Chapter V

APPLICATION IN AUTOMATED KARYOTYPING

Chapter 5 – APPLICATION IN AUTOMATED KARYOTYPING

Take risks: if you win you will be happy, if you lose, you will be wise.

-Peter Kreeft

5.1 Introduction

This chapter discusses the application perspective of the proposed methodology for incremental learning. We divide the chapter into three sections. The first section introduces human chromosomes and karyotyping. In the next section we apply a simple pairing mechanism for automated karyotyping without incremental learning and extend it to incremental update with a windowing technique. The last section discusses identification of new class.

SECTION I

Introduction to Chromosomes and karyotyping

5.2 Human cell

A human cell is the simplest living structure which is capable to live independently. This was first identified in 1663 by R. Hooke and later on in 1838 Schleiden and Schwann identified that this unit functioned according to some definite laws. The cell comprises of five parts: (i) cytoplasm (ii) mitochondriya (iii) ribosomes (iv) nucleus (v) chromosomes. Figure 5.1 shows the organization of human cell.

The membrane of a cell defines the boundary for it. The cytoplasm is a combination of water along with organic and inorganic material. The mitocondria are called as the energy generators where as the ribosomes are protein boosters whereas the [45, R17]. The nucleus of the cell belongs to the category of Eukaryotic cell owing to the presence of membrane compartment structure. The nucleus consists of DNA- Deoxyribonucleic Acid. The DNA composition for every cell is similar but may vary in the form of genomes.

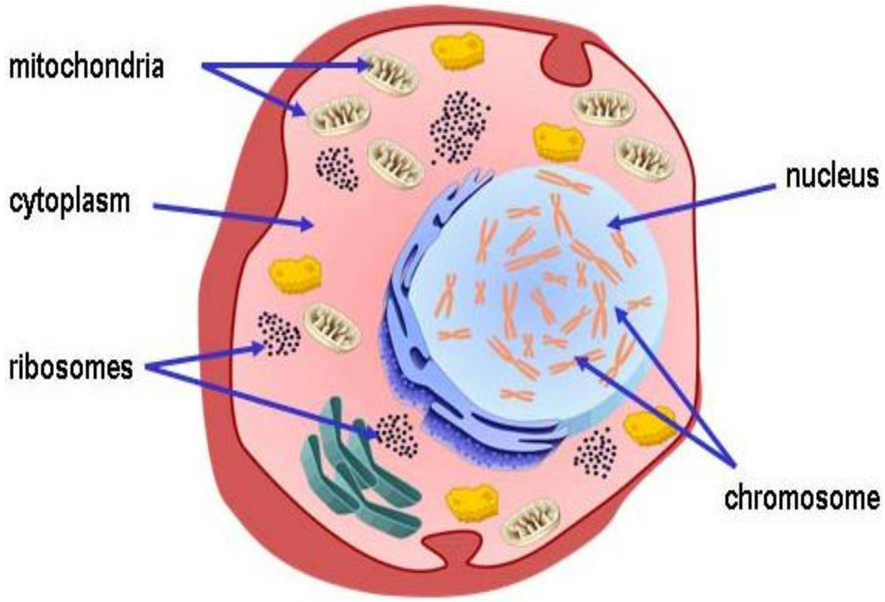


Figure 5.1 Human cell
[Source: Genetics Education, Murdoch Children's Medical Research Institute]

For growth of a human, the cells have to multiply. This process of growth results in the increase in the cell mass, duplications of the cell material and also cell division takes place. When the cell division takes place, the DNA and the protein condense to form chromosome. In cell division takes place with two processes in action. The first is 'Mitosis' and the second is 'Cytokinesis'. In Mitosis, the nucleus is divided and in the Cytokinesis, changes occur in the cytoplasm and the cell division takes place. There are four phases of Mitosis – (i) prophase, (ii) metaphase, (iii) anaphase and (iv) telophase. Amongst the phases, the Metaphase is the phase where the chromosomes of the cell can be aligned with mitotic apparatus and best studied and counted [R17].

5.3 Human chromosome

Chromosomes are located in the nucleus of the cell. They are formed of 'genes' which make up the human genome. The physical characteristics of a human being are controlled by the genomes. The living organisms belonging to same species have same number of chromosomes [R18]. A cell- generally called as somatic cell has 46 chromosomes. The chromosomes are divided into Autosomes and Sex chromosomes. In all 22 pairs of autosomes and one pair is of sex chromosome – X or Y exist. Females have two X chromosomes where as males have X and Y.

Under normal circumstances, the chromosomes are long and thin and invisible. But during the ‘metaphase’ stage they contract and become short. They generally take size of 2-10 µm with diameter of 1-2 µm. It is at this stage that they are stained to make them visible under the microscope. The cells to get the images of these phases are obtained through blood samples or bone marrow [R19]. The features of chromosome are explained below:

1.Size: The chromosomes differ in their sizes, where the 1st chromosome being the longest and the 21st one been the shortest.

2.Centromere: A centromere gets created in the cell division process when the spindle of fibres gets attached to each other. Thus it forms a neck of the chromosome that joins two arms. Figure 5.2 shows the centromere position and the chromosome types. They are –

(a) Metacentric - The arms of the chromosome have same length.

(b) Sub-metacentric - The arms nearly have same length.

(c) Acrocentric – One arm is shorter, the centromere is thus near to one end of the chromosome.

(d) Telocentric – One arm, where the centromere is at the end of the chromosome.

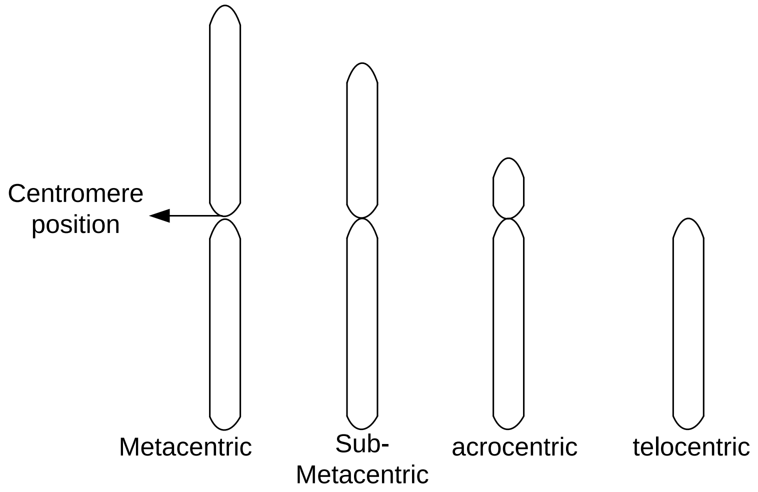


Figure 5.2 Chromosome types

3.Banding pattern: When the chromosomes are stained, they reveal a banding pattern on them. The positions and the band size provide useful information that can be used in learning. There are various types of bands that are obtained – Q-bands, G-bands and R-bands and C-bands. They are obtained by different staining mechanisms. The bands are light and dark strips that appear which help in the identification of the characteristics of the chromosome. Figure 5.3 shows different banding techniques.

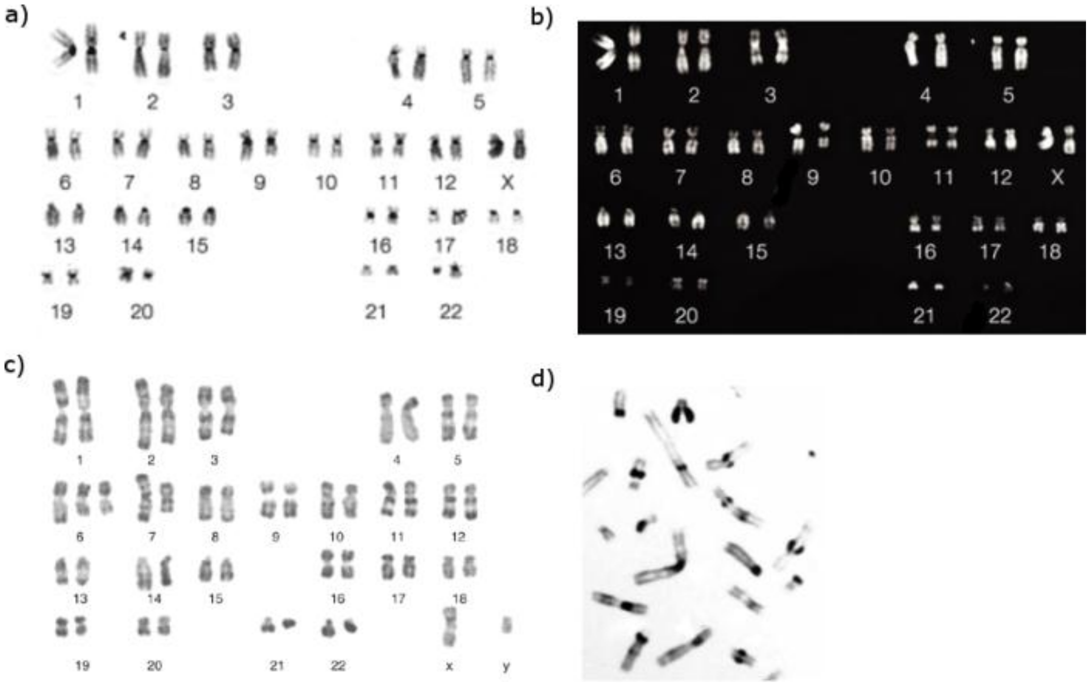


Figure 5.3 Banding techniques[1]

(a) G-bands, (b) Q-bands, (c) R-bands, (d) C-bands.

5.4 Ideogram

An ideogram is a standard for representation of the chromosomes showing the size and band pattern. A nomenclature has been developed for the chromosome identification and at present the system in use is the International System for Human Cytogenetic Nomenclature- called as ISCN 1995 [R20]. Figure 5.4 shows an ideogram [86]. The system rules are:

- The chromosomes numbering starts with the centromere.
- They have long and short arms namely 'p' and 'q' where the top arm is always referred to as p.
- Each arm is assigned to a region, where the distance of the arm increases with respect to the centromere position.
- Within each region, there are bands that are identified based on the staining used.

[1] *(a and b) Copyright 2001 Nature Publishing Group. From: Chromosome translocations: dangerous liaisons revisited. Rowley, J. D. Nature Reviews Cancer 1, 245-250. c) Copyright 2007 Nature Publishing Group. From: Stamatoullas, A., et al. Conventional cytogenetics of nodular lymphocyte-predominant Hodgkin's lymphoma. Leukemia 21, 2064–2067. d) Copyright 2005 Nature Publishing Group. From: Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. Vega, H.et al., Nature Genetics 35, 468-470.)*

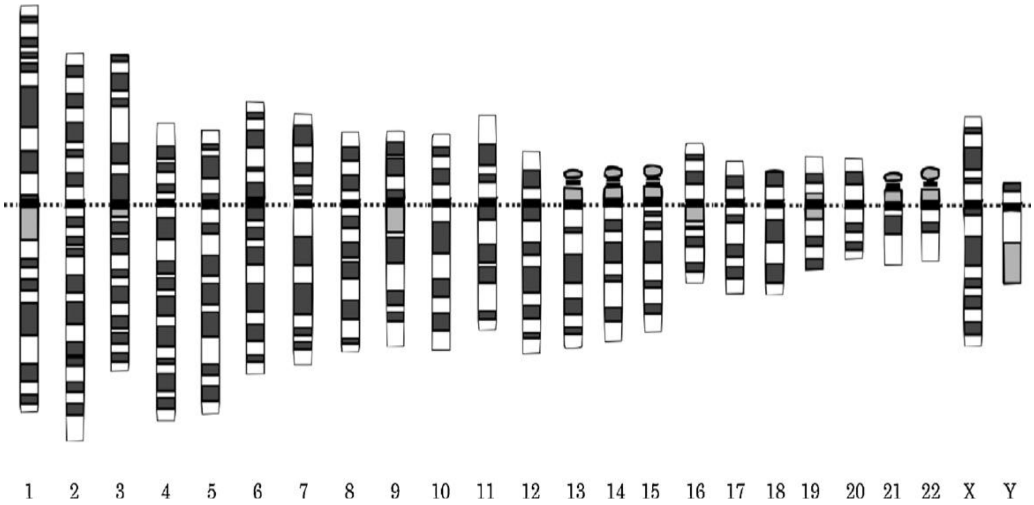


Figure 5.4 Ideogram for chromosomes

5.5 Features of chromosomes

The features that are used in the classification of the chromosomes are categorized as (i) geometrical and (ii) morphological. The features are depicted in figure below.

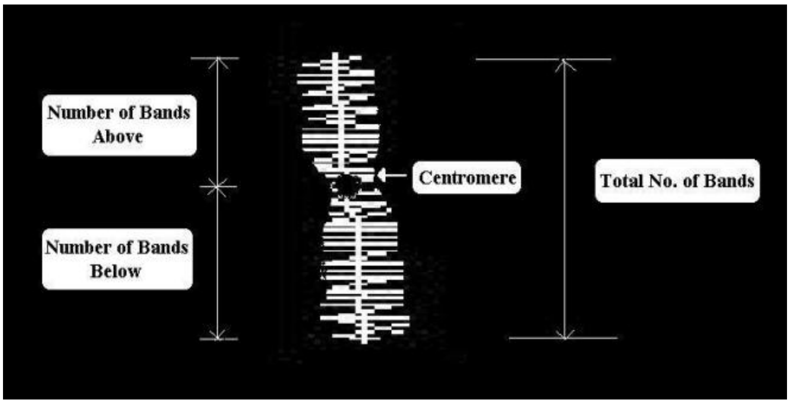


Figure 5.5 Features of the chromosomes

The geometric features involve –

(i) Chromosome length – the length of the medial axis of the chromosome forms the length of the chromosomes. It is in number of pixels.

(ii) Area of the chromosome – pixels enclosed in the closed boundary.

(iii) Perimeter of the chromosome – pixels lying on the boundary.

(iv) Number of bands that exist above the centromere and below the centromere.

The chromosomes have intensity bands as said earlier along their major axis. These bands – the number is an identified feature for karyotyping (explained in the following section)

Morphological features include-

(i) Centromeric-index – ratio of the arms of the chromosome; long: short

(ii) Ratio of length of the arms of the chromosome

(iii) Band profile density – Band of a chromosome is defined to be a visibly separable part among the varied intensity ranges on the chromosomes. Thus the chromosome can be viewed to be a series of dark and light intensities. Density profile is a single dimensional vector obtained by intensity sampling.

(iv) Width of the bands

5.6 Karyotyping

A karyogram represents a stained image of the chromosomes. This is obtained with the G-stain and the chromosomes are arranged in the decreasing size. Karyotype is extracting set of characteristics that are useful to detect the abnormalities in the karyogram. 'Karyotyping' is classification of the chromosome. This classification is extremely essential for the analysis of the chromosomes. Based on Denver classification standard, the chromosomes are classified as shown in table 5.1.

Table 5.1 Denver group classification

Group	Sub-Class	Size	Centromere position
A	1-3	Large	Metacentromeric
B	4-5	Large	Submetacentric
C	6-12, X	Medium	Submetacentric
D	13-15	Medium	Acrocentric
E	16-18	Relatively short	Meta/ Submetacentric
F	19-20	Short	Submetacentric
G	21-22,Y	Short	Acrocentric

According to the Denver standard, they are classified into Groups A to G. The standard specifies the subclasses, the length and the centromere position. A karyotype and metaphase image is shown in the figure 5.5 form publicly available database [R22].

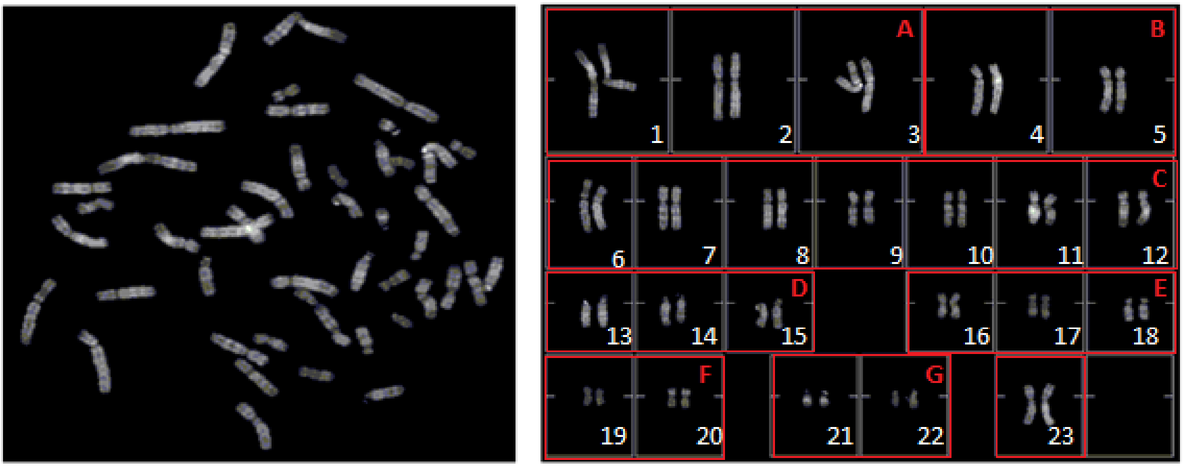


Figure 5.6 (a) A sample metaphase image and (b) Karyotype

5.6.1 Manual karyotyping

In manual karyotyping, the entire process of classification is done by hand. A cytogeneticist is the one who has to carry out the entire analysis by hand which is quite time consuming. It takes hours to have the karyotype identified. The expert has to identify the picture of each and every chromosome and using his knowledge performs the classification. Thus manual karyotyping methods have issues –

- The process is very tedious
- It requires human expert knowledge
- It is expensive and time consuming
- It is laborious and error prone

5.7 Need for incremental learning in Karyotyping

Automated methods that are efficient are now looked upon [R23]. Many Automated Karyotyping Systems (AKS) have been proposed till today performing Visual imaging, Cytovision, Integrated Computer aided detection and so on. The methods have offered advantages over the traditional karyotyping in terms of fast acquisition of the sample, been inexpensive, reduced labour costs and better interpretation of the image data. These methods though claiming to have an improved accuracy, the systems are not routinely accepted in the clinical laboratories.

With the increase in the unlabeled images for the karyotyping, classifiers that can have the learning with this new data is the need of the time. There has to be ‘expert’ knowledge built in the software for the automation, thus initiating the learning to be incremental.

Most of the recent history of automated karyotyping research has focused on designing an intelligent and a fully AKS [5, 39], that must ideally have no human interference but must still deliver a human-like performance. An AKS must therefore necessarily be consistent with human learning. Like human beings, intelligent systems must also make decisions and form inferences [121] using the knowledge acquired from the learned patterns. Knowledge acquisition is fundamental to both human and machine learning, and since the brain is often confronted with new environments containing information that may conflict with its prior knowledge or experience, connections between machine and human learning needs to be considered for development of human centric systems [103]. Another interesting point to be considered about human learning is that humans learn new information without forgetting previously acquired knowledge. This however, raises the so-called stability-plasticity dilemma [109, 143] one of the key problems in knowledge management when designing the classifiers for the automated systems. In an artificially intelligent system, some information may sometime be lost to learn new information, as learning new patterns will tend to overwrite formerly acquired knowledge [56, 90]. The dilemma points out the fact that a completely stable classifier will preserve existing knowledge, but will not accommodate any new information, whereas a completely plastic classifier will learn new information but will not conserve prior knowledge [90, 143].

Most of the typical neural network based classifiers such as multilayer perceptron, radial basis function, wavelet networks, Kohonen networks reported in the literature for the application of AKS [9, 139] involve discarding the existing classifier and retraining the classifier using all of the data that have been accumulated thus far. This approach, lying on the "stability" end of the spectrum, however, suffers from "Catastrophic Forgetting", which is the loss of all previously acquired information [109]. Moreover, in practical application like AKS, the collection and analysis of training data is expensive and time consuming [146]. It is common to acquire additional training data from the environment at some point in time after an AKS has originally been trained and deployed for classification. Since limited training data is typically employed in practice [48, 58] and underlying testing data may also have reasonable tangible knowledge worth affecting further decision making, the classifiers designed for AKS must allow for adaptation in response to newly acquired training data from the operational environment or other sources. The classifiers with such an ability of effective adaptation would certainly be an undisputed asset for sustaining a high level of performance of AKS.

Thus, despite the large variety of the classifiers that have been experimented and implemented for the task of chromosome classification there still remains a wide opportunity to explore the usefulness of other approaches. It is thus necessary to update the existing classifiers in AKS to accommodate new data without compromising classification performance on old data and without the need of retraining it from scratch. This research experiments new class evolution along with classification for chromosome classification based on incremental learning to maintain stability–plasticity balance meeting the following criteria that has stated earlier as well:

(i) It should be able to learn additional information from new data.
(ii) It should not require access to the original data, used to train the existing classifier.
(iii) It should preserve previously acquired knowledge (that is, it should not suffer from catastrophic forgetting).
(iv) It should be able to accommodate new classes that may be introduced with new data.

An algorithm that possesses these properties would be an indispensable tool for pattern recognition and machine learning researchers, since virtually unlimited number of applications can benefit from such a versatile incremental learning algorithm [143].

The next section discusses experimentation for the same with proposed incremental learning methodology.

SECTION II

Incremental learning with windowing technique

5.8 Automated karyotyping

Automated karyotyping of the homologue chromosomes is an active area of research and a number of similarity measures have been reported in the literature for the same. Similarity measures, based on cross-correlation, mutual information and Fourier techniques are computationally efficient and robust against outliers but it is sensitive to interpolation artifacts. Other similarity measures, using a covariance matrix instead of a high-dimensional histogram lead to computational complexity impact with the increase in the neighborhood. Automatic karyotyping of the chromosomes is a difficult task because the chromosomes appear blurred with undefined edges and low level of band pattern information. Further the availability of new unlabeled data set increases the complexity.

The experimentation shows that the proposed approach of i-learn with efficient feature extraction techniques can perform accurate classification. The learning takes into account the incremental update mechanism with windowing technique.

The initial part of the proposed algorithm deals with the extraction of the band patterns, i.e. intensity values along the axis of the chromosome. The generated sequence of the intensity values is used to have the karyotyping for the chromosomes. This karyotyping is for the sub-classes within the Denver Group.

5.8.1 Feature extraction

The Denver classification is based in length and centromeric index and the classification within the Denver groups is based on the band pattern of the chromosomes. Every Denver class has a pair of chromosomes with identical band patterns. Chromosomes are non rigid in nature and therefore exhibit high degree of shape variability [2, 9, 54]. They may be highly curved or bent. But irrespective of these unpredictable shapes of the chromosomes, the sequence of the band patterns must necessarily be a correctly identified.

The procedure used to extract the sequence of band patterns i.e. intensity values on the chromosomes is:

- Conversion to binary form
- Dilation of the input image
- Boundary detection of the chromosome
- Thinning of the chromosome to get ‘Axis of the chromosome’
- For band plotting, identification of three consecutive pixels on the axis
- Plotting of perpendicular line at the line joining 1st and 3rd pixel at the 2nd pixel.
- Extending the perpendicular line till the boundary of the chromosome
- Calculation average of each band on the chromosomes
- Plotting of graph of average intensities against band number.

Figure 5.7 shows the band plotting and the average intensities vs. the band number.

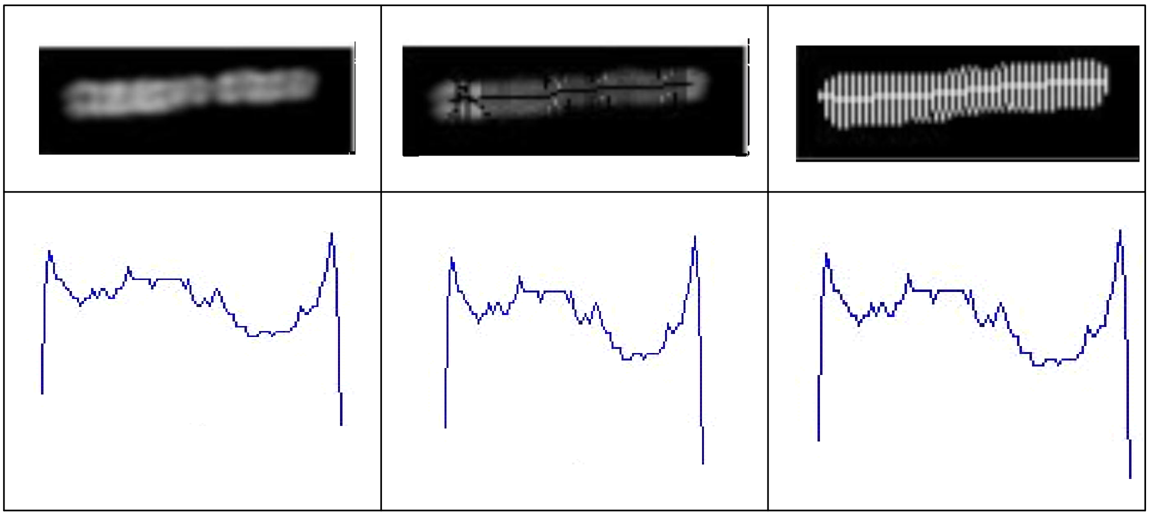


Figure 5.7 Band plotting (top row) and average intensity vs. band number (bottom)

Figure 5.8(a) shows a chromosome belonging to Group A. The chromosome is thinned to mark the axis and perpendicular segments to the axis are plotted at an interval of every three pixels as seen in figure 5.8 (b). Figure 5.8 (c) shows the axis and the perpendicular segments superimposed on the original chromosome image. The perpendicular segment with the least width is the centromere of the chromosome. (Centromere is the narrowest part of the chromosome which connects its two arms). The sequence of the band patterns on the chromosome is derived with respect to the centromere. Figure 5.8 (d) shows the plot of the intensity value versus the perpendicular row number. The 8th perpendicular row indicates the position of the centromere of the chromosome. The average intensity value at this position is 63. Every intensity value, plotted for each perpendicular row is the average of the intensity values of all the pixels lying on the corresponding perpendicular segment. The band pattern sequences for each chromosome of every group are derived using a similar procedure. This sequence of band pattern was successfully derived even for curved and bend chromosomes.

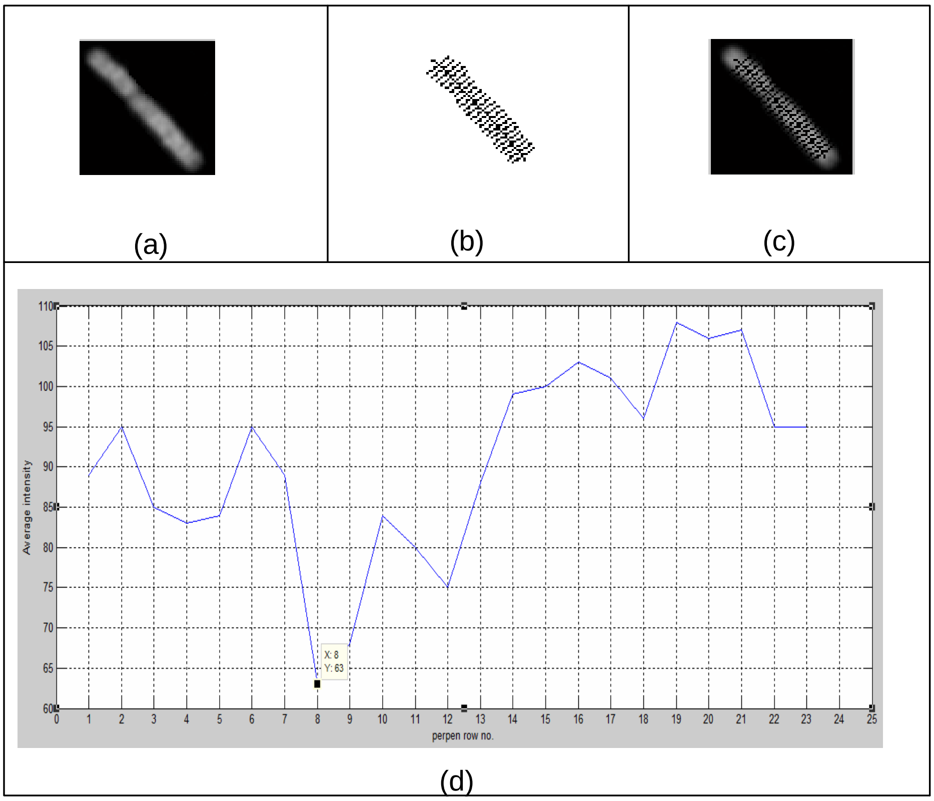


Figure 5.8 (a) Chromosome of Group A (b) Axis with perpendicular segments (c) Superimposed on the original image (d) Centromere identification

For automated karyotyping, we performed the experimentation with two different variants for comparative analysis: (i) simple pairing and (ii) Incremental update. It is assumed that the chromosomes are segmented and classified as per the Denver rule in groups A to G. The proposed approach describes the further classification i.e. pairing the chromosomes in each Denver group into their classes (1-23). The band pattern of the chromosomes (intensity values along the axis of the chromosome) is the most dominant feature in subgroup classification and the generated sequence of the intensity values is therefore selected as the feature used to pair the chromosomes.

5.8.2 Related work in AKS

Khemlinski et al. [5] used supervised linear classifier together with combinatorial optimization algorithm to compute the pairing of chromosomes. Mutual information is also one of the metric proposed by him for the automated pairing [3, 112]. Feng et al., proposed a similarity matching algorithm applying gradient profile for the automated pairing achieving 91.3 % and 100% of classification accuracies in male and female cells respectively [140]. Reel et al. proposed a new mutual information based similarity measure for the registration of medical images [96]. Similarity measures, based on cross-correlation, mutual information and Fourier techniques are computationally efficient and robust against outliers, yet they are sensitive to interpolation artifacts. Other similarity measures, using a covariance matrix instead of a high-dimensional histogram lead to computational complexity impact with the increase in the neighborhood. Techniques generating pairs or clusters of identical or similar objects are required to be robust in terms of outliers, initialization [98, 23]. Such new techniques with methods of matrix factorization [82, 19] have engaged attention and are experimented on a variety of data. Despite the efforts by researches in improving the automated systems, robust and efficient incremental methods are need of the time.

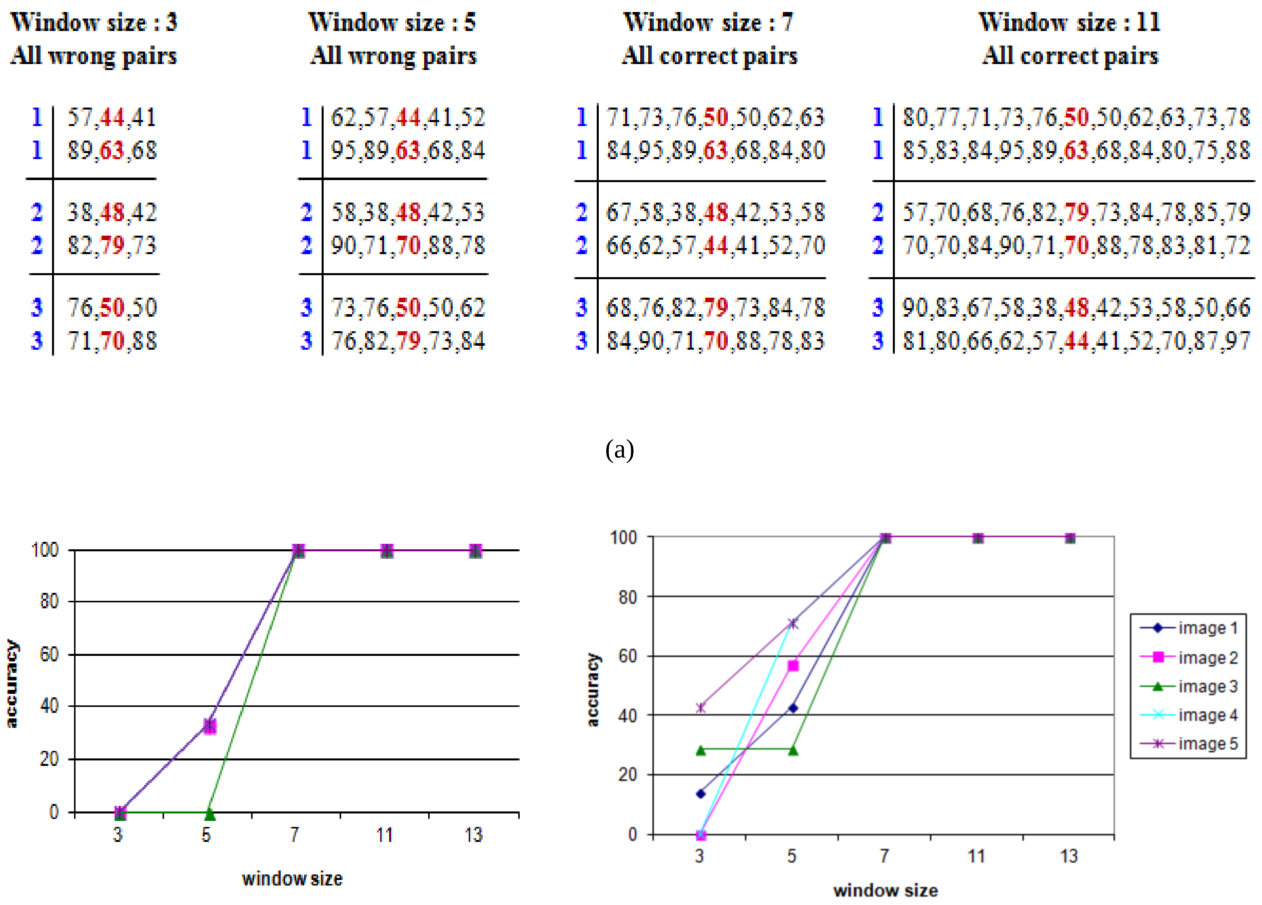
5.8.3 Experimentation and results on pairing

The experimentation for simple pairing is carried out here is for the automatic karyotyping for chromosomes in the metaphase images. Consider Ch_1 and Ch_2 to be the two chromosomes on which decision is to be taken regarding the formation of pair. Accordingly, any chromosome 'i' corresponds to a sequence -

$$Ch_i = \{c_{i1}, c_{i2}, ...c_{im}\}$$

Where c_{i1} is the intensity value at first perpendicular row. The value of m is set according to the window size selected, thus m would be varying as per i. Selection of the window size is the based on the intensity values that appear above and below the centromere. The figure 5.9 shows the results of simple technique for Groups A and C using N-factor.

The performance of the pairing was tested and analyzed using 50 images of group 'A' and group 'C' from Bio Image Lab Database. The results demonstrate an overall accuracy of 100 % for group 'A' and 97 % for group 'C'. The experimentation demonstrates that the proposed approach achieves higher performance for classification of chromosomes. Group A and Group C, groups with different chromosome characteristics were chosen for initial experimentation. Group A has large, metacentromeric chromosomes and Group C has maximum number of classes with medium, submetacentric chromosomes. Initial experimentation was restricted to only these two groups.



Window size : 3 All wrong pairs	
1	57,44,41
1	89,63,68
2	38,48,42
2	82,79,73
3	76,50,50
3	71,70,88

Window size : 5 All wrong pairs	
1	62,57,44,41,52
1	95,89,63,68,84
2	58,38,48,42,53
2	90,71,70,88,78
3	73,76,50,50,62
3	76,82,79,73,84

Window size : 7 All correct pairs	
1	71,73,76,50,50,62,63
1	84,95,89,63,68,84,80
2	67,58,38,48,42,53,58
2	66,62,57,44,41,52,70
3	68,76,82,79,73,84,78
3	84,90,71,70,88,78,83

Window size : 11 All correct pairs	
1	80,77,71,73,76,50,50,62,63,73,78
1	85,83,84,95,89,63,68,84,80,75,88
2	57,70,68,76,82,79,73,84,78,85,79
2	70,70,84,90,71,70,88,78,83,81,72
3	90,83,67,58,38,48,42,53,58,50,66
3	81,80,66,62,57,44,41,52,70,87,97

(a)

(b)

(c)

Figure 5.9 (a) Results with varying window sizes demonstrated for Group A (3 classes , 6 pairs) (b) The results for 5 images of Group A indicating the correct pair formation when the window size is 7 and above. (c) The results for 5 images of Group C indicating the correct pair formation when the window size is 7 and above

Table 5.2 Comparison of the proposed approach with other methods

Image Number	Results for (Group A, class 1,2,3)			
	Dot Product achieved / ideal	Euclidean Distance achieved / ideal	Mutual Information achieved / ideal	Proposed Approach achieved / ideal
8	1/3 [1A , 1B]	0/3	1/3 [3A , 3B]	3/3
14	1/3 [1A , 1B]	1/3 [1A , 1B]	3/3	3/3
19	1/3 [3A , 3B]	1/3 [2A , 2B]	1/3 [2A, 2B]	3/3
49	1/3 [2A, 2B]	1/3 [1A , 1B]	3/3	3/3
79	3/3	1/3 [2A , 2B]	3/3	3/3

The selection of the window size i.e. the number of the intensity levels of the chromosomes above and below the centromere was initially experimented. Figure 5.9 demonstrates the pairing results for 5 images of Group A and Group C from the dataset used for testing the developed algorithm. Correct pairing of the chromosomes in both the groups was achieved with the window size of 7 which indicates that the intensity value of the chromosome at the centromere and the intensity values of three bands above and below the centromere are enough for obtaining the correct pairing within Denver groups.

The experiments also comprised a comparative study of the other methods based on various distance measures reported in literature. It included the dot product and Euclidean distances between the chromosomes of each class. Additionally, the most commonly used measure, Mutual Information, which signifies the amount of information that one variable contains about another random variable (in this case chromosome) is also used to compare the results of the proposed algorithm. The mutual information *I(Chp: Chq)* , between two candidate chromosomes *Chp* and *Chq* is calculated as the relative entropy between the joint distribution and the product distribution as detailed in [96].

Table 5.2 summarizes the results obtained using various distance measures reported in the literature and the proposed approach as experimented for Group A. As group A has only 3 classes, impairing or wrong pairing of any one pair directly reduced the pairing accuracy to 33%. Thus the pairing results are either 1/3 or 3/3 i.e. either only one pair is correctly formed out of three or all six chromosomes form three correct pairs. Similar results were also obtained for Group C. But Group C has 7 classes with 14 chromosomes and so gave varying accuracies in few images. The overall accuracy pairing accuracy of 100 % is achieved for group 'A' and 97 % for group 'C'. The proposed approach has given encouraging results when the chromosomes were bent and curved and also it demonstrates that the proposed measure achieves higher performance in pairing chromosomes.

5.9 Incremental approach

In an artificially intelligent system, some information may sometime be lost to learn new information, as learning new patterns will tend to overwrite formerly acquired knowledge [114, 111]. To the best of our knowledge, most of the reported pairing and classification methods do not use the previously learnt knowledge for further classification. These methods tend to depend on the training data and suffer from catastrophic forgetting [109]. Moreover, the earlier reported approaches of chromosome classification including simple pairing [82, 98] require the entire set of chromosomes to be available at once in the memory. The issue that they can be evolved at later stages remains untouched.

This issue is addressed using the learning methods that use the learnt knowledge for decision making. We hence experimented to retain the previously acquired knowledge with incremental learning along with classification using N-factor. The non-incremental method discussed in the previous section though showed good accuracies for Group A, it lowered the accuracies for Group C. To improve on the same and to accommodate new classes with the new chromosomes, we applied the incremental approach for update of the knowledge with initial knowledge base (KB) availability. The approach works in semi-supervised way. The approach is an extension to the windowing technique for incremental learning.

5.9.1 Algorithm and Experimentation for incremental update

The algorithm is detailed below:

Stage 1: Building the KB:

Input: Labeled data

Output : Initial KB

- For every labeled data the representative series is:

  { $(Ch_1 , SC_1), (Ch_2 , SC_1), \ldots\ldots\ldots\ldots\ldots\ldots\ldots (Ch_n , SC_m)$ }

  where, chromosomes $Ch1,Ch2,\ldots..Ch_n \in A$ *or* B *or* $C \mid n \leq 24$

  and $SC_1\ldots SC_m$ are their respective classes; $m \leq 12$

Here n is the maximum number of chromosomes in the labeled data and m is the respective sub group or class. Three groups A, B and C would all together have 12 classes and 24 chromosomes.

- Set $Ch\_ctr =1$, $KB\_init =\{ \}$
- Do until Ch_ctr $\leq$ n

For $\{ch_p , ch_q\} \in SC_x$ where $x \in 1\ldots m$ for Denver Group and $p,q \leq n$

- Compute N-factor between ch_p , ch_q
- $KB\_init_{lab} = Rep(avg\{ ch_p , ch_q \}, N\text{-}factor)$

Stage 2: Incremental Learning

The next part of the algorithm uses the KB for classification

Input: Unlabeled data

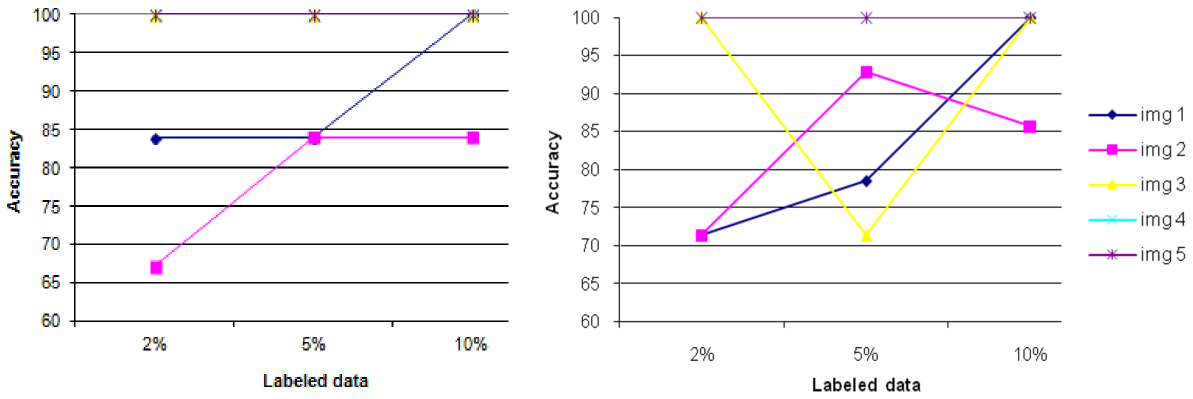
Output : Labeled data and selective Updated KB

Ch_{unlab} is the unlabeled chromosome that needs to be classified and $KB\_init_{class\_x\_i}$ represents the chromosome representative for Group x with sub-class i.

- For $Ch_{unlab} \in Class_x$ where $x \in A$ or B or C
- Do until $Ch\_ctr \leq m$ where m is the sub-classes of Denver

  $$\text{N-val} = \text{N-factor}(KB\_init_{class\_x\_i}, Ch_{unlab})$$
- Select min N-val
- Classify into group with min N-val
- Update KB selectively

  $$KB\_init_{class\_x\_i} = avg[(prev(KB\_init_{class\_x\_i}), Ch_{unlab},); prev(N\text{-}val, new\_N\text{-}val))]$$



(a) (b)

Figure 5.10 Classification Accuracies (a) Group A (b) Group C

The performance of the proposed algorithm was tested and analyzed using 75 images of group A, B and C (1800 chromosomes) from publically available data set at *http://biolab.dei.unipd.it* developed by Enea et al. [86]. 'X' chromosome of class C has not been included in the data set.

Figure 5.10 depicts the effect of available labeled data on the classification accuracies of group A and C. Initially only 2% of the entire dataset was used for building the KB. Rest of the data was used as unlabeled for testing the classification accuracies. Gradually the labeled data for initial learning was increased from 2% to 10% of the entire data set of 1800 chromosomes. Figure 5.10 (a) shows the classification accuracy as obtained for classification of chromosomes of group A into their respective 3 subclasses. Figure 5.10 (b) illustrates the accuracies as obtained for classification of the chromosomes into six sub groups (class 6-12) of group C. For the same number of unlabeled data, the indecisiveness in group C is obvious because group C is the largest group with maximum number of sub groups. The probability of misclassifications is bound to increase with number of classes. Group B has large chromosomes and has only two subgroups. 100 % accuracy was obtained for group B even with just 2% of initial labeled data. Overall average classification accuracy of 97 % is obtained using the proposed approach.

5.9.2 Discussion on incremental update

The performance of the algorithm was tested and analyzed using 75 images of Group A, Group B and Group C from publicly available data set giving an overall pairing accuracy of 97 %. The experimentation demonstrates that the proposed approach of incremental learning showed acceptable accuracies. The approach was based on the point values of the band pattern. The window size plays a significant role in the classification process. The accuracy is dominated by this selection and a stable positioning of the window has given us better results owing to appropriate feature selection done with the window. The window size selection affects the learning and would not be applicable for other Groups. The reason behind the same is there are varying sizes of the chromosomes. This is the case in chromosomes of group F and G where the position of the centromere lies at the extreme end of the chromosome. So, it will be difficult to extract a sequence of 3 bands above and below the centromere to form a window of 7. This is one of the considerations when extending the proposed approach to all the groups. Selecting the window size with centromere as the centre of the sequence ruled out the possibility of erroneous classification due to broken edges of chromosomes or unseen band patterns at the extreme ends of the chromosomes. The performance of the approach is experimented only within each Denver group. It can however also be directly applied to all the segmented chromosomes in the metaphase image and the problem of forming 22 classes can also be addressed.

The impact of learning with selective incremental methodology for automated karyotyping is proven. Artificial neural network based approach [139] is a reported popular approach for automated karyotyping. These approaches need training of the neural network with large data bases, setting of the parameters of the neural network and choosing the hidden layers of the network. The proposed approach eliminates the necessity of such crucial requirements and simplifies the process of classification substantially still retaining high classification accuracy.

The probability of learning through a misclassification in the proposed approach yet cannot be eliminated. This learning impacts the KB building to some extend and may result in a lower classification rate. However, the same can be overruled with highly effective and well defined labeled data that results in getting better accuracy.

The intention of the work is to put forth the need of an incremental approach in the karyotyping process so as to enhance and make optimal and simple decision making process. The incremental approach shows that with a robust KB and effective labeled data, the required benchmark is reached. The incremental process thus helps in building a computerized approach that will have the karyotyping an automated one. Further research in the method aims to have minimal dependency on the labeled data set availability along with introduction of threshold based techniques and extension to evolve a new class.

SECTION III
Incremental learning with new class evolution

5.10 New class generation in AKS

This section of the chapter shows experimentation of the novel approach of **I**ncremental **L**earning for **C**hromosome **C**lassification (ILC^2) for automated karyotyping of metaphase chromosomes. It addresses the issue of catastrophic forgetting with the generation of new class and performs knowledge amassing to classify the chromosomes in Denver groups (A -G).

This section of the chapter, proposes a semi-supervised incremental learning algorithm satisfying all of the previously mentioned criteria for efficient karyotyping of metaphase chromosomes. The main advantage of incremental learning is the ability to sustain a high level of performance yet bound the memory requirements [78]. Furthermore, since training is only performed on the new training sequences and not on all accumulated data (as the availability of threshold value makes it possible); on-line learning would also lower time complexity needed to learn new data. Finally, incremental learning may provide a powerful tool in a human-centric approach and is therefore an undisputed asset for sustaining a high level of performance [138]. The novelty and the contribution of this work lie in:

(i) Most of the previously reported methods utilize various similarity and pattern matching algorithms to classify and pair the chromosomes [98, 114]. The proposed semi-supervised method, ILC^2 uses N-factor only for initial decision making and further validates the decision based on the criteria of threshold value, which governs the belongingness of a particular class. This leads to appreciable increase in the classification accuracy.

(ii) To the best of our knowledge, the classifiers based on thresholding have been of static approach with heuristic calculations. The present work proposes the tuning of the threshold value, making it dynamic. The threshold value changes with every new learning. The knowledge base (KB) is also, therefore updated with the each new learning.

(iii) To further enhance the classification accuracy, finally a deviation factor delta δ is also introduced to accommodate the variations in the feature vector yielding appropriate classification.

(iv) Most importantly, the incremental learning approach performs knowledge amassing with every dataset and still retains the earlier learned knowledge.

5.10.1 Proposed approach of ILC^2 for chromosome classification

The focus of this study is to propose a reliable and an efficient method for the classification of the chromosomes using the approach of semi supervised incremental learning. The various steps in the proposed approach can be envisioned in two parts:

- Extraction of the chromosome features and building of the feature vectors.
- Generating, updating and evolving the KB to classify the unlabelled data using the proposed approach of ILC^2.

The geometric features used for chromosome classification include the length, arm ratio and the centromeric index of the chromosome. These are reported to be the most dominant features for Denver Group classification. The two latter indices are related algebraically quite simply, but are used with the aim of having a concrete feature vector, which further plays a vital role in the generation of the knowledge base and the decision making process. This process is already explained in the previous section for feature extraction. Table 5.3 shows the extracted features.

The proposed algorithm, ILC^2 is a semi-supervised approach for the efficient karyotyping of chromosomes in their respective Denver Groups. ILC^2 is comprised of two major steps which include initially the building of KB and further its usage for the classification of the unlabelled data followed by updating of the KB with every new learning. The generated feature vector along with few additional parameters, as will be detailed in the algorithm will jointly form the KB.

Table 5.3: The extracted features of the chromosome no. 5-1a from the database.
($C_{Gr\,:}$ Denver group of the chromosome ; C_{cl} : Class of the chromosomes ; C_L: Length of the chromosomes ; C_I : Centromeric index; C_R : Centromeric ratio)

C_{Gr}	A						B				C	
C_{cl}	1A	1B	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B
C_L	67	62	70	75	50	51	56	50	51	52	50	46
C_I	0.48	0.46	0.36	0.48	0.40	0.47	0.21	0.25	0.23	0.23	0.35	0.33
C_R	1.07	1.16	1.72	1.06	1.50	1.10	3.60	3.00	3.20	3.20	1.85	2.00
C_{Gr}	C											
C_{cl}	7A	7B	8A	8B	9A	9B	10A	10B	11A	11 B	12 A	12B
C_L	40	45	38	42	37	43	38	31	36	40	39	39
C_I	0.33	0.27	0.21	0.31	0.28	0.41	0.28	0.36	0.30	0.40	0.26	0.40
C_R	2.00	0.26	3.66	2.20	2.50	1.42	2.50	1.75	2.25	1.50	2.75	1.50

5.10.2 Algorithm and experimentation for ILC^2

The first stage of ILC^2 deals with the customization of the training samples S for the KB representation, where $S = \{(ch_1,y_1),(ch_2,y_2)\ldots ch_i,y_i.. ..(ch_m,y_m)\}$ where ch_i and y_i are the chromosomes are their corresponding Denver classes, CL respectively. S_{rep} is the representative vector of extracted features. The feature vector, FV is formed using S_{rep}, threshold value *th*, deviation factor δ, density D of each class and number of comparisons ɳ, and is given as input to the incremental approach with the unlabeled data set UL to classify. UL_i represents an unlabeled chromosome in UL.

ALGORITHM

Step I: KB Building

```
Input:
Training samples S = {(ch$_1$,y$_1$),(ch$_2$,y$_2$)…..(ch$_m$,y$_m$)}
Output:
FV for KB
Begin
Do for LD$_{CL}$ where CL ϵ S
        Set N-$factor$-$_{CL}$ = $0$
        Initialize ch-no to 1, next-ch-no = ch-no+1
        Do until next-ch-no<=n where n is the total no. of chromosomes in CL
                - Compute N-modified( Chromo$_{ch\text{-}no}$, Chromo$_{next\text{-}ch\text{-}no}$ )
                -N-$factor$-$_{CL\text{-}all}$ = N-$factor$-$_{CL}$ +$N$-$modified$
                -$N$-$factor$-$_{CL=}$ $N$-$factor$-$_{CL}$ +$N$-$modified/2$
                -$th$=N-$factor$-$_{CL\text{-}all}$ / ( n-1)+(n-2)+…1
                -$\delta$ = N-$factor$-$_{CL\text{-}all}$/D where D is the density of CL.
        - Compute $F$ and S-val and S-avg
        -Set FV = { $th$, $\delta$, D, S$_{val}$ , $F_{val\_all}$ , S$_{avg\_all}$ , N-$_{factor}$ , S$_{rep,}$ ɳ}
End
```

Step II : Classification And Knowledge Augmentation

```
Input:
UL = {UL$_1$ ,UL$_2$,…,UL$_n$ }
Output:
CL= Classification of UL
Begin
Do For every UL$_i$ ϵ UL
        Do For L$_i$ϵ FV where i ={1,2,…m}; m is the number of classes
                - Compute N-factor(UL$_i$, L$_{CL}$) where L$_{CL}$ ϵ FV
                - Predict probable class, CL$_{pr}$ | N-factor(UL$_i$, L$_{CLpr}$) < (N-factor(UL$_i$ ,
                  LC$_x$ )
                  Where CL$_{pr}$ $\neq$ x and x = {1,,2…m}

                - If (N-factor) $\leq$ $th$ $_{CLpr}$
                          Classify UL$_i$ϵ CL$_{pr}$
```

```
            Else
                        If(N-factor  $\leq$ $th_{CLpr}+\delta_{CLpr}$)
                                Classify the $UL_i \in CL_{pr}$
                        Else
                                Generate new class.
        -   FV= old(FV) + new(FV) for $UL_i$ classified; tune $th$ and $\delta$; update KB
END
```

The ILC^2 approach performs a three-step decision making prior to prediction of class for UL_i. The initial decision making happens by selection of lowest N-val. Further the decision making occurs at threshold and delta check points. However, it must be noted that the incremental step of evolving a new class happens only after this three-fold validation. With every new unlabeled data that is classified using ILC^2, the last step of the algorithm updates and refines the KB by tuning of *th* and δ. Every learning process has an impact on the N-val, ɳ and D resulting in updation of previously formulated FV. Thus ILC^2 performs dynamic tuning of threshold and delta making the classification more accurate. The adaptive nature of the algorithm adds benefit to enhance the Karyotyping. The parameters in the formulation of the feature vectors selected are influential in classification as discussed earlier and play a significant impact on the classification accuracy.

The publically available data set at *http://biolab.dei.unipd.it* developed by Enea et al. [86] is used to validate the efficiency of the algorithm. The performance of the proposed algorithm is tested and analyzed using 1800 chromosomes from 75 images. From every image, 24 chromosomes (group A: 6, group B: 4 and group C: 14) are selected. The entire dataset thus comprises of:

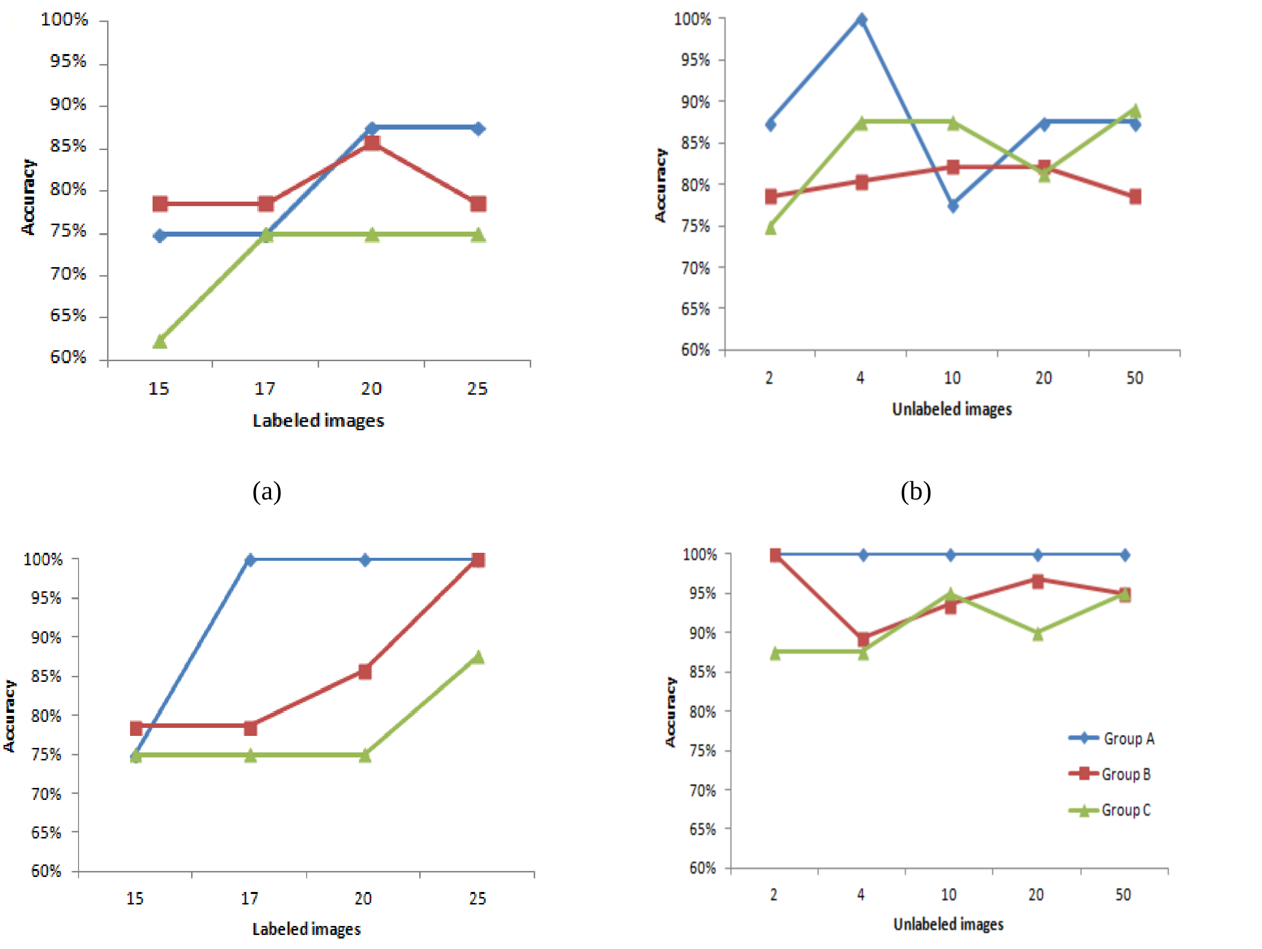
- 25 labelled images (600 chromosomes)
- 50 unlabelled images (1200 chromosomes)

Group A, B and C are the groups with different chromosome characteristics having large, and medium sized, metacentric and sub-metacentric chromosomes. Group C is the largest group, with the maximum number of chromosomes. These groups are therefore chosen for experimentation. For testing the efficiency of the proposed algorithm, two labeled classes namely group A and group B are considered to be initially available. ILC^2 evolves the new class: group C incrementally and also correctly classifies the unlabelled chromosomes of groups A, B and C. The initial KB is generated using only 15 labeled images (90 chromosomes of Group A and 60 chromosomes of Group B) and gradually it is increased to 25 labeled images.

Figure 5.11 depicts the experimental results of testing ILC^2 on the unlabeled chromosomes considering the variations in the number of available labeled images along with static and dynamic approaches. Figure 5.11 (a) and (b) demonstrate the experimental results with static *th* and δ. The KB built using the labeled images plays a significant role in the classification of unlabeled images. It is therefore necessary to estimate the number of labeled images for the initial KB building. Figure 5.11 (a) illustrates the accuracies for 2 unlabeled images considering variations in the set of labeled images. In order to improve the accuracy for other Groups, the number of the labeled images is increased and the final KB is generated using the FV derived out of 25 labeled images. After setting the baseline of 25 labeled images, performance of the algorithm is examined for 50 unlabeled images. Fig. 5.11 (b) depicts this experimentation for unlabeled images in the static environment.

The key factor of ILC^2 is knowledge refinement with every new learning. This is possible with dynamic values of *th* and *δ*. Figure 5.11 (c) shows the experimentation to evaluate the approximate number of labeled images essential for achieving good classification in dynamic environment. When tested on two unlabelled images, 100% accuracy is achieved with 17 labeled images, but only for group A. With 25 labeled images, 100% accuracies are achieved for groups A and B. After the pre-learning from 25 labeled images, the testing is carried with unlabeled images as illustrated in Figure 5.11(d). It is worth noting that the *th* and *δ* values of class C are estimated only after accumulation of 5 images of Class C. Prior to this accumulation, S_{rep} is the decision maker for the classification. With increase in the labeled samples for group A and B, the *th* and δ values appropriately tuned into the suitable range which has a great impact in new class generation. This led to initial variations in classification accuracy for group C and later with the stable volume of data, the knowledge augmentation of newly evolved class: group C is possible. This enhanced its classification accuracy. Adaptation to the changing environment is only possible due to dynamic *th* and *δ* of ILC^2 and thus plays a significant role in knowledge amassing.

Higher classification accuracies are achieved with 20 labeled images. But the number of labeled images, used in KB formation is still intentionally increased from 20 to 25 so as to achieve improved classification of unlabelled images. The KB built with lesser labeled images resulted in lower classification accuracies, when experimented on unlabelled images. It must however be noted that even with 25 labeled images, the results obtained with dynamic values are superior as compared to static approach. One of the most important reasons responsible for less accuracy is the pre-bound values of *th* and δ. In the static approach the values of *th* and δ remain unaltered with the new learning whereas in dynamic scenario, the adaptation to the new dataset occurs. Thus it is observed that static values of *th* and δ confined the role of the classifier only to the stability spectrum, hindering the new learning.



(a)

(b)

(c)

(d)

Figure 5.11: Experimental results to validate the efficiency of ILC^2. The classification accuracy for: (a) Evaluation of number of labeled images (static *th* and δ, 2 unlabeled images) (b) Test data set of unlabeled images (static *th* and δ, 25 labeled images) (c) Evaluation of number of labeled images (dynamic *th* and δ; 2 unlabeled images). (d) Test data set of unlabeled images (dynamic *th* and δ, 25 labeled images).

5.10.3 Discussion on ILC^2

Table 5.4 compares the results of the proposed approach with the ones reported in literature. It must be emphasized, that the previously reported methods are all tested on independent data sets and also on different groups. A direct comparison may therefore not necessarily justify their effectiveness and efficiency but a general idea of the performance might be obtained. The developed algorithm is a successful attempt to address most of the observations mentioned in table 5.4. The issue of CF from the perspective of karyotyping systems is as of now, to the best of our knowledge unreported in the literature. This research identifies this issue and provides a novel solution for it. The developed incremental classifier, ILC^2 being self evolving, learning occurs with both labeled and unlabeled data, without the necessity of re-training and therefore justifies the stability-plasticity balance. The rigidity of earlier classifiers in terms of CF is overruled. Moreover, the algorithm does not require access to previously used data during subsequent incremental learning sessions, because it does not forget previously acquired knowledge.

Table 5.4 Comparison of ILC^{2}with existing methods

Methodologies		Accuracy (%)	Observations
DTW	[61]	81	Heavy computation load
HMM	[50]	≤ 97	Dependency on parameter estimation
ANN -	[86]	94	Need of optimization
	[5]	67 - 97	Low accuracy of group C
	[139]	98	Performed only on group E
GA	[79]	91 - 95	Unable to handle incomplete cells and bend chromosomes
Similarity	[140]	90	Shape dependability
Proposed ILC^2		**97**	

The developed algorithm avoids the necessity of estimating and optimizing the ANN parameters like: number of output neurons, hidden neurons, steepness of the activation function, learning rate, momentum and others. Estimating these parameters needs rigorous experimentation due to unavailability of underlying thumb rules and is extremely tedious.

It will be difficult and time consuming to repeat the experimentations for every new training that is required for every new learning. Another difficulty in repeated retraining is the usage of memory. Recently, AKS is also presented as an image analysis pipeline of banded human chromosomes, where the pipeline is composed of three different stages: an image segmentation step, a feature extraction procedure and a final pattern classification task [39]. Such an approach will be bounded by memory handling constraints and retraining would further complicate the problem. With the proposed method, the since it requires class feature to be preserved in the KB, the memory size is no more an issue. With such recent developments in AKS and also with the use of digital media for biomedical image archiving, storage, and communication, efficient memory handling techniques are highly desirable to accommodate the rapid growth of chromosome image data. The entire process of retraining in ANN, which obviously would require larger usage of memory, needs to be minimized by addressing the problem of CF. Literature reports popular and successful ANN based classifiers for AKS, but to the best of our knowledge the issue of CF is unaddressed and its impact on AKS is unexplored.

The focus of this research is to build an efficient incremental classifier that addresses the issue of CF by avoiding the step of retraining and evolves new classes if required. The developed classifier needs less volume of data for learning. The obtained results are independent of the sequence of images used. Interplays between the current classification and the knowledge update impact the decision making time. Though this factor did not affect the performance of ILC^2, it would certainly be an issue when learning is required from huge data sets. In some cases, incomplete cells (Metaphase image with less than 46 chromosomes) may exist due to the segmentation challenges in the image processing operations or genetic disorders. ILC^2 can also handle karyotyping of such incomplete cells successfully. However, there still remains a scope of improvement with the efficient feature extraction methods which indirectly contribute to the performance of ILC^2.

Summary

An application perspective for incremental learning in AKS is experimented. Karyotyping is a very challenging task and the evolution of new data requires minimal storage requirements with incremental learning. With the growing size of the images and to perform automation in chromosome analysis, new class generation in the dynamic scenario necessarily will be a boon in the AKS.

The proposed algorithm showcases its strength in terms of new class generation and can be reliably extended further for rest of the classes in Denver group. Another serious issue in AKS is identification of the cluster of touching and overlapping chromosomes from the isolated ones. ILC^2 can be very much looked upon to evolve such a cluster as a new class. The broader scope of the ILC^2 includes selective incremental learning that ranks the required feature sets and the images for learning. Selective in terms of features for learning will ensure optimum generation of KB thus boosting its performance significantly.

Another factor mentioned and reported in literature is the CF issue in the karyotyping process. In the next chapter we evaluate method to address the same with a cloning approach for chromosomes.