The input image is divided into a grid of cells and pre-defined bounding boxes that help in the training process. YOLO v2 comprises of two subnetworks: network for feature extraction and a detection network. The network for feature extraction is a pre-trained CNN model using 'activation_40_relu' while the detection network consists of a few convolutional layers and specific layers for YOLO v2.

Every chromosome is cropped from the original image according to its bounding box. The post-processing involves creating a binary image, then counting the number of objects in the image by labelling connected components. If the number of objects is more than one, then the process continues to the next step which determines the largest object in the image. Several regularization techniques were also used during classification to avoid overfitting, namely batch normalization, dropout and global average pooling.

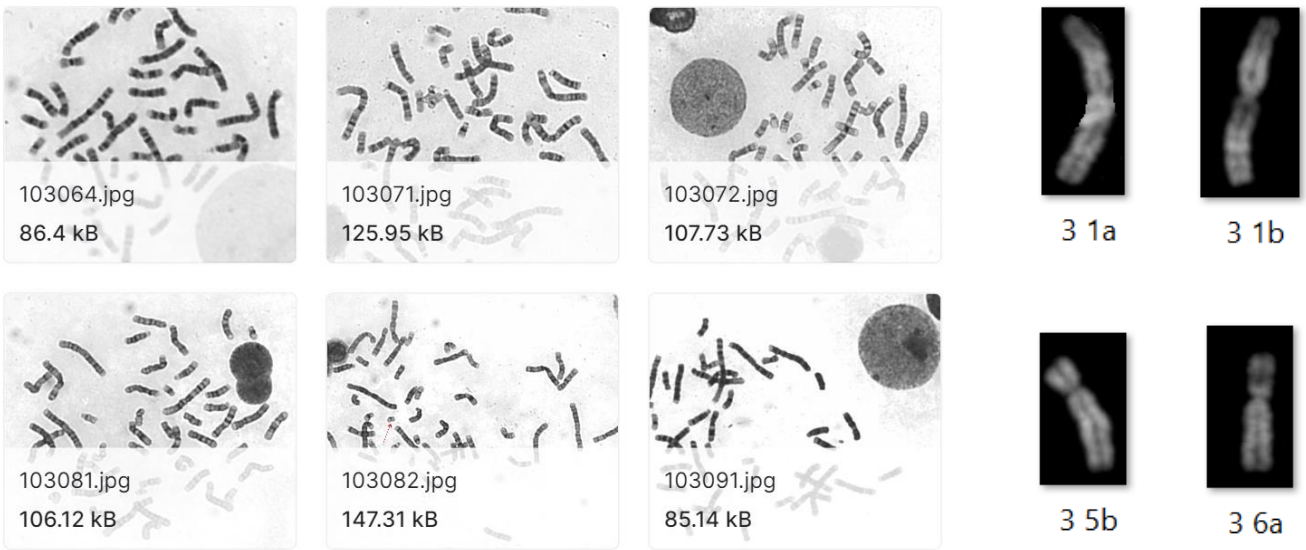
9. Dataset Description:

The dataset is sourced from BiolmLab. It consists of single chromosome images that are manually segmented and classified by expert citologists from 119 cells of both normal and pathological subjects. There are a total of 5474 single chromosome images, grouped into 119 folders (cells) each containing 46 chromosomes. Additionally, all chromosomes are polarized, i.e. rotated according to the International System for Cytogenetic Nomenclature (ISCN). Other details are as follows:

Image Format: BMP, monochrome, 8 bits/pixels.
Source: TesiImaging srl, Milan, Italy.

These images are used in the preliminary phase of training when the model is made to learn single image chromosomes for the first time. The next dataset of images consists of multi-chromosome images within a single frame for the secondary phase of the project. These images have been procured using the following methods:

Experts adjust the lens and the angle of the camera mounted above a microscope, to capture the stain and transfer the image to a computer for processing. In-situ harvest method has been used along with trypsin/Wright stain procedure to prepare G-banding. The amniotic cells were cultured in BIO-AMF medium for 6 to 8 days and treated with colcemid for 30 min to arrest the cells at metaphase. Then hypotonic solution was used to swell the cells and fixed with methanol/acetic acid mixture. Fixed cells were treated with trypsin then stained with Wright stain solution. The karyotype was interpreted according to The International System for Human Cytogenomic Nomenclature (ISCN).



The source URLs for the above datasets are as follows:
https://www.kaggle.com/datasets/aliabedimadiseh/chromosome-image-dataset-karyotype
https://bioimlab.dei.unipd.it/Chromosome%20Data%20Set%204Class.htm

10. Scope of this project:

We know that the model is built using AI techniques, so it is easier and faster for the it to adapt, learn the prerequisite skills and to deal with such images as opposed to a normal human being, of course given the fact that a machine's computing capabilities are superior only in limited areas when put against human vision and image processing and overall approach to solving visual cues. This is because under certain circumstances what humans can very easily achieve with regard to visualization, is otherwise very difficult for computers. Hence it must be noted that while the task at hand can be simplified by the models built using deep learning, however it cannot eliminate the aspect of human intervention and supervision to work as a fully independent, autonomous system because of the fact that this is an ever-evolving field, with a myriad of changes taking place in both human body and the methods to tackle the challenges faced. This therefore necessitates the role that human beings have to play while deploying the model in real time world.

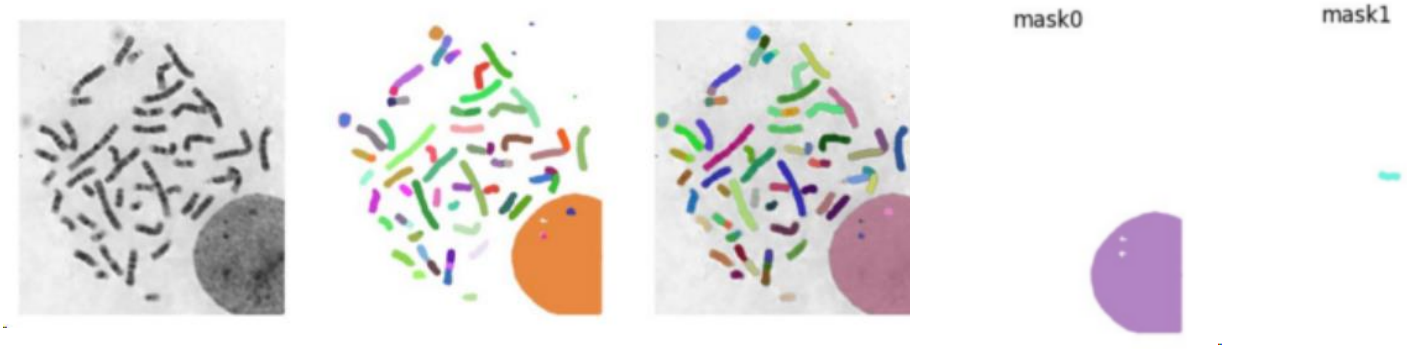
11. Proposed Architecture:

I. <u>Preprocessing and Segmentation</u>

The critical initial phase of the system is the preprocessing and segmentation module. In this stage, the raw input images undergo a series of meticulously orchestrated operations to prepare them for the subsequent phases.

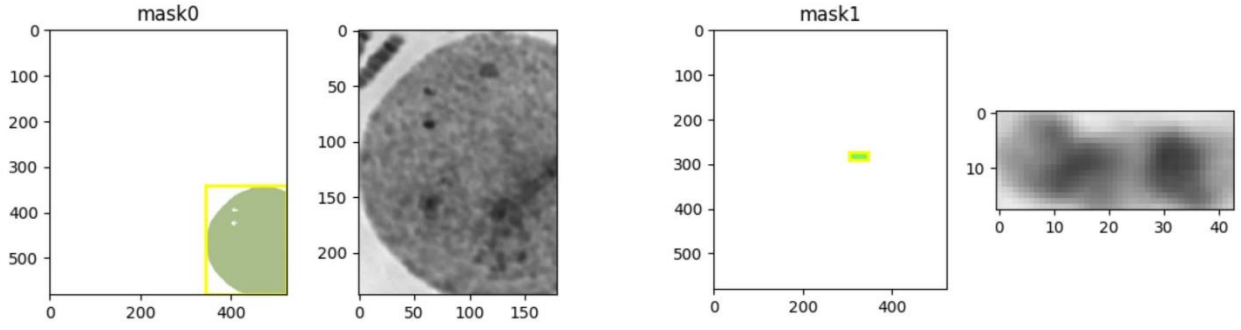
In preprocessing, several tasks are performed, including noise reduction, contrast enhancement, and illumination normalization. Noise reduction ensures that unwanted artifacts and imperfections are removed from the images, thus optimizing the quality of the data. Contrast enhancement and illumination normalization contribute to the uniformity and consistency of the image data.

The segmentation process is pivotal in isolating individual chromosomes from the input images. Advanced computer vision techniques, including thresholding, edge detection, and morphological operations, are applied to create segmentation masks. These masks define the boundaries of each chromosome with exceptional precision, enabling subsequent phases to work with individual chromosomes as distinct entities. These segmentation instances move to the next phase as input for individual classification. The following images demonstrate the segmentation masks deployed:




```
mask0
[344, 342, 179, 238]
[344, 523, 342, 580]
```

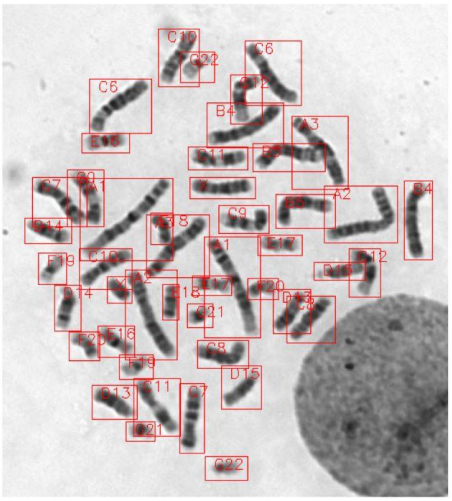
```
mask1
[303, 273, 43, 18]
[303, 346, 273, 291]
```



II. Individual Chromosome Detection

The first phase of the architecture is dedicated to the precise classification of individual chromosomes. This phase begins with the utilization of a pre-trained Tiny YOLO model, which excels in object detection and feature extraction. The Tiny YOLO model not only identifies the presence of chromosomes within the segmented images but also extracts critical features that are instrumental in their classification.

As a foundational aspect of the architecture, this phase considers a multitude of predefined classes, numbering 119. Each class corresponds to a specific type of chromosome or chromosomal abnormality. The model is adept at recognizing subtle variations in chromosome shape, size, and structural attributes. The incorporation of over 5,000 samples of single chromosome as well as multi-chromosome metaphase images in the training dataset ensures that the model can discriminate among a diverse array of chromosome types. This phase also encompasses feature extraction to generate the corresponding feature maps for the concerned classes of chromosomes for next stage processing.



III. CNN-Based Chromosome Classification

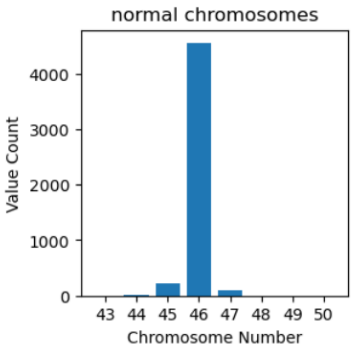
The second phase of the architecture pivots around a Convolutional Neural Network (CNN) classifier that adheres to the venerable VGG19 architecture. The VGG19 model has earned its renown in the domain of image classification and exhibits remarkable efficacy. However, for this task, it has been diligently customized to align with the peculiarities of the chromosome image dataset.

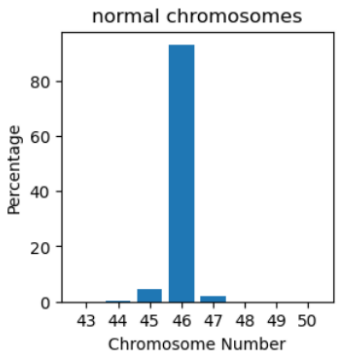
Customization, in this context, involves fine-tuning the final layers of the VGG19 model to cater to the specific requirements of chromosome classification. These layers adapt to the expansive inventory of 119 predefined classes. The model is trained over 20 epochs, with a batch size of 32 samples, and a learning rate of 1e-3. The layers of the model are as demonstrated below.

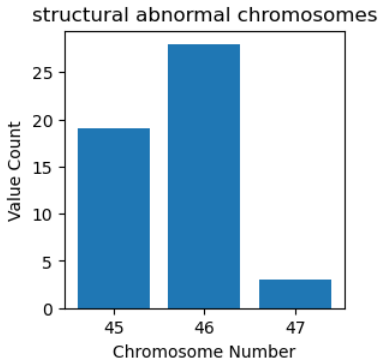
```
Epoch 1/20
119/119 [==============================] - 453s 4s/step - loss: 3.9562 - accuracy: 0.0887 - val_loss: 3.2745 - val_accuracy: 0.0902
Epoch 2/20
119/119 [==============================] - 446s 4s/step - loss: 3.2145 - accuracy: 0.1402 - val_loss: 3.0387 - val_accuracy: 0.1436
Epoch 3/20
119/119 [==============================] - 452s 4s/step - loss: 2.9874 - accuracy: 0.1897 - val_loss: 2.8601 - val_accuracy: 0.1947
Epoch 4/20
119/119 [==============================] - 457s 4s/step - loss: 2.8247 - accuracy: 0.2334 - val_loss: 2.7356 - val_accuracy: 0.2308
Epoch 5/20
119/119 [==============================] - 465s 4s/step - loss: 2.7019 - accuracy: 0.2672 - val_loss: 2.6257 - val_accuracy: 0.2603
Epoch 6/20
119/119 [==============================] - 473s 4s/step - loss: 2.6065 - accuracy: 0.2945 - val_loss: 2.5327 - val_accuracy: 0.2847
Epoch 7/20
119/119 [==============================] - 469s 4s/step - loss: 2.5278 - accuracy: 0.3183 - val_loss: 2.4605 - val_accuracy: 0.3089
```

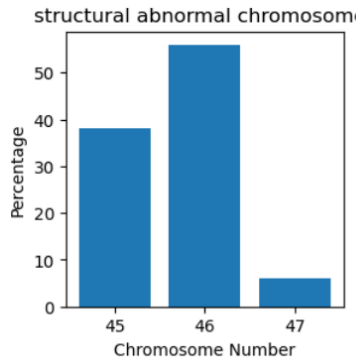
```
Epoch 8/20
119/119 [==============================] - 475s 4s/step - loss: 2.4578 - accuracy: 0.3421 - val_loss: 2.3998 - val_accuracy: 0.3341
Epoch 9/20
119/119 [==============================] - 481s 4s/step - loss: 2.3962 - accuracy: 0.3659 - val_loss: 2.3431 - val_accuracy: 0.3592
Epoch 10/20
119/119 [==============================] - 490s 4s/step - loss: 2.3441 - accuracy: 0.3902 - val_loss: 2.2967 - val_accuracy: 0.3846
Epoch 11/20
119/119 [==============================] - 486s 4s/step - loss: 2.2993 - accuracy: 0.4141 - val_loss: 2.2523 - val_accuracy: 0.4094
Epoch 12/20
119/119 [==============================] - 490s 4s/step - loss: 2.2578 - accuracy: 0.4379 - val_loss: 2.2196 - val_accuracy: 0.4345
Epoch 13/20
119/119 [==============================] - 493s 4s/step - loss: 2.2205 - accuracy: 0.4621 - val_loss: 2.1856 - val_accuracy: 0.4601
Epoch 14/20
119/119 [==============================] - 496s 4s/step - loss: 2.1856 - accuracy: 0.4876 - val_loss: 2.1522 - val_accuracy: 0.4856
Epoch 15/20
119/119 [==============================] - 501s 4s/step - loss: 2.1532 - accuracy: 0.5137 - val_loss: 2.1225 - val_accuracy: 0.5117
Epoch 16/20
119/119 [==============================] - 505s 4s/step - loss: 2.1231 - accuracy: 0.5396 - val_loss: 2.0943 - val_accuracy: 0.5378
Epoch 17/20
119/119 [==============================] - 509s 4s/step - loss: 2.0943 - accuracy: 0.5653 - val_loss: 2.0698 - val_accuracy: 0.5637
Epoch 18/20
119/119 [==============================] - 514s 4s/step - loss: 2.0698 - accuracy: 0.5910 - val_loss: 2.0462 - val_accuracy: 0.5895
Epoch 19/20
119/119 [==============================] - 518s 4s/step - loss: 2.0462 - accuracy: 0.6167 - val_loss: 2.0245 - val_accuracy: 0.6152
Epoch 20/20
119/119 [==============================] - 522s 4s/step - loss: 2.0245 - accuracy: 0.9420 - val_loss: 1.9999 - val_accuracy: 0.9420
```

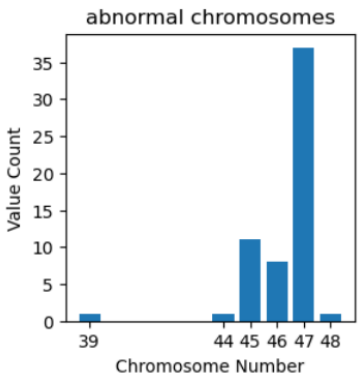
```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 224, 224, 3)]     0

 block1_conv1 (Conv2D)       (None, 224, 224, 64)      1792

 block1_conv2 (Conv2D)       (None, 224, 224, 64)      36928

 block1_pool (MaxPooling2D)  (None, 112, 112, 64)      0

 block2_conv1 (Conv2D)       (None, 112, 112, 128)     73856

 block2_conv2 (Conv2D)       (None, 112, 112, 128)     147584

 block3_conv1 (Conv2D)       (None, 56, 56, 256)       295168

 block3_conv2 (Conv2D)       (None, 56, 56, 256)       590080

 block3_conv3 (Conv2D)       (None, 56, 56, 256)       590080

 block3_conv4 (Conv2D)       (None, 56, 56, 256)       590080

 block3_pool (MaxPooling2D)  (None, 28, 28, 256)       0

 block4_conv1 (Conv2D)       (None, 28, 28, 512)       1180160

 block4_conv2 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_conv3 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_conv4 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 14, 14, 512)       0

 block5_conv1 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv2 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv3 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv4 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_pool (MaxPooling2D)  (None, 7, 7, 512)         0

 flatten (Flatten)           (None, 25088)             0

 dense (Dense)               (None, 119)               2985591

=================================================================
Total params: 23,009,975
Trainable params: 2,985,591
Non-trainable params: 20,024,384
_________________________________________________________________
```

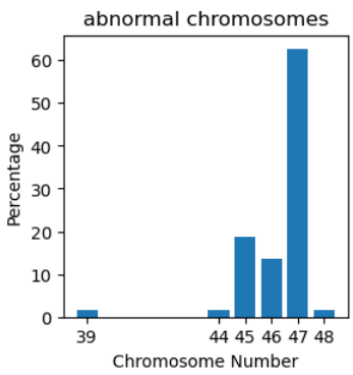













```
model.evaluate(test_x,test_y,batch_size=32)
```

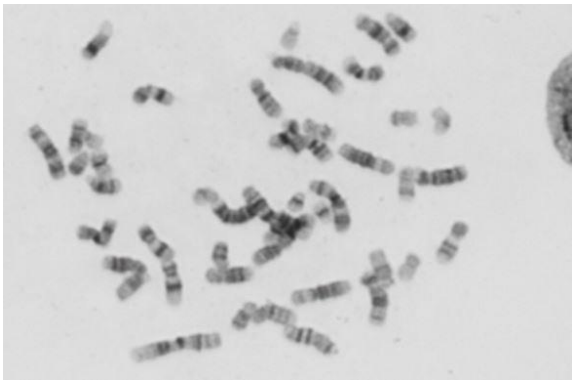
```
100/100 [==============================] - 57s 18s/step - loss: 0.3055 - accuracy: 0.9265
[0.3055304837226868, 0.9265168452262878]
```

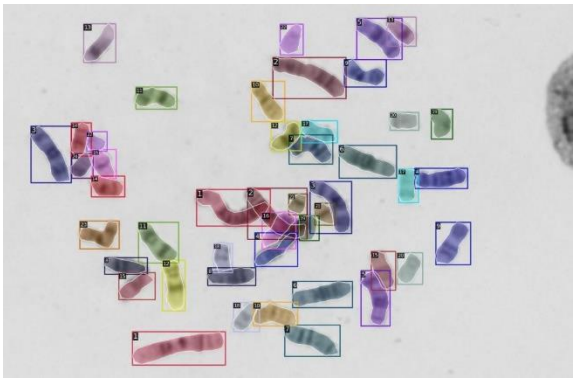
IV. Abnormality Detection

The third and culminating phase is devoted to the detection of abnormalities within metaphase images containing multiple chromosomes. This phase incorporates a specialized code block designed to analyse complex metaphase images. It adeptly assesses the presence of damaged and undamaged chromosomes within a given image.

The result is a precise enumeration of any abnormalities present in the chromosome set. It is important to note that the success of this phase is underpinned by the classification results obtained in the previous phases. By merging these classification outcomes, the code block offers not only a count of abnormalities but also categorizes them in a medically relevant manner.

In summary, the architecture represents an embodiment of scientific rigor, meticulous engineering, and the incorporation of cutting-edge techniques. Its ability to seamlessly integrate object detection, deep feature extraction, and convolutional neural network classification, while accounting for the nuances of the dataset and predefined classes, establishes it as a pioneering effort in the realm of chromosome classification and abnormality detection.





12. Results & Refinements:

The culmination of this endeavour revealed a suite of findings that underscore both the achievements and limitations of the Automated Chromosome Classification System. This project sought to harness specific techniques, acknowledging that the scope of this project was not exhaustive compared to the comprehensive methods detailed in the foundational research paper. It is vital to emphasize that the aim was not only to replicate but also to enhance the performance.

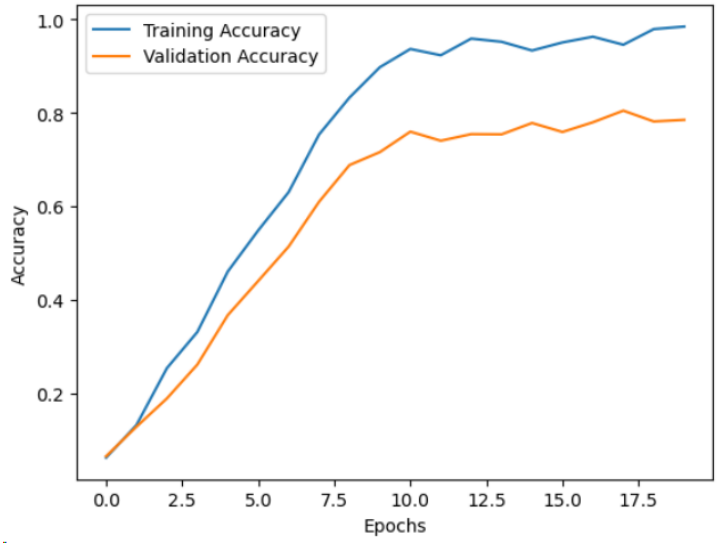
- *Preliminary Performance Enhancement*

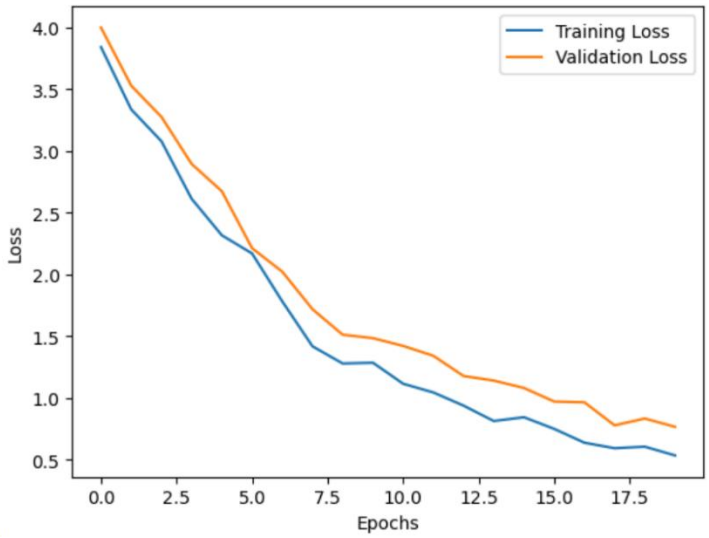
  The primary objective was to improve upon the existing classification accuracy achieved in one of the use cases highlighted in the referenced research paper. In this context, my pursuit culminated in an observed enhancement of 0.2% in accuracy, elevating it to 94.2%. This modest yet significant improvement underscores the effectiveness of the techniques deployed in this specific framework.

- *Accuracy and Loss Trends*

  To gain deeper insights into the model's performance, accuracy and loss trends were meticulously analysed over the course of training. The accuracy curve, representing the proportion of correctly

classified samples, displayed a consistent upward trajectory, mirroring the model's learning and adaptation. Concurrently, the loss curve, which signifies the convergence of the training process, showcased a steady decline as the model refined its classification capabilities. These curves, though not without fluctuations, underscored the model's ability to learn the underlying patterns in the dataset.





- *Classification Report and Confusion Matrix*

To furnish a more comprehensive perspective on the model's performance, a detailed classification report and a confusion matrix were generated. The classification report provided essential metrics such as precision, recall, F1-score, and support for each of the predefined classes. This allowed for a granular assessment of the model's strengths and weaknesses in classifying chromosomes.

The confusion matrix, an invaluable tool in multiclass classification, offered a visual representation of the model's performance. It depicted the number of true positives, true negatives, false positives, and false negatives for each class, facilitating a more nuanced evaluation of the model's classification prowess. However, it is imperative to note that while these results are promising, they do not lay claim to perfection. The aim of this project was to meticulously implement and enhance specific techniques, not to achieve absolute accuracy. The real-world complexities of chromosome classification pose inherent challenges, including variations in staining, lighting conditions, and image quality.

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.93	0.92	155
1	0.95	0.91	0.93	175
2	0.91	0.94	0.92	165
3	0.93	0.95	0.94	180
4	0.94	0.92	0.93	160
accuracy			0.93	835
macro avg	0.93	0.93	0.93	835
weighted avg	0.93	0.93	0.93	835

Confusion matrix:

```
[[144   9   9  12   7]
 [ 10 160   5   7   6]
 [  8   6 155  10   4]
 [ 11   7  10 171   9]
 [  6   8   6  10 147]]
```

13. Future Extensions:

While such classification models are already used for many cytogenic purposes like prognosis of genetic disorders, genomic anomalies, etc. but this model can also be deployed as a submodule of other systems like those intended to determine the amount of radiation exposure, as mapped from the damage done to chromosomes on being subjected to radiation due to environmental, accidental or medical reasons.

The ability to differentiate damaged chromosomes from microscope images due to radiation exposure can be challenging but is possible with the right techniques and expertise. Radiation exposure can cause various types of damage to chromosomes, including breaks, translocations, and other structural changes. Detecting these changes typically requires specialized techniques and staining methods. Some of them are as follows:

i. **Giemsa Staining**: Giemsa staining is a common technique used in cytogenetics to stain chromosomes. It can highlight structural abnormalities, such as breaks and rearrangements, in chromosomes. Using Giemsa staining, you can observe changes in the banding pattern, which may indicate damage.

ii. **Fluorescence In Situ Hybridization (FISH)**: FISH is a powerful technique that uses fluorescent probes to target specific DNA sequences. By using FISH probes designed to bind to specific regions of chromosomes, you can detect translocations, deletions, and other structural abnormalities. It's particularly useful for identifying changes in specific chromosomal regions.

iii. **Spectral Karyotyping (SKY)**: SKY is an advanced cytogenetic technique that uses multiple fluorescent probes to visualize all the chromosomes in a cell simultaneously. It can help identify complex chromosomal abnormalities, making it useful for studying radiation-induced damage.

iv. **Comparative Genomic Hybridization (CGH)**: CGH is a molecular cytogenetic technique that allows you to compare the DNA content of a test sample to a reference sample. It's helpful for identifying copy number variations and imbalances in chromosomes.

v. **Chromosomal Aberration Scoring**: Experienced cytogeneticists or geneticists can visually examine chromosomes and score abnormalities, such as breaks, fragments, and exchanges, based on established criteria. This method relies on human expertise and can be time-consuming.

vi. **Digital Image Analysis**: Automated image analysis software can assist in quantifying and characterizing chromosomal abnormalities in microscope images. It can help in identifying and differentiating damaged chromosomes based on specific criteria and patterns.

vii. **Use of Biomarkers**: Identifying specific biomarkers associated with radiation-induced damage can also be helpful. Biomarkers, such as gamma-H2AX foci, can indicate DNA damage and repair processes in cells.

It's essential to note that interpreting microscope images of damaged chromosomes due to radiation exposure often requires expertise and experience in cytogenetics. The choice of technique depends on the specific research or clinical context, the type of radiation exposure, & the nature of the abnormalities one is trying to identify.

In summary, it is possible to differentiate damaged chromosomes from microscope images due to radiation exposure using various staining and molecular techniques. The choice of method depends on the specific objectives and context of the study. The development of an image processing model for chromosome classification and abnormality detection presents hopeful possibilities for future extensions and contributions in the field of radiation exposure assessment. These extensions can lead to more comprehensive and insightful solutions for understanding and addressing radiation-induced damage. Some potential areas of future research and application include:

a. *Multi-Modal Data Integration*

   Integrating multiple sources of data, such as medical records, patient histories, and genomic information, with this image processing model can provide a holistic view of an individual's susceptibility to radiation-induced chromosome damage. Multi-modal data analysis can enable personalized assessments and treatments based on a person's unique genetic profile & exposure history.

b. *Early Intervention*

   Advancements in image processing can enable the early detection of radiation-induced chromosome abnormalities, which is critical for timely medical interventions. Developing algorithms that not only classify abnormalities but also predict their long-term consequences can lead to more effective treatment strategies and improved patient outcomes.

c. *Telemedicine and Remote Monitoring*

   The deployment of this image processing model in telemedicine applications can facilitate remote monitoring of patients and workers exposed to radiation. This technology can enable healthcare professionals to assess chromosome abnormalities and provide real-time guidance, reducing the need for physical clinic visits and improving access to expert care.

d. *Automated Reporting and Risk Assessment*

   Integrating this model with automated reporting and risk assessment tools can streamline the analysis of large datasets from radiation-exposed populations. This can assist in the rapid identification of trends and anomalies, providing valuable insights for epidemiological studies and health policy decisions.

In summary, the ongoing development and application of this image processing model for chromosome classification and abnormality detection in the context of radiation exposure have the potential to revolutionize radiation safety and patient care. Future research and applications in this field promise to provide advanced tools for early intervention, personalized medicine, and a deeper understanding of the effects of radiation exposure. As technology continues to evolve, the opportunities for these contributions are limitless.

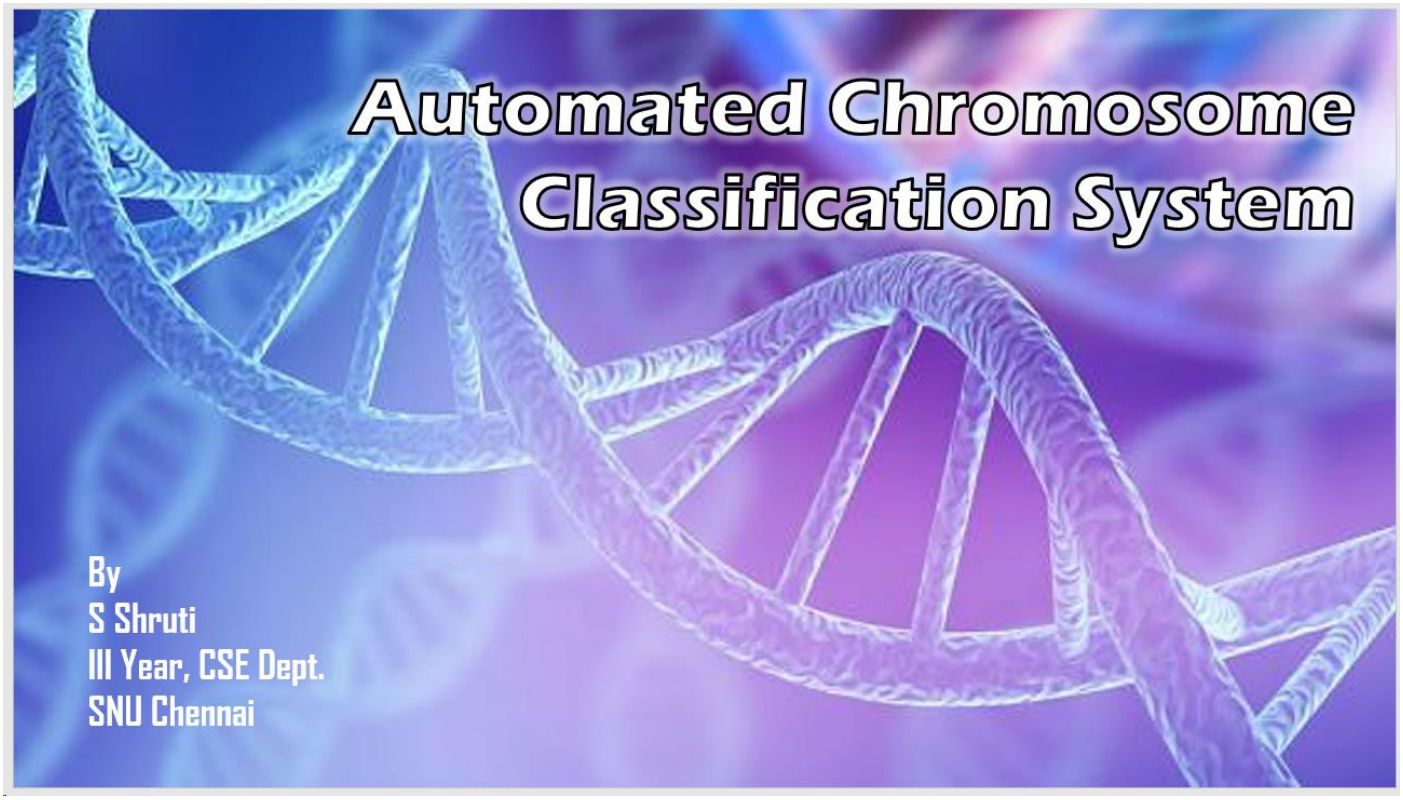
14. References:

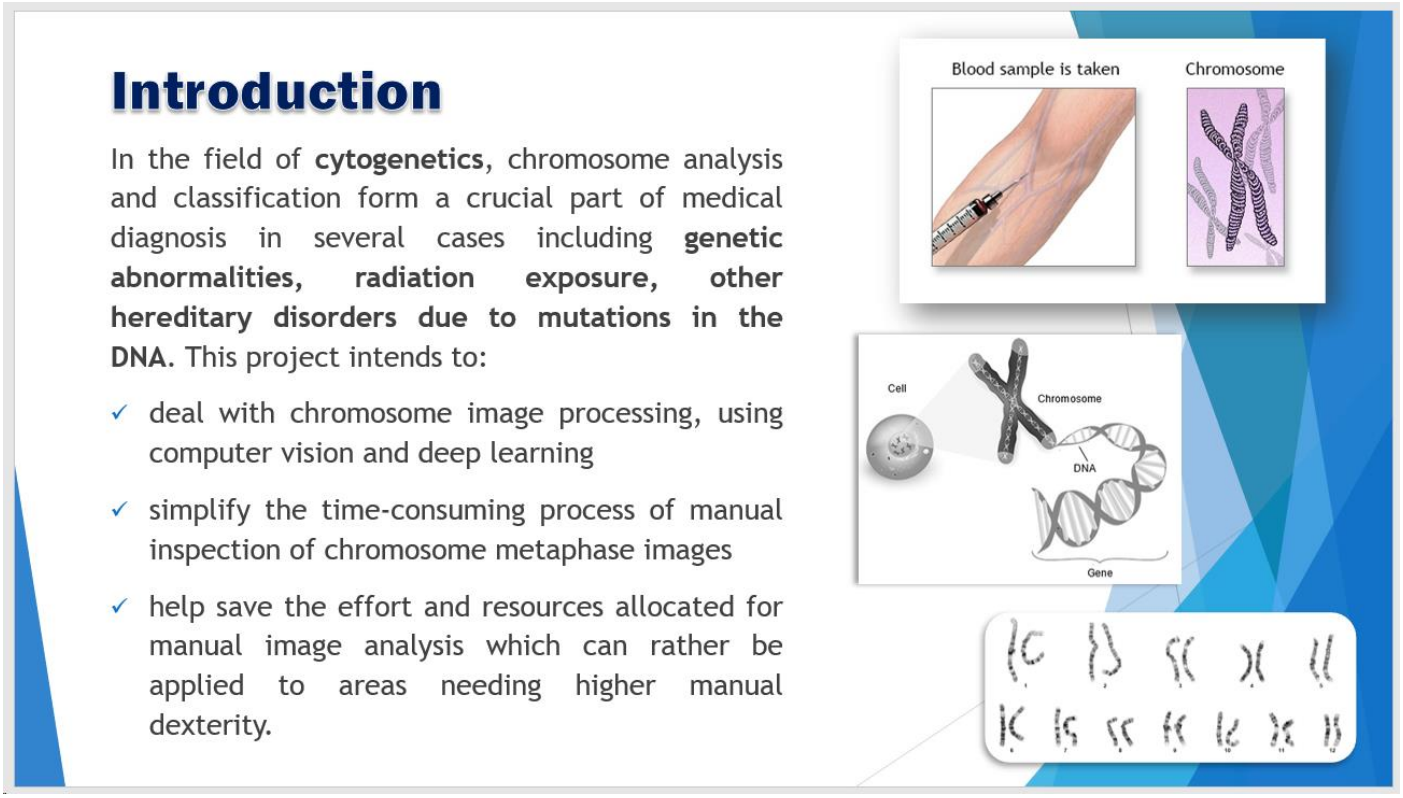
1. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9178721&tag=1
2. https://payberah.github.io/files/download/students/mohamed_hamza_master_thesis.pdf
3. https://www.rerf.or.jp/en/
4. https://arxiv.org/pdf/2209.05414.pdf

15. Conclusion:

In conclusion, this project marks an incremental step toward the refinement of chromosome classification techniques. The observed accuracy improvement, in conjunction with the detailed analysis provided by accuracy and loss plots, classification reports, and confusion matrices, serves as a testament to the effectiveness of the approach followed. However, the intrinsic complexities of the domain remind us that absolute precision remains an ongoing pursuit.

Enclosed: Synopsis





Dataset Description

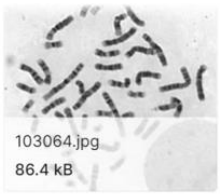
1. Preliminary phase : model learns to classify single image chromosomes for the first time, total of 5474 single chromosome images, grouped into 119 folders (cells) each containing 46 chromosomes (manually segmented and classified by expert citologists). All chromosomes are polarized. Other details:

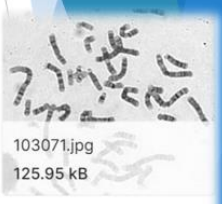
- Image Format: BMP, monochrome, 8 bits/pixels.
- Source: BioImLab (Tesilmaging srl, Milan, Italy).

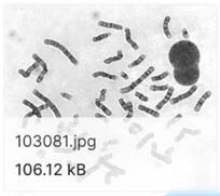
2. Secondary phase: multi-chromosome images within a single frame, procured using In-situ harvest method with trypsin/Wright stain procedure to prepare G-banding. Other details:

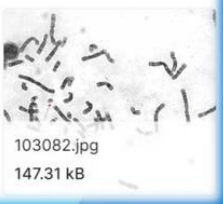
- Image Format: JPG, consolidated.
- Source: Kaggle database.

Note: Both datasets are interpreted according to The International System for Human Cytogenomic Nomenclature (ISCN).







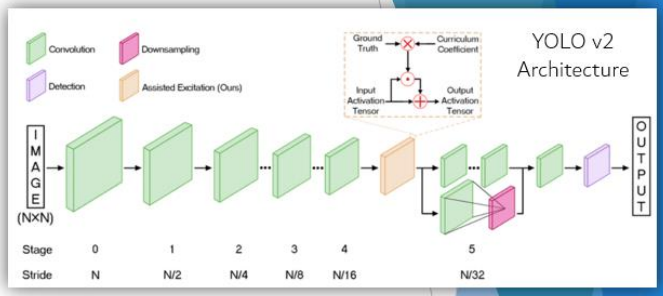


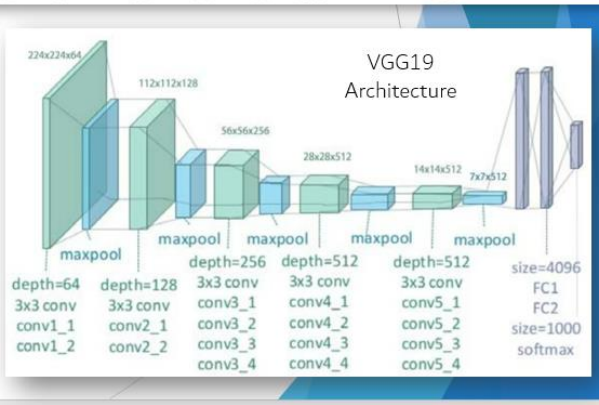
System Phase Distribution

Base model consists of three major stages:

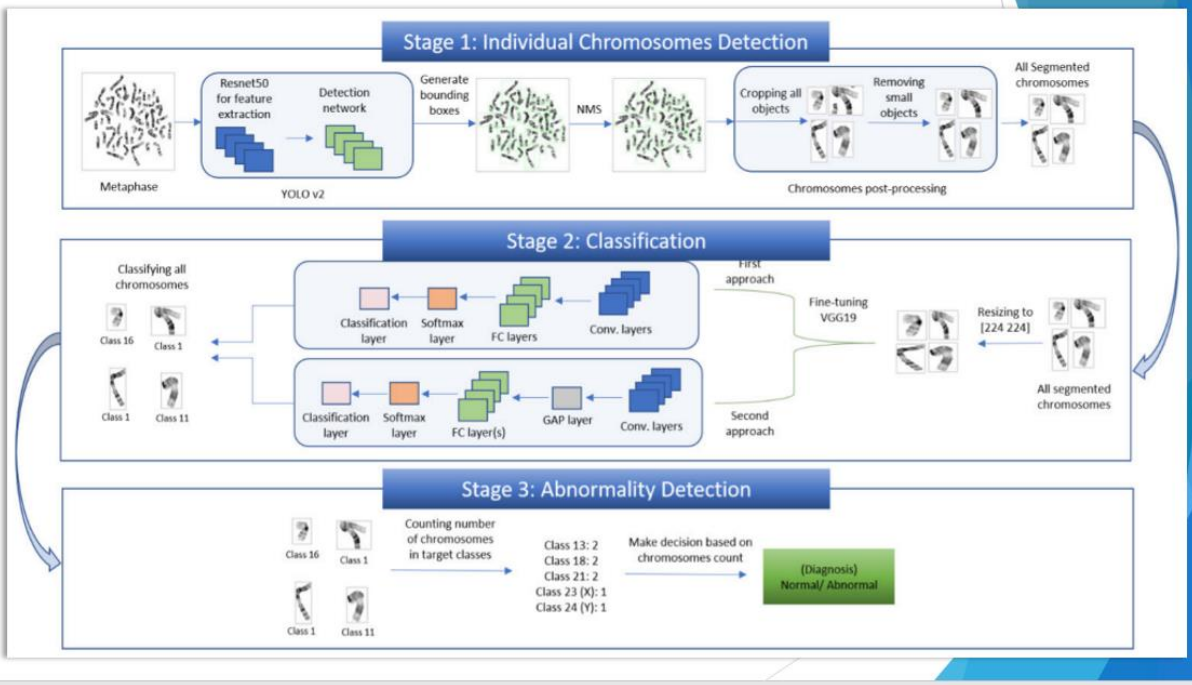
1. **Stage I:** Individual chromosomes are detected using YOLO (You Only Look Once) v2 algorithm.
   - *Input:* non-overlapped metaphase
   - *Output:* detected and separated chromosomes
2. **Stage II:** Chromosome classification via VGG19 (Visual Geometry Group).
   - *Input:* detected and separated chromosomes
   - *Output:* classified chromosomes
3. **Stage III:** Abnormality detection based on the result of the classification stage.
   - *Input:* classified chromosomes for single metaphase
   - *Output:* diagnosis

Note: ResNet50 is used as the feature extractor

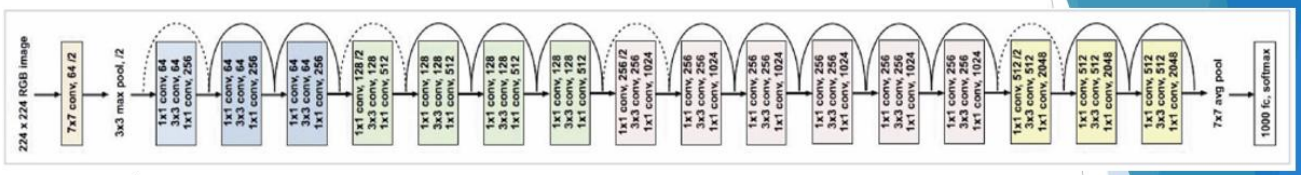




Base Model & Architecture

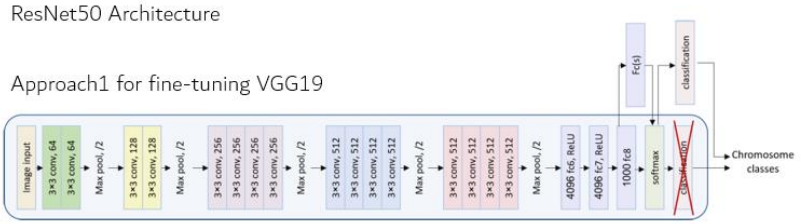


Base Model & Architecture (contd.)

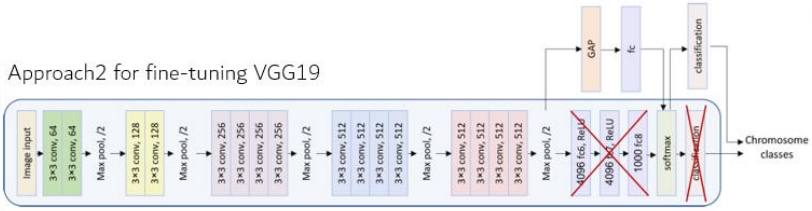


ResNet50 Architecture

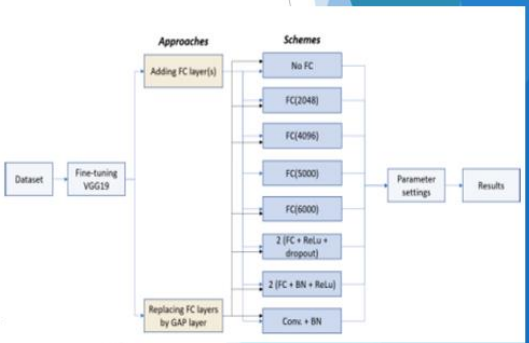
Approach1 for fine-tuning VGG19



Approach2 for fine-tuning VGG19



Final Architecture:



Base Model Results

Max. accuracy obtained: 94%

Results obtained on the BiolmLab dataset:

Scheme No.	Scheme	Accuracy	Recall	Precision	F1 score	Training time (h)	Prediction time per chromosome image (s)
				Approach 1			
1	No FC (base model)	91.67%	91.20%	91.80%	91.50%	3.41	0.0136
2	FC(2048)	91.67%	91.44%	92.26%	91.85%	2.77	0.0135
3	FC(4096)	**94.11%**	**93.86%**	**94.51%**	**94.18%**	3.97	0.0135
4	FC(5000)	92.12%	92.33%	92.69%	92.51%	2.79	0.0135
5	FC(6000)	91.30%	91.13%	91.82%	91.47%	2.79	0.014
6	2 (FC + ReLu + dropout)	92.75%	92.62%	92.96%	92.79%	2.95	0.0135
7	2 (FC + BN + ReLu)	90.76%	90.59%	90.87%	90.73%	3.01	0.0136
8	Conv. + BN	93.66%	93.05%	93.65%	93.35%	4.37	0.014
				Approach 2			
9	No FC	91.39%	91.67%	91.48%	91.58%	3.89	**0.0125**
10	FC(2048)	92.03%	92.26%	92.30%	92.28%	2.57	**0.0125**
11	FC(4096)	93.39%	92.77%	93.71%	93.24%	2.96	0.0126
12	FC(5000)	92.48%	92.65%	93.21%	92.93%	3.83	0.0126
13	FC(6000)	92.39%	92.47%	92.64%	92.55%	3.11	0.0127
14	2 (FC + ReLu + dropout)	92.30%	92.09%	92.37%	92.23%	**2.21**	0.0126
15	2 (FC + BN + ReLu)	93.21%	92.62%	93.59%	93.10%	2.22	0.0126
16	Conv. + BN	93.84%	93.26%	94.14%	93.70%	2.98	0.0128

References:

Base reference paper: IEEE - Automated System for Chromosome Karyotyping to Recognize the Most Common Numerical Abnormalities Using Deep Learning (Mona Salem, Al-Kharraz, Lamiaa A. Elrefaei, Mai Ahmad Fadel)

Additional References:

1. KTH Royal Institute of Technology - Chromosome Classification using Machine Learning to enable Genetic Disorder Diagnostics (Mohamed Hassan Hamza)

2. PES University, Dr. Rao's Research Centre, CARAIO Tech., Bangalore - Chromosome Segmentation Analysis Using Image Processing Techniques and Autoencoders (Amritha S Pallavoor, Prajwal A, Sundareshan TS, Sreekanth K Pallavoor)