Adaptive Chromosome Diagnosis Based on Scaling Hierarchical Clusters

Muhammed Akif Ağca, Cihan Taştan, Kadir Üstün, and Ibrahim Halil Giden

1 Introduction to Feature Extraction and Chromosome Classification on Massive Data

Healthcare analytical applications generate massive data including text/image/video in both batch and streaming contexts, which consist of a growing volume in petabytes for a single average-size city hospital that possesses micro-macro scale real-time tracking/monitoring devices. Furthermore, such hospital systems require tracking real-time responses both during diagnosis and historical analytics for trusted results. Therefore, the veracity of the growing datum is highly controversial, in which trust is an obligation to give vital decisions.

In this study, we unify the resources and apply real-time trusted analytical models just by focusing on scalable chromosome diagnostic models. Picture archiving and communication systems (PACS) are considered as storage data formats and HL7/similar ones to unify the resources. PACS are a prerequisite in data-exchange mission in modern hospitals to enable routing, retrieving, and storing medical images [1]. Specifically, images from radiology department such as X-ray, magnetic resonance imaging, and computed tomography are the main tasks of PACS systems to facilitate the workflow. Here, we integrate the chromosome karyotyping

M. A. Ağca (✉)
TOBB ETU/Computer Engineering, Ankara, Turkey
e-mail: akif.agca@etu.edu.tr

C. Taştan
Acıbadem Labcell Laboratory,İstanbul, Turkey
e-mail: cihan.tastan@acibademlabcell.com.tr

K. Üstün · I. H. Giden
TOBB ETU/Electronics Engineering, Ankara, Turkey
e-mail: k.ustun@etu.edu.tr; igiden@etu.edu.tr

H. R. Arabnia et al. (eds.), *Advances in Artificial Intelligence and Applied Cognitive Computing*, Transactions on Computational Science and Computational Intelligence, https://doi.org/10.1007/978-3-030-70296-0_45

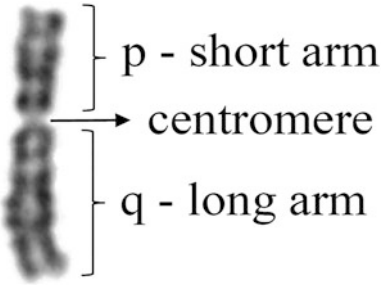

Fig. 1 Chromosome image with its geometrical features

resources, which are used for disease characterization, and our analytical models in a trusted manner to co-operate with other resources by considering the trustworthiness of the overall system for the vital decisions.

Chromosomes are organic structures found in the cell nucleus that carries genetic information. A healthy human cell includes 23 chromosome pairs (22 pairs of autosomes, classes 1–22; a sex genosome, either XX for females or XY for males). Microscopic images of the chromosomes are taken at the metaphase stage of the cell division for identification of chromosomes as well as their classification, i.e., karyotyping, since the chromosomes can be easily distinguished at that stage [2]. Genetic disease diagnosis at an early stage is an important task for proper medical treatment processes, and thus, morphological analysis of chromosomes becomes very crucial. Chromosomal abnormalities that cause genetic diseases can be exemplified as having an improper number of chromosomes (monosomy/trisomy), translocation, deletion, or inversion of chromosomes. Such morphological abnormalities of chromosomes could be associated with genetic diseases and, especially, with many cancer diseases [3].

In feature extraction systems, a reliable chromosome classification is possible if feature vectors are defined appropriately. The common features used in automatic chromosome classification algorithms are mainly geometrical and banding pattern-based features [4]: The length of the chromosome and its centromeric index (CI), a ratio of the short arm (p) with total length of the chromosome ($p + q$), are widely used geometrical features. As an example, an individual chromosome is given in Fig. 1 with its geometrical features $\{p,q\}$.

Efficient chromosome diagnosis systems require dynamic feature extraction and object detection mechanisms. Multiple-object detection is one of the challenging problems in computer vision application domains. Furthermore, real-time object images may not be sufficiently clear and separation of the objects from the background may usually not be adequate, which can be considered as other arising problems in computer vision systems. In order to circumvent such limitations, different object detection algorithms have already been developed in the literature [4–9]. Adaptive sub-modularity can be considered as one of prominent image processing techniques for the detection of multiple objects [10]. As a solution to dealing with large-scale data sets, a submodular function maximization method is proposed in a distributed fashion, whose effectiveness in massive data such as sparse

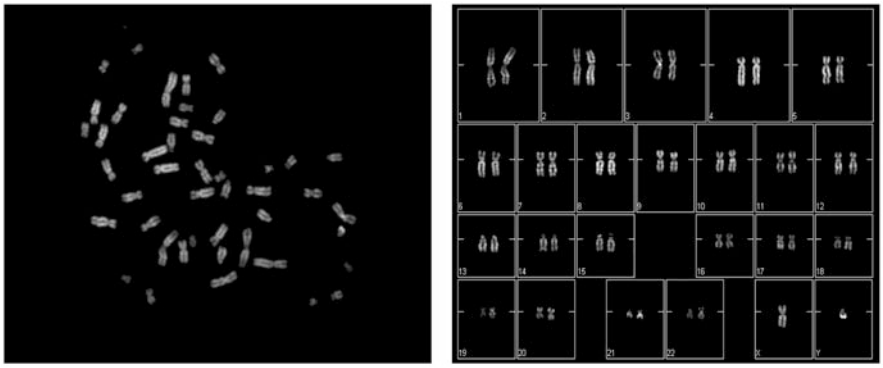

Fig. 2 (**a**) Typical Q-band chromosome images taken at metaphase of cell division and (**b**) their karyotyping [15]

Gaussian process inference and exemplar-based clustering is demonstrated in [11]. As additional types of chromosome classifiers, different algorithms have already been studied in literature such as artificial neural network-based classifiers [12], support-vector machine [13], and fuzzy logic–based classifiers [14].

In this study, a scalable and dynamic feature-extraction model is developed to manage the specified features in chromosome dataset. A distributed clustering system and a memory-centric analytical infrastructure are implemented for high-speed data processing of chromosome images, which can be built on the current PACS systems in hospitals. Every class of chromosomes in different cells is compared in terms of their features and with its homolog chromosome and the cell having potential genetic disorder is diagnosed, accordingly. The chromosomal features are firstly extracted from publicly downloadable (BioImlab) chromosome-image dataset [15], which includes 119 human cells with 5474 individual Q-band chromosome images. A typical chromosome cell at mesophase and its karyotyping is demonstrated in Fig. 2.

The remainder of the chapter is as follows: Sect. 2 discusses the scalable trusted computing technique briefly employed in this study for the classification process of the chromosomes. In Sect. 3, the feature extraction model is discussed in detail; the raw chromosome images, feature selection techniques, and the establishment classification models based on the features are presented. Finally, a conclusion is provided in the last section.

2 Proposed Chromosome Classification Model

In this study, a hierarchical multi-layer neural networks and tree structures are considered as a chromosome classifier that processes the genome data in hierarchical fashion. Each layer gathers the chromosome image features as an input data; automatically vectorizes and merges the matrices; analyses the feature of genes and encodes the data; and sends the feature sets to the next layer.

As universal health standardization, Health-Level Seven (HL7) information modeling is implemented for the genome analyses [16]. HL7 follows an object-oriented modeling technique that is organized into class diagrams (graphs) with attributes of classes. The extracted genomic data in our study will be transported in the distributed system via predefined encapsulating HL7 objects and the raw genomic data will be sorted out for the selection of specific genes at the memory-centric section of the system. Later, the feature extraction of the specified genes will be provided with the distributed system via training algorithms, and then, new genetic data will be classified comparing with formerly collected genetic features.

As a network communication model in HL7 standards, Open System Interconnection (OSI) reference model is used. As a hierarchical network architecture, the OSI model includes seven layers in a logical progression (see Fig. 3). Briefly stated, the seven OSI layers have the following assignments [17]:

- *Application layer* includes network software (user interface and application features) directly served to users.
- *Presentation layer* converts the information into specific formats (encryption, compression, conversion, and so on). In our case, the raw genomic data is processed.
- *Session layer* controls end-to-end connections between the applications in different nodes.
- *Transport layer* converts the received genome data from the upper layers into segments.
- *Network layer* is responsible for path determination and datagrams generation.
- *Data links* provides error control as well as "transparent" network services to physical layer.
- *Physical layer* enables direct communication with the physical media.

Comparing both TCP/IP and OSI reference models, TCP/IP has four layers—combining the application, presentation, and session layers into one top layer; taking both data-link and physical layers as a bottom (Network) layer (see Fig. 3).

Storing, printing, and transmitting of chromosome image information is handled by the Digital Imaging and Communications in Medicine (DICOM) protocol, which includes application protocol in TCP/IP model to communicate among the system. Each chromosome image is defined as a data object and the data exchange is conducted in image format. Complete encoding of medical data is standardized by DICOM protocol with attributes named as "DICOM data dictionary."

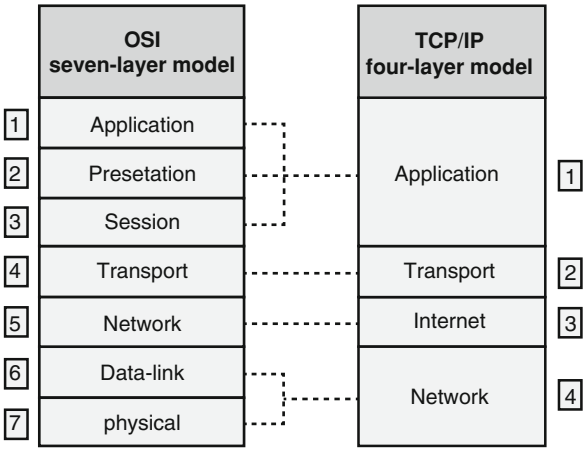

Fig. 3 Hierarchical architectures of OSI and TCP/IP reference models

DICOM stores the images from the archive, and if they are needed, DICOM provides association between service class users (SCUs) and service class providers (SCPs). Different protocols are also available for specific purposes like disease names, clinical context management, and hospital data acquisition aims. The current global standards can be exemplified as ICD (International Classification of Disease), NIST (National Institute of Standards and Technology), HL7 CDA (Clinical Document Architecture), CCR (Continuity of Care Record), CCOW (Clinical Context Management Specification), LOINC (Logical Observation Identifiers Names and Codes), ELINCS, EHR-Lab (Electronic Health Record - Lab), X12, SNOMED (Systematized Nomenclature of Medicine Clinical Terms), NCPDP (National Council for Prescription Drug Programs), IHE (Integrating the Healthcare Enterprise), CCHIT (Certification Commission for Healthcare Information Technology), HITSP (Healthcare Information Technology Standards Panel), CMMI (Capability Maturity Model Integration), ISO 27001, TS 13298, OHSAS 18001, Medical Data Interchange Standard (MEDIX), ASTM E1238, IEEE 1073 and ASC X12N. Furthermore, ASTM E31.17 and ASTM E31.20 standards are applied for data security and privacy. The standards are explored to manage the transactions in a scalable/trusted manner.

The studied feature extraction method will be described in detail in the following subsections.

2.1 Proposed Feature Extraction Method

For high-precision classification, it is very important to have best-characterized chromosome images. For that reason, better features should be selected to improve classification performance. It is also critical to note that well-described features

Fig. 4 Original image of Chromosome 1 pair [15]

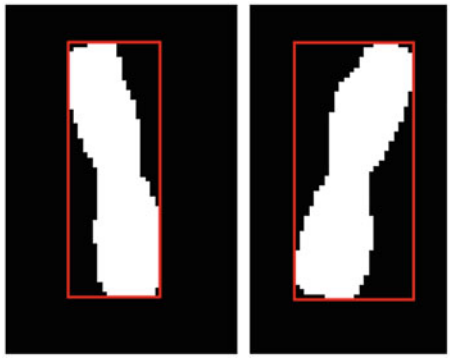

Fig. 5 Chromosome 1 pair image (in Fig. 4) after binarization. Solid rectangles indicate the region above the predefined threshold

reduce the processing workload and thus offer greater success of chromosome classification in less processing time.

2.1.1 Adaptive Thresholding to Individual Chromosome Images

An adaptive threshold value must be determined because the quality of each chromosome image may differ. A boundary detection-based thresholding is applied for efficient feature extraction. Chromosome 1 pair is chosen as an example for the application of the proposed feature extraction method (see Fig. 4).

2.1.2 Chromosomes-Image Binarization and Skeletonization

After thresholding with an adaptive value for each chromosome image as in Fig. 4, binarization of the image is conducted and the chromosome region is detected from the image (see Fig. 5). Later, a skeletonization algorithm is carried out for the binary version of the chromosome image, which can be considered as a landmark for the feature extraction (see Fig. 6).

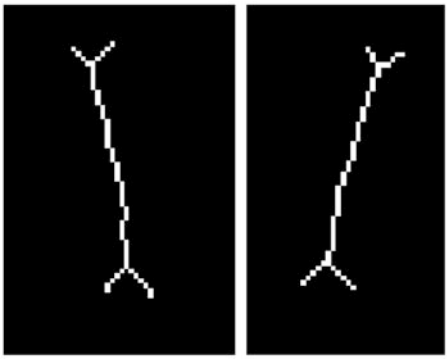

Fig. 6 Skeletons of Chromosome 1 pair image (in Fig. 4) extracted by using the skeletonization algorithm

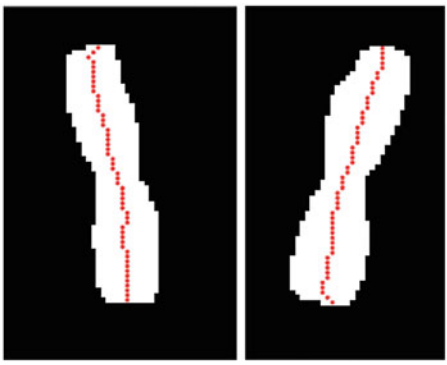

Fig. 7 Binary images of Chromosome 1 pair with its medial axes

2.1.3 Medial Axis Estimation of Chromosome Pair

Determination of medial axis is an essential task for the feature extraction of chromosome pairs. Medial axis estimation provides detection of centromere location and its indexing, chromosome length estimation, and density profile extraction [18]. For that purpose, the chromosome image in Fig. 5 is scanned with lines to determine boundaries of the binary image. The mid-points of every scanned line are found using a mid-point algorithm, which enables an accurate medial axis estimation. Estimated medial axes for the binary images in Fig. 5 are drawn and represented in Fig. 7.

2.1.4 Gray Level Extraction Along Medial Axis

After the determination of medial axes, a curve-fitting is performed with a 5th-degree polynomial function. Using this approach, the chromosome alignment can be mathematically derived. The next step is to resolve the chromosome image by slicing an equal number of perpendicular lines to the medial axes (see Fig. 8 for detailed representation). For that purpose, the norm of the polynomial curve is

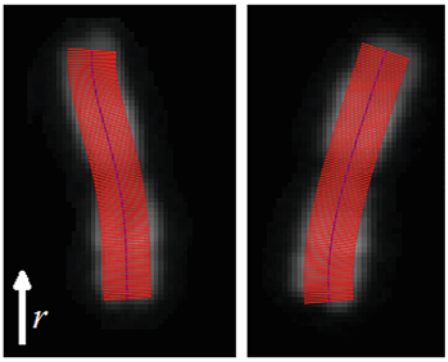

Fig. 8 Polynomial curve-fitting of medial axes on the chromosome images and their slicing with perpendicular lines

determined, and the curve is sliced equally for every adjacent perpendicular line. It should be noted that the solid lines are in parallel, and hence, they do not intersect with each other. The distance vector r is given as an inset in Fig. 8.

The next step is extraction of the gray levels through the norm of medial curve. To do this, a new parameter called Histogram, $H(r)$, of the chromosome image is calculated by the following equation:

$$h(r) = \sum_{k=1}^{n} \frac{\overrightarrow{g_k} \cdot \overrightarrow{u_k}}{n}, \tag{1}$$

where $\overrightarrow{g_k}$ denotes the gray-level vector along the perpendicular line $\overrightarrow{u_k}$ and n is the number of perpendicular lines, viz. pixels. Corresponding histogram function dictates the average density of pixels along the medial axis. The $h(r)$ calculation is conducted for the chromosome pair in Fig. 4 and plotted in Fig. 9. Some important remarks could be gathered from the histogram plot: (1) End-points of the chromosome pairs could be inferred from the histogram plot; (2) centromere position, which is the lowest point of the curve, could be precisely determined by using the histogram plot. In this case, corresponding centromere-index (CI) could also be calculated from Fig. 9, which is found to be CI = 0.46 for chromosome 1a and CI = 0.49 for chromosome 1b.

2.1.5 The Feature Extraction from Histogram Information

As a common shape signature extraction of image data, frequency response calculation is implemented in image processing systems [19]. Therefore, discrete Fourier transform (DFT) of the histogram function is taken as a feature extractor $H(f)$ for precise chromosome classification. The feature extractor function is calculated via the following equation:

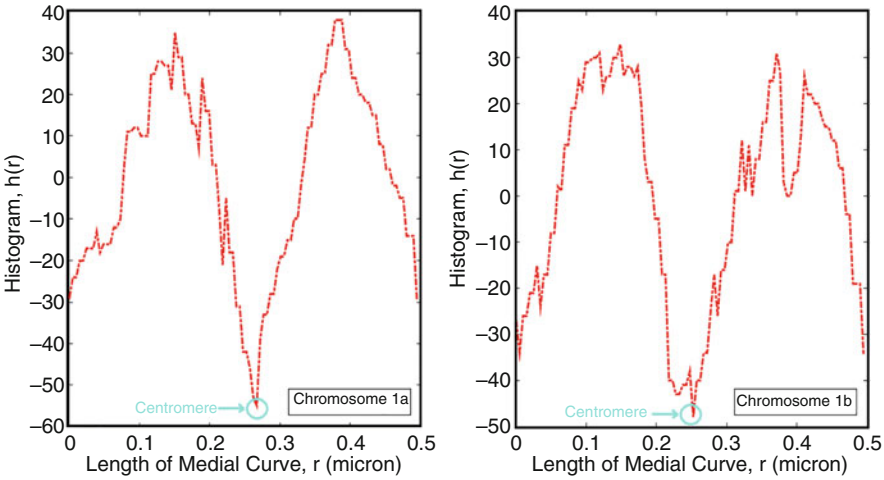

Fig. 9 Histogram function calculation along the length of the medial curve

$$H(f) = \sum_{k=1}^{n} h_k(r) e^{-i2\pi kr/n}. \tag{2}$$

Absolute values of $|H(f)|$ is plotted in terms of frequency (see Fig. 10). There are two critical points to infer from Eq. (2): The centromere location, i.e., CI of the chromosome, directly influences the frequency of the main peak, which is found to be 4 Hz for Chromosome 1a whereas that value equals 4.02 Hz for Chromosome 1b (see Fig. 10). It is important to note that the reason of being the two frequencies nearly the same is due to the nearly equal values of CI. Another important remark can be inferred while comparing Figs. 9 and 10 that the average value of histogram function $h(r)$ affects the absolute value of corresponding feature extractor, $|H(f)|$.

As an alternative, the feature extraction model can be extended by using hyper-spectral images of the chromosomes. Namely, chromosomes images under different wavelengths of illumination can give us additional information for the diagnosis of disease carrying/faulty chromosomes. It is known that biological molecules and cells have footprints in the mid-infrared and terahertz parts of the electromagnetic spectrum [20]. Combining the image and spectral data at these wavelengths would improve the performance of the classification algorithm significantly. We discuss further improvements on feature extraction and classification in detail in Sect. 2.3.

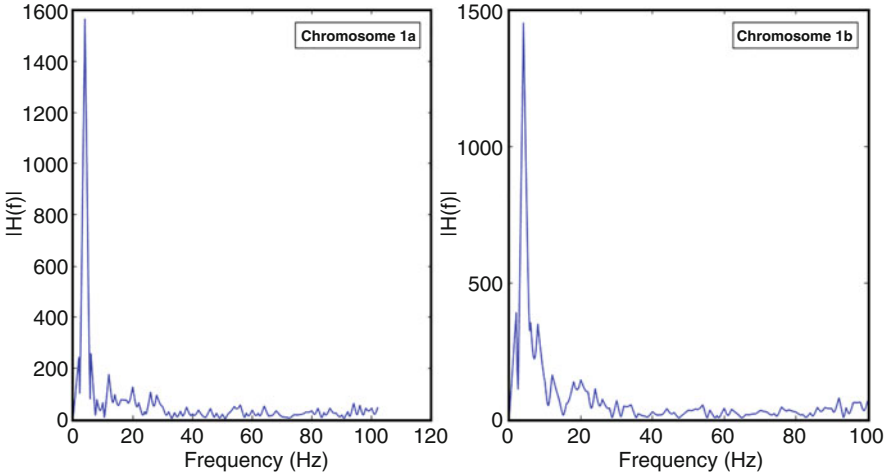

Fig. 10 DFT function $|H(f)|$ as a feature extractor calculation for the Chromosome 1 pair in Fig. 4

2.2 Chromosomal Abnormalities–Related Disease Diagnosis by Spectral Chromosome Information and Other Novel Features

Humans have 23 pairs of chromosomes which have various numbers of genes and differ in their structures and sizes (i.e., male Y and female X chromosome). Structural and numerical chromosomal alterations are identified in ~0.6% of births [21], which frequently cause developmental disabilities, mental retardation, birth defects, and dysmorphism [22]. Many genetic or rare diseases are caused by the chromosomal abnormalities, which are diagnosed with a determined gain or loss of some genomic materials coding hundreds of genes. Unlike single-gene mutations, the gain or loss in chromosome complexes directly influences gene expression doses which can cause imbalances in phenotype and human body functions.

Alternative abnormalities can be microdeletion or microduplication of chromosomes, all of which can be identified via conventional visual analysis of chromosomes under microscope (karyotyping). In contrast to numeric chromosomal abnormalities, which occur when three copies of chromosome present rather than two such as Down syndrome (trisomy of chromosome 21), structural abnormalities are caused from the breakage or rejoining of chromosome arms. This can result in deletions of chromosome segments.

The chromosomal alterations can cause the imbalances in a contiguous gene syndrome, in which multiple genes are negatively influenced, resulting in combinatorial abnormalities in clinical phenotype [23]. On the other hand, most of the chromosomal deletion syndromes (such as Williams syndrome, Miller-Dieker syndrome and DiGeorge syndrome) result in haploinsufficiency of a gene or genes

in the lost segment where single copy of other allele gene does not express sufficient doses of the gene product for a normal and healthy state phenotype [22]. Chromosomes can be visualized as lighter and darker bands which are considered as sections/segments following certain staining protocols and classical microscopy.

However, the newly developed spectral chromosome technologies enable the characterization of the disorders in a high throughput manner via super-resolution visualization with lab-on-a-chip (LOC) microfluidics and microscopy in a real-time fashion. The chromosome imaging and spectral analysis platform can address disease-related abnormalities with respect to healthy and homolog chromosomal information. Therefore, any subtle alterations or rearrangements of the segments in a chromosome (micro-level deletions or reunions syndromes) can be easily imaged and detected using high-resolution chromosome spectral analysis, enabling the identification of unbalanced abnormalities that are invisible under conventional microscopy technique.

The use of high-resolution LOC spectral chromosome analysis can allow cytogenetic field both to diagnose the pre-characterized genetic syndromes and to identify new disorders. With an automated chromosome spectral analysis through an adaptive optimization method of karyotyping schemes, numerical or structural abnormalities in chromosomes can be detected highly efficient and robust manner while diagnosing cancers and other genetic disorders. For instance, in diagnosed chronic myeloid leukemia (CML) patients, abnormal pattern named *t* (9;22) chromosomal translocation is identified in one chromosome 9 and one chromosome 22 [24].

A study reported a developed computerized scheme to automatically identify the unbalanced alteration in one chromosome 22 of abnormal cells in CML patients by analyzing and image-processing karyotypes of metaphase cells from bone marrow patient specimens [25]. Extracting novel spectral features for real-time diagnostic requires using maximum available computational resources in minimum physical space. The minimization of physical space can be achieved via various innovations such as 3D-Stack packaging OBCs and ASICs including micro-channel cooling/transfer, nano-sensing, custom processor CPU/GPU/FPGA on a single board, which we discussed in the last chapter.

Extracted genetic disorder features can be efficiently characterized and diagnosed using a genetic algorithm, artificial neural networks and statistical models along with graph construction for genetic syndrome-related chromosomal spectral information, extended data flow blocks and computational units' path tracking like in biological system design tool, BioGuide [26]. The software tool standardizes genetic elements and provides graph data structures of all possible genetic element interactions to allow molecular biologists to estimate and characterize phenotypically functional genetic circuits. BioGuide constructs a massive graph structure $G = (V,E)$, where $V = \{g_1, g_2, g_3, \ldots, g_n\}$ each element is a part of genetic circuit. $E = \{e_1,e_2,e_3,\ldots,e_m\}$ each edge $e_m = \{g_s,g_t\}$ represents the possible connections between the elements, which build a genetic circuit. In this way, complex sequence relations can be modelled effectively.

Nevertheless, tracking the complex sequences requires scalable data flow and transaction flow management in a trusted manner. Both analytical models and the system are needed to scale at a massive level. As illustrated in Fig. 14, the transaction is managed by layered system components. Scaling the models and the system requires memory-centric analytical approaches, which is tracking the lineage of datum, transaction, and the models. In another study, we proposed a Memory Centric Analytics (MEMCA) system that provides a holistic abstraction to ensure trusted scaling and memory speed trusted analytics. Initial results are shared in [27] MEMCA used for satellite and space data. The results evidenced the potentials for spectrum analysis of satellite and space datum. Thus, the use of MEMCA approach for other kinds of high-resolution spectral imagery data may have great potential for novel studies.

The proposed scalable analytical model for chromosome diagnosis can be utilized in cytogenetic fields as well to diagnose the pre-characterized genetic syndromes and to identify new disorders, which will be further discussed in the next section. The system architecture will be explored in Sect. 2.3 to track the complex sequences and to apply the analytical models on the massive data at memory speed in a trusted manner.

2.3 Genetic Disorder Detection via Spectrum Analysis and Scalable Multi-layer Neural Networks

In this study, an open-public data set is [15] used for single-chromosome images of different human cells. The chromosomes of the same class are analyzed with the abovementioned spectral analysis approach in the cases of different cells, and the obtained FFT spectra are compared altogether. Important outcomes are gathered from the spectral information comparison: The shape of chromosome as well as density profile of genes directly influences the spectral feature. For that reason, comparing the FFT spectral feature enables to detect corresponding genome diseases, chromosome diagnostic, and so on. Spectral feature extraction could be performed via OBCs and ASICs (2D/3D Stacks) including nano-sensor plates, which will be discussed later in detail.

Combining the image and spectral data from different types of measurement systems would improve the performance of the classification algorithm significantly. Specifically, the feature extraction stage can be extended by using hyper-spectral images and other spectroscopic data related to chromosomes. Namely, images of the chromosomes under different wavelengths of illumination can give us additional information for the diagnosis of decease carrying/faulty chromosomes. It is known that biological molecules and cells have foot-prints in the mid-infrared and terahertz parts of the electromagnetic spectrum [20].

Another option is to use Raman spectral response of chromosomes to gather extra features for the classification algorithm. Indeed, Ojeda et al. [28] have

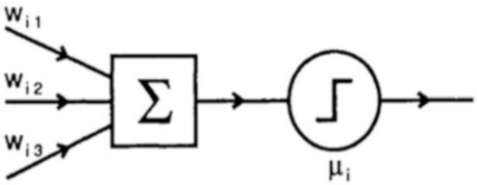

Fig. 11 A simple analytical model of neurons as a binary threshold unit

managed to differentiate the Raman spectroscopy response of three different types of chromosomes by optical tweezing and exposing the chromosomes to a laser beam; an interesting method for differentiating colon cancer cells is just measuring the dry mass of chromosomes by tomographic phase microscopy (actually the refractive index distribution is measured) [29].

Another important method which is very beneficial for analysis of chemical and biological substances is Fourier transform infrared spectroscopy (FTIR). FTIR is already proposed for the diagnosis of cancer [30]. In one of the studies, FTIR is used for comparing the chromosomes taken from breast cancer cells and healthy cells [31]. As another genetic diagnostic imaging approach, quantum dots such as gold nanoparticles are used for in vivo bio-imaging of subcellular organelles as well as genetic diagnosis [32]. Nanowires are another alternative platform for efficient label-free nano-sensing of DNAs and selective electrical detection of genetic disorders [33]. Graphene-based bio-devices can also be considered as a powerful biosensor for monitoring the chromosomes' behaviors as well as extracting their electro-chemical features [34].

The datum, including the features, has varying resources and computational complexity/cost increases the complexity of analytical model and training process. We define a multi-layer neural network to classify and embed the new features dynamically obtained from available data sources. Neural computation is an efficient method to represent complex sequences inspired from neuroscience. Analytically, a neuron can be simply modeled as a binary threshold unit, w: Each neuron model computes the weighted sum of its inputs from other connected units and outputs either "0" or "1" according to the threshold level, μ (see Fig. 11). The output state of neurons, n, is calculated via the following formula:

$$n_i\,(t+1) = \Theta\left(\sum_j w_{ij} n_j(t) - \mu_i\right). \tag{3}$$

Here, Θ is a unit step-function representing the following characteristic:

$$\Theta(x) = \begin{cases} 1, & \text{for } x \geq 0 \\ 0, & \text{elsewhere.} \end{cases} \tag{4}$$

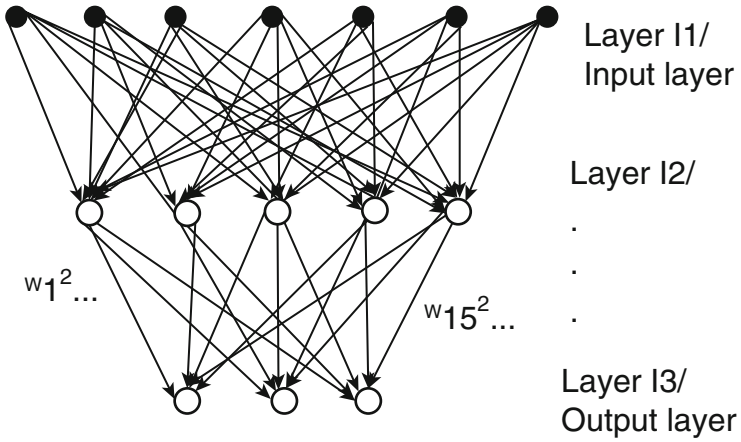

Fig. 12 Illustration of MNNs intra-node and intra-layer interaction

The output state n_i becomes either "1" (*firing* state) or "0" (*not firing* state) depending on the inputs as well as the threshold level. In this way, real neurons could be translated as computing units since the output states are described as binary units [35].

In reality, it is not as simple as in Eq. (3) to determine the behavior of real neurons since (1) the input/output relationships are nonlinear, (2) neurons produce continuous output pulses rather than a simple output level, and (3) neurons do not works synchronically as Eq. (3) states. For that reason, the output state, n_i, should be defined as continuous-valued function like the following formula:

$$n_i = g\left(\sum_j w_{ij} n_j - \mu_i\right). \tag{5}$$

In this equation, a new function g called activation function is added to the neural system.

Human brain itself can be described as a parallel system of billions of processors (neurons), which are highly connected into each other via synapses and operate simultaneously. Therefore, neural networks can be considered as an efficient way for parallel processing of massive data. For that purpose, neural network systems are described as multi-layer architecture (perceptron) to make applicable in computational networks (see Fig. 12).

In the studied feature extraction model, each chromosome is taken as a neuron and operates as a feature vector generator in the multi-layer neural network system. The implemented image feature classifiers such as binarization, skeletonization, gray-level and medial axis detection, thresholding, histogram, and DFT vectors are set to be multi-layers of the proposed neural network system. At every layer, the

activation function g is adaptively selected to infer spectral chromosome feature for real-time diagnostic purposes.

Each neuron takes as input w_{i1}, w_{i2}, w_{i3}, …, w_{in}, feature sets and outputs an activation function g(.), triggering the function according to μ_i threshold value. As illustrated in Fig. 12 perceptron/multi-layer architecture is defined to denote layer-wise implementation. There are L number of scalable layers denoted as $L = \{l_1, l_2, \ldots, l_n\}$. Connection between layers is denoted as $w_{ij}^{(l)}$, where unit j in layer l is connected with unit i in layer $l + 1$. The activation of unit i in layer l is denoted as $\mu_i^{(l)}$. Activation function computation is as follows:

$$\begin{aligned}
\mu_1^{(2)} &= g\left(w_{11}^{(1)} + w_{12}^{(1)} + w_{13}^{(1)} + \ldots\right) \\
\mu_2^{(2)} &= g\left(w_{21}^{(1)} + w_{22}^{(1)} + w_{23}^{(1)} + \ldots\right) \\
\mu_3^{(2)} &= g\left(w_{31}^{(1)} + w_{32}^{(1)} + w_{33}^{(1)} + \ldots\right) \\
&\ldots
\end{aligned}$$

Each layer is trained dynamically; training sets, test sets, and activation functions are updated according to upcoming extracted features. Below is generalized sudo-code for the multi-layer neural network building/training process.

1. **Set** initial feature sets $\boldsymbol{w_{ij}^{(l)} = 0}$ for each neuron and all $\boldsymbol{l}$
2. Embed new features to each neuron
3. Extract new features with CNN/RNN
4. For $\boldsymbol{i = 1}$ to $\boldsymbol{n}$
    1. Update activation function g(.) for each $\boldsymbol{w_{ij}^{(l)}}$.
    2. Update features $\boldsymbol{w_{ij}^{(l)}}$ for each neuron
    3. Update output state $\boldsymbol{n_i}$ for each neuron

    Repeat until $\sum_j w_{ij}^{(l)}$ of each neuron >= $\boldsymbol{\mu_i}$

Activating the functions $g(.)$ in appropriate layer, hidden layer or upper ones, improves the training performance and decreases computational cost notably. Reference [36] is discussing the placement of activation functions, recommending the replacement of activation functions of lower layers with MNNs in a limited computing resources case.

To extract new features based on the available sets, limit the interaction between the neurons to computationally feasible level, we restrict the connections between hidden layers inter/intra layers, and define stimulus locations for the images and activation functions. Layer-wise implementation of multi-layer neural networks enables efficient abstraction of stimulus features like centromere indexes, skeleton, polynomial curves, histograms, discrete Fourier transform (DFT), and additional spectral features we pursue research. Convolutional/subsampling layers are defined in the fully connected MNN structure; 2D structural advantage of images/signals is used with the local connections.

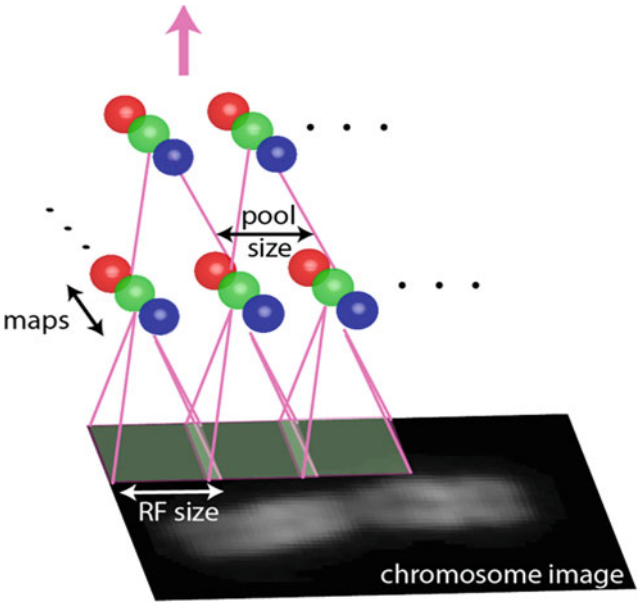

Fig. 13 First layer of CNN with mean/max pooling

The image sets are given as input to convolutional layer as $m*m*r$ image, where m = height/width, and $r = 3$, RGB channels in this set (see Fig. 13) for the illustration of CNN multi-layer system architecture. Other channels with spectral and signal features can also be added into the system modeling. There are kernels/filters k have size $n*n*q$, where the size of kernel $n < m$ and $q \leq r$. Kernel size raises the locally connected structure, convolved with the image to produce k feature maps with the size of $m - n + 1$. The maps are sub-sampled with mean or max pooling over $p*p$ contiguous regions. Size of pool is variant, $2 \leq p \leq 5$ up to size of image, greater for larger input images. Bias and sigmoidal nonlinearity is applied to each feature map, before or after subsampling layers. After the convolutional layers, there may be any number of fully connected layers.

The implemented neural network system needs a training feature set $\{t_1,t_2,\ldots,t_n\}$; in our case, t_n denotes known image features of either healthy chromosomes or with genetic diseases. A cost function can be described with the following equation to prevent possible over-fitting conditions:

$$J_n^{(l)} = \left\| \mu_n^{(l)} - t_n \right\|^2. \tag{6}$$

Equation (6) implies that the implemented CNN model tries to converge the output feature set to the training set by minimizing the differentiation between the two sets and thus reducing the overall cost function $J_n^{(l)}$. The train and output feature datasets/matrices are vectorized dynamically and served to a scalable memory-

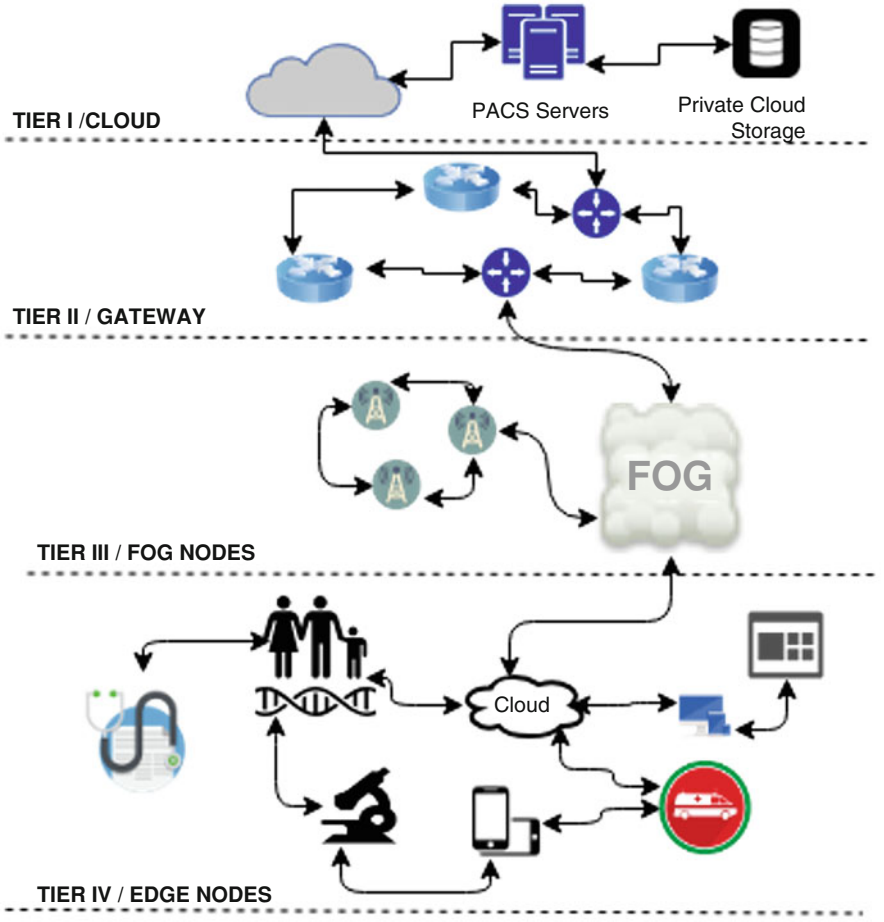

Fig. 14 A healthcare analytical cloud system overview

centric clustering algorithm, which is implemented at the output and training feature datasets in a distributed and trusted manner. In this way, new anomaly detection and classification is provided with the help of disjoint computational nodes, and hence, the requested genetic feature comparison can be conducted in real time with faster data processing. Further hierarchical system architecture is schematically represented in Fig. 14.

Placement of activation functions is a non-deterministically triggering computational resource. We implemented our memory-centric analytical approach for trusted analytics as illustrated in Fig. 14. MEMCA is a holistic abstraction to ensure trusted scaling and memory speed trusted analytics [37]. Using Markovian chains and distributed ledger-based structures is an effective way to keep the data sources fresh and to execute the transactions trustfully. Managing non-linearly growing

transaction in trusted manner ensures scalability and elasticity of the system as much as possible up to currently available hardware and physical constraints. Trusted execution of the system enables trusted scaling of algorithms in real time. We extend the analytical models in trusted manner via the MEMCA abstraction. Large feature sets/matrices are dynamically embedded to the growing models. The transactions including private data are verified with checksum values at appropriate check-pointing locations. Thereby, the integrity between cloud, gateway, fog nodes, and edge nodes layers is ensured. The proposed system is highly scalable and trusted to manage the growing datum and emerging edge devices to extract new features like spectral data and other custom ASIC and OBC devices.

ASIC designs congregate different types of detectors and processing units. These detectors may belong to different physical domains such as pressure, heat, and electromagnetics. Moreover, electromagnetic wave detectors can also be classified according to the wavelength (frequency) band of operation, lying between microwave regime and up to ultraviolet and X-rays. Especially, infrared and terahertz bands are extremely useful for the characterization of chemical and biological samples due to their specific footprints [38]. Their integration to ASIC systems is very important in that respect. Infrared detectors are mostly based on semiconductor technology and hence their integration to the ASIC integration is not very challenging, except in some cases where the detectors should be cooled down to ultra-low temperatures.

In the terahertz detector case, there are many different detector types with different operation principles. Here we are interested in the photoconductive antennas, because of its simple design and low cost. These antennas would occupy a planar region in the order of the wavelength of operation (e.g., 1 THz corresponds to a wavelength of 300 μm). However, a lens system and a detector array may be needed for far-field imaging systems, which can be useful for defense and testing. The size of the detector array and the lens may differ according to the target resolution and image size, where increasing the size may show some integration problems to the ASIC architecture.

On the other hand, such far-field detection systems may become separated from the ASIC system and a communication link may be established for the control and date transfer between these modules. 3D-Stack packaging enables novel miniaturized OBCs and ASICs into highly space-efficient architectures, where the micro-channel cooling/transfer can remedy any extreme heat generation and hot-spot issues (references above). The space efficiency can be further improved by nano-sensing, custom processor CPU/GPU/FPGA/Storage on a single board, and memory speed compact storage given the available physical limits [39]. The novel micro-machines packaged with the innovative method is handling computational bottlenecks in limited physical space and enabling to extract novel features for real-time diagnostic applications. We develop customs ICs for specific purposes up to requirements, which will be discussed in detail in another study.

The datum/transaction flowing in the system is verified with periodical checksums at available check-pointing locations to enhance overall trustworthiness of the system. Lineage data enhances fault recovery and decreasing up-time to

milliseconds in case of any failure, which enables memory speed scalable/trusted analytics on the massive datum processed by thousands of transactions in real time.

3 Conclusions

Chromosomal abnormalities are identified using conventional microscopy techniques after classic karyotyping protocols. Fast, efficient, and trustable bioinformatical tools can help genetic diagnosis laboratories to characterize genetic disorders with a high-throughput manner. Here, we developed an adaptable and dynamic feature-extraction model, which utilizes scalable and hierarchical chromosomal data sets. Chromosomal alterations include numerical (i.e., trisomy) or structural rearrangements (i.e., micro-deletions or rejoining) of chromosomes, which can be identified by comparing with healthy homolog chromosome. Our study characterized chromosomal structures by spectral analysis in detail, where solid rectangles, skeletons, medial axes, polynomial curve-fitting of medial axes, length of the medial curve, and DFT functions of regions of the chromosome pair were extracted. This deep information of the chromosomal spectral analysis for a chromosome pair provides a massive data set for an efficient characterization of a chromosome pair. Different soft computing techniques are developed to classify chromosomes based on the extracted features using the Copenhag data base.

We further improved our classification and feature extraction algorithms with multi-layer networks and convolutional neural networks. The proposed classification system is compatible with HL7 standards and efficient for massive data acquisition and storage as well as real-time data-stream processing. It is also compatible with spectral analyses and edge computing methods, which could be considered as another superiority of the proposed chromosome classifier tool. The proposed system operates via a distributed and scalable algorithm, which provides elasticity on the number of edges/clients.

Placement of activation functions differentiates the behavioral pattern of computational resource triggering non-deterministically. We implemented our MEMCA (memory-centric analytical) abstraction for trusted analytics so that chromosome feature data is transferred to clients in a trusted manner and the behavioral pattern of computational resources is optimized most efficiently up to the resource requirements. Initial results indicate a promising performance of trusted scaling with the trust metrics to ensure the trustworthiness of the overall system. Furthermore, ASICs and OBCs could be developed to extract new electrochemical features and identified in the case of using nanoscale sensors based on nanomaterials like graphene and quantum dots, which is the future direction of our study besides from memory speed trusted and scalable analytic models of the massive datum.

Acknowledgment Thanks to TOBB University of Economics and Technology Distributed Data Analytics Research Laboratory and IBM for providing test clusters and research infrastructure. Thanks to industrial collaborators Dr. Bruno Michel and Dr. Atakan Peker for discussions about

the 3D-Stack Packaging and ASIC design technologies. Thanks to *YONGATEK Embedded Systems* (https://yongatek.com/) *and SilTerra* (https://www.silterra.com/index.php#homepage) *Incorporations* for partial funds for the research studies and proof of concept (POC) implementation supports.

References

1. Y.X. Ho, Q. Chen, H. Nian, K.B. Johnson, An assessment of pharmacists' readiness for paperless labeling: A national survey. J. Am. Med. Inform. Assoc. **21**(1), 43–48 (2014)
2. S. Jahani, S.K. Setarehdan, Centromere and length detection in artificially straightened highly curved human chromosomes. Int. J. Biomed. Eng. **2**(5), 56–61 (2012)
3. X. Wang, B. Zheng, M. Wood, S. Li, W. Chen, H. Liu, Development and evaluation of automated systems for detection and classification of banded chromosomes: Current status and future perspective. J. Phys. D. Appl. Phys. **38**, 2536–2542 (2005)
4. J.M. Cho, Chromosome classification using back propagation neural networks. IEEE Eng. Med. Biol. **19**, 28–33 (2000)
5. B. Lerner, Towards a completely automatic neural-network-based and human chromosome analysis. IEEE. Trans.. Syst. **28**(4), 544–552 (1998)
6. M. Moradi, S.K. Setarehdan, S.R. Ghaffari, Automatic landmark detection on chromosomes' images for feature extraction purposes, in *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, (2003), pp. 567–570
7. M.R. Mohammadi, Accurate localization of chromosome centromere based on concave points. J. Med. Signals Sens. **2**(2), 88–94 (2012)
8. E. Poletti, E. Grisan, A. Rugger, A modular framework for the automatic classification of chromosomes in Q-band images. Elsevier Comput. Methods Prog. Biomed. **105**, 120–130 (2012)
9. N. Madian, K.B. Jayanthi, Analysis of human chromosome classification using centromere position. Measurement **47**, 287–295 (2014)
10. Y. Chen et al., Active detection via adaptive submodularity, in *ICML*, (2014)
11. B. Mirzasoleiman, A. Karbasi, R. Sarkar, A. Krause, Distributed Submodular Maximization: Identifying Representative Elements in Massive Data. In NIPS. (2013, December). pp. 2049-2057
12. S. Rungruangbaiyok, P. Phukpattaranont, Chromosome image classification using a two-step probabilistic neural network. J. Sci. Technol. **32**(3), 255–262 (2010)
13. A.S. Arachchige, J. Samarabandu, J.H.M. Knoll, P.K. Rogan, Intensity integrated Laplacian-based thickness measurement for detecting human metaphase chromosome centromere location. I.E.E.E. Trans. Biomed. Eng. **60**(7), 2005–2013 (2013)
14. H. Choi, K.R. Castlman, A.C. Bovik, Segmentation and fuzzy logic classification ofM-FISH chromosomes images, in *Proceedings of Image Processing, 2006 IEEE International Conference*, (2006), pp. 69–72
15. BioImlab [Available on line 23 Sept 2019] http://bioimlab.dei.unipd.it.
16. HL7 [Available on line 23 Sept 2019] https://site.hl7.org.au/standards/international-published/ . http://www.hl7.com.au/HL7-Tools.htm.
17. D. Wetteroth, *OSI Reference Model for Telecommunications*, vol 396 (McGraw-Hill, New York, 2002)
18. F. Abid, L. Hamami, A survey of neural network based automated systems for human chromosome classification. Artif. Intell. Rev. **49**(1), 41–56 (2018)
19. S. Prakash, N.K. Chaudhury, Dicentric chromosome image classification using Fourier domain based shape descriptors and support vector machine, in *Proceedings of International Conference on Computer Vision and Image Processing*, (Springer, Singapore, 2017)
20. S.S. Dhillon et al., The 2017 terahertz science and technology roadmap. J. Phys. D. Appl. Phys. **50**(4), 043001 (2017)

21. L.G. Shaffer, J.R. Lupski, Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annu. Rev. Genet. **34**, 297–329 (2000)
22. A. Theisen, L.G. Shaffer, Disorders caused by chromosome abnormalities. Appl. Clin. Genet. **3**, 159–174 (2010). https://doi.org/10.2147/TACG.S8884
23. R.D. Schmickel, Contiguous gene syndromes: A component of recognizable syndromes. J. Pediatr. **109**(2), 231–241 (1986)
24. P.C. Nowell, D.A. Hungerford, A minute chromosome in human chronic granulocytic leukemia. Science **142**, 1497 (1960)
25. X. Wang, B. Zheng, S. Li, J.J. Mulvihill, X. Chen, H. Liu, Automated identification of abnormal metaphase chromosome cells for the detection of chronic myeloid leukemia using microscopic images. J. Biomed. Opt. **15**(4), 046026 (2010)
26. M.A. Agca, C. Tastan, et al., Biological system design tool, application paper, in *iGEM*, (2011)
27. M.A. Ağca, E. Baceskı, S. Gökçebağ, MEMCA for satellite and space data: MEMCA [Memory centric analytics] for satellite and space data, in *Recent Advances in Space Technologies (RAST), 2015 7th International Conference on*, (IEEE, 2015)
28. J.F. Ojeda et al., Chromosomal analysis and identification based on optical tweezers and Raman spectroscopy. Opt. Express **14**(12), 5385–5393 (2006)
29. Y. Sung et al., Stain-free quantification of chromosomes in live cells using regularized tomographic phase microscopy. PloS One **7**(11), e49502 (2012)
30. P.D. Lewis et al., Evaluation of FTIR spectroscopy as a diagnostic tool for lung cancer using sputum. BMC Cancer **10**(1), 640 (2010)
31. D.J. Lyman, J. Murray-Wijelath, Fourier transform infrared attenuated total reflection analysis of human hair: Comparison of hair from breast cancer patients with hair from healthy subjects. Appl. Spectrosc. **59**(1), 26–32 (2005)
32. P.C. Chen, S.C. Mwakwari, A.K. Oyelere, Gold nanoparticles: from nanomedicine to nanosensing. Nanotechnol. Sci. Appl. **1**, 45 (2008)
33. J.-i. Hahm, C.M. Lieber, Direct ultrasensitive electrical detection of DNA and DNA sequence variations using nanowire nanosensors. Nano Lett. **4**(1), 51–54 (2004)
34. L. Feng, L. Wu, Q. Xiaogang, New horizons for diagnostics and therapeutic applications of graphene and graphene oxide. Adv. Mater. **25**(2), 168–186 (2013)
35. J.A. Hertz, *Introduction to the Theory of Neural Computation* (CRC Press, Boca Raton, 2018)
36. W. Sun, S. Fei, L. Wang, Improving deep neural networks with multi-layer maxout networks and a novel initialization method. Neurocomputing **278**, 34–40 (2018)
37. M.A. AGCA, A holistic abstraction to ensure trusted scaling and memory speed trusted analytics, in *Developments in eSystems Engineering (DESE), 2018 Eleventh International Conference on*, (IEEE, Cambridge University, 2018)
38. B. Stuart, *Infrared Spectroscopy* (Wiley, New York, 2005); P. Jepsen, D. Cooke, M. Koch, Terahertz spectroscopy and imaging—Modern techniques and applications. Laser Photon. Rev. **5**(1), 124–166, (2011)
39. Y. Yin et al., 3D integrated circuit cooling with microfluidics. Micromachines **9**(6), 287 (2018)