(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 May 2020 (14.05.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/097302 A1

(51) International Patent Classification:
G06F 19/20 (2011.01) *G06T 7/30* (2017.01)
G06T 7/00 (2017.01) *G06T 7/32* (2017.01)
G06T 7/136 (2017.01) *G06T 7/33* (2017.01)
G06T 7/143 (2017.01) *G06T 7/35* (2017.01)
G06T 7/174 (2017.01)

(21) International Application Number:
PCT/US2019/060225

(22) International Filing Date:
07 November 2019 (07.11.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/756,958 07 November 2018 (07.11.2018) US
62/772,327 28 November 2018 (28.11.2018) US

(71) Applicant: **TRUSTEES OF TUFTS COLLEGE** [US/US]; Ballou Hall, Medford, MA 02155 (US).

(72) Inventors: **SOKOLOV, Igor**; 29 Winford Way, Medford, MA 02155 (US). **MILJKOVIC, Milos**; c/o Trustees of Tufts College, Ballou Hall, Medford, MA 02155 (US).

(74) Agent: **LICHAUCO, Faustino**; Occhiuti & Rohlicek LLP, 50 Congress Street, Suite 1000, Boston, MA 02109 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title:** ATOMIC-FORCE MICROSCOPY FOR IDENTIFICATION OF SURFACES

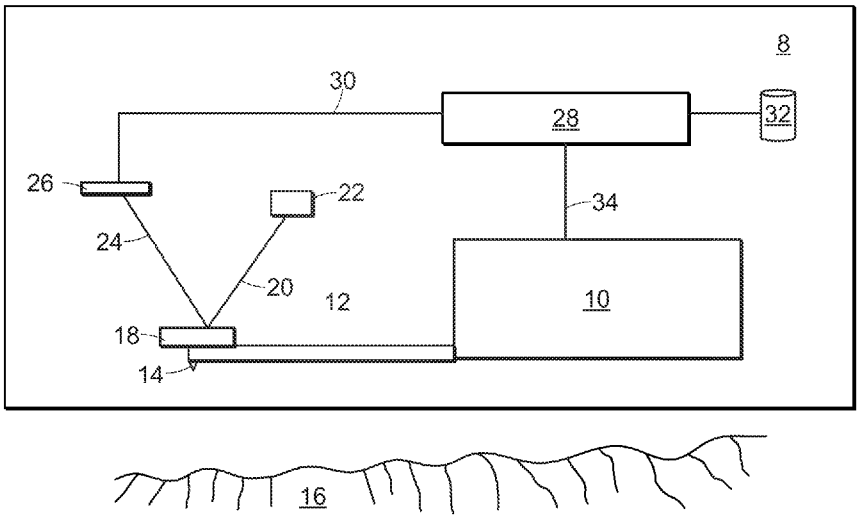


FIG. 1

(57) **Abstract:** A method comprises using an atomic-force microscope, acquiring a set of images associated with surfaces, and, using a machine-learning algorithm applied to the images, classifying the surfaces. As a particular example, the classification can be done in a way that relies on surface parameters derived from the images rather than using the images directly.

WO 2020/097302 A1

ATOMIC-FORCE MICROSCOPY FOR IDENTIFICATION OF SURFACES

RELATED APPLICATIONS

This application claims the benefit of the Nov. 7, 2018 priority date of U.S. Provisional Application 62/756,958 and the Nov. 28, 2018 priority date of U.S. Provisional Application 62/772,327, the contents of which are incorporated herein by reference.

FIELD OF INVENTION

The invention relates to the use of atomic force microscopy and machine learning in connection with using features of a surface to classify or identify that surface, and in particular, to using features to identify or classify biological cells.

BACKGROUND

In atomic force microscopy, a probe attached to the tip of a cantilever scans the surface of the sample. In one mode for operation, the probe taps the surface as it scans. As the probe scans the sample, it is possible to control the magnitude and direction of the force vector associated with a loading force that the probe exerts on the sample.

The deflection of the cantilever from its equilibrium position provides a signal from which a great deal of information can be extracted. As an example, by keeping either the loading force or the cantilever's deflection constant, it is possible to obtain the sample's topology at various points on the sample. The values collected at each point are then organized into an array in which the row and column identifies the location of a point in a two-dimensional coordinate system and the value at the row and column is representative of a property measured at that point. The resulting array of numbers can thus be viewed as a map. This makes it possible to make a map of the sample in which each point on the map indicates some property of the sample's surface at that point. In some examples, the property is the height of the surface above or below some reference plane.

However, an image of the surface's height is not the only image that can be recorded when scanning. The cantilever's deflection can be used to collect multiple images of the sample's surface, with each image being a map of a different property

of the surface. Examples of just a few of these properties include adhesion between the probe and the surface, the stiffness of the surface, and viscoelastic energy loss.

SUMMARY

The invention provides a method for identifying a surface using multidimensional images obtained by an atomic force microscope and for using information from those images for classifying a surface into one of several classes. According to the invention, it is possible to obtain a multi-dimensional image of a surface with two of the dimensions corresponding to spatial dimensions and additional dimensions corresponding to different physical and spatial properties that exist at the coordinate identified by the two spatial dimensions. In some embodiments, the dimensions are lateral dimensions.

A question that arises is how one chooses and uses these different physical and spatial properties for identification and classification of a surface. According to the invention, the properties that will be used for identification and classification of a surface are not pre-determined. They are calculated based on the result of machine learning applied to a database of images and their corresponding classes. They are learned. In particular, they are learned by machine learning.

Among the embodiments of the invention are those that include using an atomic force microscope to acquire different maps corresponding to different properties of the surface and using combinations of these maps, or parameters derived from those maps, to identify or classify a sample surface. Such a method comprises recording atomic force microscope images of examples of surfaces that belong to well-defined classes, forming a database in which such atomic force microscope maps are associated with the classes to which they belong, using the atomic force microscope maps thus obtained and the combinations thereof to learn how to classify surfaces by splitting the database into training and testing data with the training data being used to learn how to classify, for example by building a decision tree or neural network or a combination of thereof, and using the testing data to verify that the classification thus learned is effective enough to pass a given threshold of effectiveness.

Another embodiment includes reducing the maps provided by the atomic force microscope to a set of surface parameters, the values of which are defined by mathematical functions or algorithms that use those properties as inputs thereof. In a preferred practice, each map or image yields a surface parameter that can then be used as, together with other surface parameters to classify or identify the surface. In such embodiments, there exists a classifier that classifies based on these surface parameters. However, the classifier itself is not predetermined. It is learned though a machine-learning procedure as described above.

The method is agnostic to the nature of the surface. For example, one might use the method to classify surfaces of paintings or currency or secure documents such as birth certificates or passports in order to spot forgeries. But one might also use the same method to classify surfaces of cells or other portions of a living body in order to identify various disorders. For example, various cancers have cells that have particular surface signatures. Thus, the method can be used to detect various kinds of cancers.

A difficulty that arises is that of actually obtaining cells to examine. In some cases, an invasive procedure is required. However, there are certain kinds of cells that are naturally sloughed off the body or that can be extracted from the body with only minimal invasiveness. An example is that of gently scratching the cervix's surface in a Pap smear test. Among the cells that are naturally sloughed off are cells from the urinary tract, including the bladder. Thus, the method can be used to inspect these cells and detect bladder cancer without the need for an invasive and expensive procedure, such as cystoscopy.

The invention features using atomic force microscope that can produce a multidimensional array of physical properties, for example, when using sub-resonance tapping mode. In some practices, acquiring the set of images comprises using an atomic-force microscope in mode to carry out nanoscale-resolution scanning of the surfaces of cells that have been collected from bodily fluids and providing data obtained from the atomic force microscope scanning procedure to a machine learning system that provides an indication of the probability that the sample came from a

patient who has cancer, hereafter referred to as a "cancer-afflicted patient." The method is applicable in general to classifying cells based on their surface properties.

Although described in the context of bladder cancer, the methods and systems disclosed herein are applicable for detection of other cancers in which cells or body fluid are available for analysis without the need for invasive biopsy. Examples include cancer of the upper urinary tract, urethra, colorectal and other gastrointestinal cancers, cervical cancers, aerodigestive cancers, and other cancers with similar properties.

Moreover, the methods described herein are applicable to detection of cellular abnormalities other than cancer as well as to monitoring cellular reaction to various drugs. In addition, the methods described herein are useful for classifying and identifying surfaces of any type, whether derived from a living creature or from non-living matter. All that is necessary is that the surface be one that is susceptible to being scanned by an atomic force microscope.

For example, the method described herein can be used to detect forgeries, including forgeries of currency, stock certificates, identification papers, or artwork, such as paintings.

In one aspect, the invention features using an atomic-force microscope to acquire a set of images of each of a plurality of cells obtained from a patient, processing the images to obtain surface parameter maps, and, using a machine-learning algorithm applied to the images, classifying the cells as having originated in either a cancer-afflicted or cancer-free patient.

Among these embodiments are those in which the microscope is used in sub-resonance tapping mode. In yet other embodiments, the microscope is used in ringing mode.

In another aspect, the invention features: using an atomic-force microscope, acquiring a set of images associated with surfaces, processing the images to obtain surface parameter maps, and, using a machine-learning algorithm applied to the images, classifying the surfaces.

Among these practices are those that include selecting the surfaces to be surfaces of bladder cells and classifying the surfaces as those of cells that originated from a cancer-afflicted or cancer-free patient.

In another aspect, the invention features a method comprising, using an atomic-force microscope to acquire a set of images associated with surfaces, combining the images, and, using a machine-learning method applied to the combined images, classifying the surfaces.

This method cannot be carried out in the human mind with or without pencil and paper because it requires an atomic force microscope to be carried out and because the human mind cannot carry out a machine-learning method since the human mind is not a machine. The method is also carried out in a non-abstract manner so as to achieve a technical effect, namely the classification of surfaces based on technical properties thereof. A description of how to carry out the method in an abstract and/or non-technical manner has been purposefully omitted to avoid misconstruing the claim as covering anything but a non-abstract and technical implementation.

In some practices, the images are images of cells. Among these are practices that further include automatically detecting that an image of a cell has an artifact and excluding that image from being used for classifying the surfaces and practices that include partitioning an image of a simple into partitions, obtaining surface parameters for each partition, and defining a surface parameter of the cell as being the median of the surface parameters for each partition.

Some practices also include processing the images to obtain surface parameters and using machine learning to classify the surfaces based at least in part on the surface parameters. Among these are practices that further include defining a subset of the surface parameters and generating a database based on the subset. In such practices, defining the subset of surface parameters includes determining a correlation between the surface parameters, comparing the correlation with a threshold to identify a set of correlated parameters, and including a subset of the set of correlated parameters in the subset of surface parameters. Also among these are

practices that further include defining a subset of the surface parameters and generating a database based on the subset. In these practices, defining the subset of surface parameters includes determining a correlation matrix between the surface parameters and wherein determining the correlation matrix includes generating simulated surfaces. Also among these practices are those that include defining a subset of the surface parameters and generating a database based on the subset. In these practices, defining the subset of surface parameters includes combining different surface parameters of the same kind from the same sample.

Practices also include those in which acquiring the set of images includes using a multi-channel atomic-force microscope in ringing mode, wherein each channel of the atomic-force microscope provides information indicative of a corresponding surface property of the surfaces.

Also among the practices of the invention are those that include selecting the surfaces to be surfaces of cells collected from urine of a subject and classifying the cells as indicative of cancer or not indicative of cancer.

A variety of ways of using the microscope are within available without departing from the scope of the invention. These include using a multi-channel atomic force microscope, wherein each channel corresponds to a surface property of the surface, using the atomic-force microscope in sub-resonant tapping mode, and using the atomic force microscope in connection with acquiring multiple channels of information, each of which corresponds to a different surface property of the surface, condensing information provided by the channels and constructing, from that condensed information, a condensed database.

Among the practices of the invention that rely on a multi-channel atomic force microscope are those that further include forming a first database based on the information provided by the channels and carrying the construction of a condensed database in any of a variety of ways. Among these are projecting the first database into a subspace of dimensionality lower than that of the first database, the projection defining the condensed database, the condensed database having a dimensionality that

is less than that of the first database. Also among these are those that include a condensed database from the first database, the condensed database having fewer indices than the first database. This can be carried out, for example, by carrying out tensor addition to generate tensor sums that combine information from the first database along one or more slices corresponding to one or more indices of the first database and forming the condensed database using the tensor sums.

In some practices of the invention, deriving a condensed database from the first database includes defining a subset of values from the first database, each of the values being representative of a corresponding element in the first database, deriving a condensed value from the values in the subset of values, and representing the corresponding elements from the first database with the condensed value, wherein deriving the condensed value includes summing the values in the subset of values. The summation can be carried out in a variety of ways, including by carrying out tensor addition to generate tensor sums that combine values from the first database along one or more slices corresponding to corresponding indices of the first database and forming a condensed database using the tensor sums.

Practices of the invention also include those in which the condensed database is derived from the first database by defining a subset of values from the first database, each of the values being representative of a corresponding element in the first database, deriving a condensed value from the values in the subset of values, and representing the corresponding elements from the first database with the condensed value, wherein deriving the condensed value includes averaging the values in the subset of values, for example by obtaining an arithmetic average or a geometric average.

Also among the practices of the invention are those in which deriving a condensed database from a first database includes defining a subset of values from the first database, each of the values being representative of a corresponding element in the first database, deriving a condensed value from the values in the subset of values, and representing the corresponding elements from the first database with the

condensed value, wherein the condensed value is one of a maximum or a minimum of the values in the subset of values.

In yet other embodiments, deriving a condensed database from the first database includes defining a subset of values from the first database, each of the values being representative of a corresponding element in the first database, deriving a condensed value from the values in the subset of values, and representing the corresponding elements from the first database with the condensed value, wherein deriving the condensed value includes passing information from the first database through a surface-parameter extractor to obtain a surface-parameter set. Among these are practices that include normalizing the surface parameters representative of the surface-parameter set to be independent of surface areas of images from which they were derived and practices that include dividing the surface parameter by another parameter of the same dimension.

Other practices include automatically detecting that an image of a sample has an artifact and automatically excluding the image from being used for classifying the surfaces.

Still other practices include partitioning an image of a sample into partitions, obtaining surface parameters for each partition, and defining a surface parameter of the cell as being the median of the surface parameters for each partition.

Some practices of the invention include g processing the images to obtain surface parameters and using machine learning to classify the surfaces based at least in part on the surface parameters and from externally-derived parameters. Among these are practices in which the surfaces are surfaces of bodies that have been derived from collected samples, at least one of the samples being a body-free sample, which means that it has no bodies. In these practices, the method further includes selecting the externally-derived parameters to include data indicative of an absence of bodies from the body-free sample. Among the practices that include a body-free sample are those that include assigning an artificial surface parameter to the body-free sample. In some practices, the surfaces are surfaces of cells derived from samples obtained from

a patient. Among these are practices that include selecting the externally-derived parameters to include data indicative of a probability that the patient has a particular disease. Examples of such data indicative of the probability includes the patient's age, the patient's smoking habits, and the patient's family history.

A variety of machine-learning methods can be used. These include the Random Forest Method, the Extremely Randomized Forest Method, the method of Gradient Boosting Trees, using a neural network, a method of decision trees, and combinations thereof.

In some embodiments, the surfaces are surfaces of a first plurality of cells from a patient, a second plurality of the cells has been classified as having come from a cancer-afflicted patient, and a third plurality of the cells has been classified as having come from a cancer-free patient. These methods include diagnosing the patient with cancer if a ratio of the second plurality to the first plurality exceeds a predetermined threshold.

In some practices, the atomic-force microscope includes a cantilever and a probe disposed at a distal tip of the cantilever. The cantilever has a resonant frequency. In these practices, using the using the atomic-force microscope includes causing a distance between the probe and the surface to oscillate at a frequency that is less than the resonant frequency.

In some practices, using the atomic-force microscope includes using a microscope that has been configured to output multiple channels of information corresponding to different physical properties of the sample surface.

Other practices include processing the images to obtain surface parameters and using machine learning to classify the surfaces based at least in part on the surface parameters and from externally-derived parameters. In these embodiments, the surfaces are surfaces of cells derived from samples obtained from a patient, at least one of the samples being a cell-free sample that has no cells from the patient. In such practices, the method further includes selecting the externally-derived parameters to include data indicative of an absence of cells from the cell-free sample. Among these

practices are those that further include assigning an artificial surface parameter to the cell-free sample.

In another aspect, the invention features an apparatus comprising an atomic force microscope and a processing system. The atomic force microscope acquires images associated with surfaces. The processing system receives signals from the atomic force microscope representative of the images and combines the images. The processing system includes a machine-learning module and a classifier that classifies an unknown sample after having learned a basis for classification from the machine-learning module.

In some embodiments, the processing system is configured to process the images to obtain surface parameters and to use the machine-learning module to classify the surfaces based at least in part on the surface parameters. Among these are embodiments in which the atomic-force microscope comprises a multi-channel atomic force microscope, each channel of which corresponds to a surface property of the surfaces. Among these are embodiments that also include a condenser that condenses information provided by the channels and constructs, from the condensed information, a condensed database.

Embodiments that include a condensed database also include those in which a classifier classifies an unknown sample based on the condensed database.

A variety of condensers are available for constructing a condensed database. Among these are condensers that construct the condensed database by projecting the first database into a subspace of dimensionality lower than that of the first database. This projection defines a condensed database that has a dimensionality that is less than that of the first database.

As used herein, "atomic force microscopy," "AFM," "scanning probe microscopy," and "SPM" are to be regarded as synonymous.

The only methods described in this specification are non-abstract methods. Thus, the claims can only be directed to non-abstract implementations. As used

herein, "non-abstract" is a deemed to mean compliant with the requirements of 35 USC 101 as of the filing of this application.

These and other features of the invention will be apparent from the following detailed description and the accompanying figures, in which:

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 shows a simplified diagram of one example of an atomic force microscope;

FIG. 2 shows additional details from the processing system of FIG. 1;

FIG. 3 shows a diagnostic method carried out by the atomic force microscope and the processing system shown in FIGS. 1 and 2;

FIG. 4 shows the view through an optical microscope built into the atomic force microscope shown in FIG. 1;

FIG. 5 shows maps of bladder cells acquired by the atomic force microscope of FIG. 1;

FIG. 6 shows details of interactions between the database and the machine-learning module in the processing system of FIG. 2;

FIG. 7 shows details of condensing the initial large database into a condensed database of smaller dimension and shows the details of interactions between the condensed database and the machine-learning module in the processing system of FIG. 2;

FIG. 8 shows examples of simulated surfaces used in connection with evaluating correlation between different surface parameters;

FIG. 9 shows a histogram plot of an importance coefficient for two surface parameters;

FIG. 10 shows a binary tree;

FIG. 11 shows a machine-learning method adapted to the data structure needed for classification;

FIG. 12 shows a representative example of the artifacts because of possible contamination of the cell surface.

FIG. 13 shows the dependences of the number of surface parameters on the correlation threshold;

FIG. 14 shows the hierarchy of importance of the surface parameters for height and adhesion properties calculated within the Random Forest method;

FIG. 15 shows accuracy for different numbers of surface parameters and different allocations of data among the training and testing database as calculated using the Random Forest method for combined channels of height and adhesion;

FIG. 16 shows receiver operating characteristics using the Random Forest method for the combined channels of height and adhesion;

FIG. 17 shows a plot similar to that shown in FIG. 16 but with artificial data used to confirm the reliability of the procedure used to generate the data in FIG. 16;

FIG. 18 shows area under the receiver operating characteristics of FIG. 17;

FIG. 19 shows accuracy for different numbers of surface parameters and different ways of allocating data between the training data and testing data using the Random Forest method for the combined channels of height and adhesion when using with five cells per patient and two cells required to be identified as having come from a cancer-afflicted patient (N=5, M=2);

FIG. 20 shows receiver operating characteristics calculated using the Random Forest method for the combined channels of height and adhesion when

using with five cells per patient and with two cells required to be identified as having come from a cancer-afflicted patient (N=5, M=2); and

FIG. 21 is a table showing statistics of the confusion matrix associated with cancer diagnosis for two separate channels, one of which is for height and the other of which is for adhesion.

DETAILED DESCRIPTION

FIG. 1 shows an atomic force microscope **8** having a scanner **10** that supports a cantilever **12** to which is attached a probe **14**. The probe **14** is thus cantilevered from the scanner **10**. The scanner **10** moves the probe **14** along a scanning direction that is parallel to a reference plane of a sample's surface **16.** In doing so the scanner **10** scans a region of a sample's surface **16**. While the scanner is moving the probe **14** in the scanning direction, it is also moving it in a vertical direction perpendicular to the reference plane of the sample surface **16.** This causes the distance from the probe **14** to the surface **16** to vary.

The probe **14** is generally coupled to a reflective portion of the cantilever **12.** This reflective portion reflects an illumination beam **20** provided by a laser **22**. This reflective portion of the cantilevered **12** will be referred to herein as a mirror **18**. A reflected beam **24** travels from the mirror **18** to a photodetector **26**, the output of which connects to a processor **28**. In some embodiments, the processor **28** comprises FPGA electronics to permit real time calculation of surface parameters based on physical or geometric properties of the surface.

The movement of the probe **14** translates into movement of the mirror **18**, which then results in different parts of the photodetector **26** being illuminated by the reflected beam **24**. This results in a probe signal **30** indicative of probe movement. The processor **28** calculates certain surface parameters based on the probe signal **30** using methods described below and outputs the results **33** to a storage medium **32**. These results **33** include data representative of any of the surface parameters described herein.

The scanner **10** connects to the processor **28** and provides to it a scanner signal **34** indicative of scanner position. This scanner signal **34** is also available for use in calculating surface parameters.

FIG. 2 shows the processing system **28** in detail. The processing system **28** features a power supply **58** having an AC source **60** connected to an inverter **62**. The power supply **58** provides power for operating the various components described below. The processing system further includes a heat radiator **64**.

In a preferred embodiment, the processing system **28** further includes a user interface **66** to enable a person to control its operation.

The processing system **28** further includes first and second A/D converters **68**, **70** for receiving the probe signal and the scanner signals and placing them on a bus **72**. A program storage section **74**, a working memory **76**, and CPU registers **78** are also connected to the bus **72**. A CPU **80** for executing instructions **75** from program storage **74** connects to both the registers **78** and an ALU **82**. A non-transitory computer-readable medium stores these instructions **75**. When executed, the instructions **75** cause the processing system **28** to calculate any of the foregoing parameters based on inputs received through the first and second A/D converters **68**, **70**.

The processing system **28** further includes a machine-learning module **84** and a database **86** that includes training data **87** and testing data **89**, best seen in FIG. 6. The machine-learning module **84** uses the training data **87** and the testing data **89** for implementing the method described herein.

A specific example of the processing system **28** may include FPGA electronics that includes circuitry configured for determining the values of the properties of the imaging services and/or the surface parameters described above.

FIG. 3 shows a process that uses an atomic force microscope **8** to acquire images and to provide them to the machine-learning module **84** to characterize the sample using the images. The process shown in FIG. 3 includes acquiring urine **88**

from a patient and preparing cells **90** that have been sloughed off into the urine **88**. After having scanned them, the atomic force microscope **8** provides images of the bladder cells **90** for storage in the database **86**.

Each image is an array in which each element of the array represents a property of the surface **16**. A location in the array corresponds to a spatial location on the sample's surface **16**. Thus, the image defines a map corresponding to that property. Such a map shows the values of that property at different locations on the sample's surface **16** in much the same way a soil map shows different soil properties at different locations on the Earth's surface. Such a property will be referred to as a "mapped property."

In some cases, the mapped properties are physical properties. In other cases, the properties are geometrical properties. An example of a geometrical property is the height of the surface **16**. Examples of physical properties include the surface's adhesion, its stiffness, and energy losses associated with contacting the surface **16**.

A multi-channel atomic force microscope **8** has the ability to map different properties at the same time. Each mapped property corresponds to a different "channel" of the microscope **8**. An image can therefore be regarded as a multidimensional image array $M^{(k)}$, where the channel index, k, is an integer in the interval $[1,K]$, where K is the number of channels.

When used in a sub-resonance tapping mode, a multi-channel atomic force microscope **8** can map the following properties: height, adhesion, deformation, stiffness, viscoelastic losses, feedback error. This results in six channels, each of which corresponds to one of six mapped properties. When used in ringing mode, the atomic force microscope **8** can map, as an example, one or more of the following additional properties in addition to the previous six properties: restored adhesion, adhesion height, disconnection height, pull-off neck height, disconnection distance, disconnection energy loss, dynamic creep phase shift, and zero-force height. This results in a total of fourteen channels in this example, each of which corresponds to one of fourteen mapped properties.

The scanner **10** defines discrete pixels on the reference plane. At each pixel, the microscope's probe **14** makes a measurement. For convenience, the pixels on the plane can be defined by Cartesian coordinates (x_i, y_j). The value of the k^{th} channel measured at that pixel is $z_{i,j}^{(k)}$. With this in mind, an image array that represents a map or image of the k^{th} channel can be formally represented as:

$$M^{(k)}=\{x_i, y_j, z_{i,j}^{(k)}\} \qquad (1)$$

where "*i*" and "*j*" are integers in the intervals [1, *Ni*]and [1, *Nj*] respectively and where *Ni* and *Nj* are the numbers of pixels available for recording an image in the *x* and *y* directions respectively. The values of *Ni* and *Nj* can be different. However, the methods described herein do not depend significantly on such a difference. Hence, for purposes of discussion, *Ni*=*Nj*=*N*.

The number of elements in a sample's image array would be the product of the number of channels and the number of pixels. For a relatively homogeneous surface **16**, it is only necessary to scan one region of the surface **16**. However, for a more heterogenous surface **16**, it is preferable to scan more than one region on the surface **16**. By way of analogy, if one wishes to inspect the surface of the water in a harbor, it is most likely only necessary to scan one region because other regions would likely be similar anyway. On the other hand, if one wishes to inspect the surface of the city that the harbor serves, it would be prudent to scan multiple regions.

With this in mind, the array acquires another index to identify the particular region that is being scanned. This increases the array's dimensionality. A formal representation of the image array is thus:

$$M^{(k,s)} = \{x_i^{(s)}, y_j^{(s)}, z_{i,j}^{(k,s)}\} \qquad (2)$$

where the scanned-region index *s* is an integer in the interval [1, *S*] that identifies a particular scanned region within a sample. Note that this causes the number of elements in the image array for a particular sample to grow by a factor equal to the number of scanned regions.

Preferably, the number of such scanned regions is large enough to be represent the sample as a whole. One way to converge on an appropriate number of scanned regions is to compare the distribution of deviations between two such scanned regions. If incrementing the number of scanned regions does not change this in a statistically significant way, then the number of scanned regions is likely to be adequate to represent the surface as a whole. Another way is to divide what is considered to be a reasonable testing time by the amount of time required to scan each scanned region and to use that quotient as the number of areas.

In some cases, it is useful to split each of the scanned regions into partitions. For the case in which there are P such partitions in each scanned region, the array can be defined as:

$$M^{(k,s,p)} = \{x_i^{(s,p)}, y_j^{(s,p)}, z_{i,j}^{(k,s,p)}\} \quad (2a)$$

where the partition-index p is an integer in the interval $[1,P]$. In the case of a square scanned area, it is convenient to divide the square into four square partitions, thus setting P to be equal to four.

The ability to divide a scanned region into partitions provides a useful way to exclude image artifacts. This is particularly important for inspection of biological cells **90**. This is because the process of preparing cells **90** for inspection can easily introduce artifacts. These artifacts should be excluded from any analysis. This makes it possible to compare one partition against the others to identify which, if any, deviate significantly enough to be excluded.

On the other hand, the addition of a new index further increases the dimensionality of the array.

To identify a class to which a sample belongs based on the image arrays $M^{(k,s)}$ acquired by the atomic force microscope **8**, the machine-learning module **84** relies in part on building a suitable database **86** that includes images of surfaces that are known a priori to belong to particular classes $C^{(q)}$. Such a database **86** can be formally represented by:

$$D_n^{(l,k,s,p)} = \{M_n^{(k,s,p)}, C^{(l)}\} \quad (2b)$$

where k is a channel index that represents a property or channel, s is a scanned-region index that identifies a particular scanned region, p is a partition index that represents a particular partition of the s^{th} scanned region, n is a sample index that identifies a particular sample, and l is a class index that identifies a particular class from a set of L classes. The overall size of the array is thus the product of the number of classes, the number of samples, the number of scanned regions, the number of partitions per scanned region, and the number of channels.

FIG. 3 shows a diagnostic method **10** that features using an atomic force microscope **8** operated using sub-resonance tapping and the machine-learning module **84** to inspect surfaces of biological cells **90** that have been recovered from urine **88** in an effort to classify patients into one of two classes: cancer-afflicted and cancer-free. Since there are two classes, L=2.

A preferred practice includes collecting the cells **90** using centrifugation, gravitational precipitation, or filtration followed by fixing, and freeze drying or subcritical drying the cells **90**.

In the example shown, the atomic force microscope **8** was operated using both sub-resonant tapping modes, such as PeakForce QMN as implemented by Bruker, Inc., and ringing modes, for example as implemented by NanoScience Solutions, LLC. Both modes allow to record height and adhesion channels. Ringing mode is, however, a substantially faster mode of image collection. As noted above, these modes allow many channels to record simultaneously. However, only two channels are used in the experiment described herein.

FIG. 4 shows the atomic force microscope's cantilever **12** together with a cell **90** obtained from a patient and prepared as described above. The view is taken through an optical microscope that is coupled to the atomic force microscope **8**.

FIG. 5 show first and second map pairs **92, 94**. The first map pair **92** shows maps of a cell **90** from a cancer-free patient. The second map pair **94** shows maps

of a cell **90** from a cancer-afflicted patient. The maps shown are those of a square scanned area that is ten micrometers on a side with a resolution of 512 pixels in both dimensions. The scan speed was 0.1 Hz when scanning in a sub-resonant tapping mode, such as PeakForce QMN mode, and 0.4 Hz when scanning in ringing mode. The peak force during scanning is five nano-newtons.

Referring now to FIG. 6, the machine-learning module **84** trains a candidate classifier **100** based on the database **86**. A particular machine learning method can be chosen from the family of machine learning methods, for example, decision trees, neural networks, or combinations thereof.

The methods shown in FIG. 6 and FIG. 7 begin by splitting the database **86** into training data **87** and testing data **89**. This raises the question of how much of the data in the database **86** should go into the training data **87** and how much should go into the testing data **89**.

In some embodiments, 50% of the database **86** goes into the training data **87** and the remaining 50% goes into the testing data **89**. In other embodiments, 60% of the database **86** goes into the training data **87** and the remaining 40% goes into the testing data **89**. In yet other embodiments, 70% of the database **86** goes into the training data **87** and the remaining 30% goes into the testing data **89**. In still other embodiments, 80% of the database **86** goes into the training data **87** and the remaining 20% goes into the testing data **89**. The candidate classifier **100** should ultimately be independent of the ratio used in the split.

In the example illustrated in FIG. 3, ten bladder cells **90** were gathered for each patient. The presence of cancer was identified using standard clinical methods including invasive biopsies and histopathology. These methods are reliable enough for the two classes to be regarded as well defined. As a result, the database **86** shown in FIG. 6 can be represented as:

$$\begin{matrix} D_1^{(1,k\times p)} = \{M^{(k\times p)}{}_1, C^{(1)}\}, D_2^{(1,k\times p)} = \{M^{(k\times p)}{}_2, C^{(1)}\} \quad .. \quad D_{N\text{class}1}{}^{(1,k\times p)} = \{M^{(k\times p)}{}_{N\text{class}1}, C^{(1)}\} \\ D_1^{(2,k\times p)} = \{M^{(k\times p)}{}_1, C^{(2)}\}, D_2^{(2,k\times p)} = \{M^{(k\times p)}{}_2, C^{(2)}\} \quad .. \quad D_{N\text{class}2}{}^{(2,k\times p)} = \{M^{(k\times p)}{}_{N\text{class}2}, C^{(2)}\} \end{matrix}, \quad (3)$$

where N_{data1} is the number of patients that are in a first class, N_{data2} is the number of patients that are in a second class, and s, which is a whole number between one and ten inclusive, identifies the particular one of ten cells collected from a single patient. It is not necessary that N_{data1} and N_{data2} be equal.

When splitting the database **86** between the training data **87** and the testing data **89**, it is important to avoid having image arrays for different scanned areas from the same sample $\{M^{(k,1,p)}, M^{(k,2,p)}, M^{(k,S,p)}\}$ be divided between training and testing data **87, 89**. Violation of this rule would result in training and testing on the same sample. This would artificially pump up the classifier's effectiveness in a way that may not be reproducible when applying the classifier **100** to independent new samples.

The machine-learning module **84** uses the training data **87** to build the candidate classifier **100**. Depending on the type of classifier **100**, the training data **87** can be a learning tree, a decision tree, a bootstrap of trees, a neural network, or combinations thereof. The classifier **100**, which is represented below as "*AI*," outputs a probability that a particular sample n belongs to a particular class l:

$$\text{Prob}_n^{(k,s,p)(l)} = \text{AI}(M_n^{(k,s,p)}|C^{(l)}) \qquad (3a)$$

where $\text{Prob}_n^{(k,s,p)(l)}$ is the probability that the image or channel defined by $M_n^{(k,s,p)}$ belongs to class $C^{(l)}$.

After having been built, a verification module **102** uses the testing data **89** to verify that the candidate classifier **100** is, in fact, sufficiently effective. In the embodiment described herein, the verification module **102** evaluates effectiveness based at least in part on a receiver operating characteristic and on a confusion matrix. The robustness of the candidate classifier **100** was verified by repeating the random splitting of the database **86** to thereby generate different testing data **89** and training data **87** and then carrying out the classification procedure to see if this made in any difference.

If the candidate classifier **100** turns out to be insufficiently effective, the machine-learning module **84** changes the parameters of the training process and generates a new candidate classifier **100**. This cycle continues until the machine-learning module **84** eventually provides a candidate classifier **100** that attains a desired threshold of effectiveness.

The process of building a suitable classifier **100** is hindered to some extent by the computational load that arises when there is more than one probability value associated with a sample n. In fact, as a result of the multidimensional nature of the image array, for any one sample, there would be K-S-P probabilities, $\mathrm{Prob}_n^{(k=s=p)(i)}$ to process. The required computational load would be impractically high for such a large database.

Another bottleneck of dealing with such large arrays of data is the large number of samples used to provide a reasonable training of the classifiers. When building decision trees, a rule of thumb requires the number of samples to be at least six times larger than the dimension of the database. Because atomic force microscopy is a relatively slow technique, it would be impractical to obtain enough samples to build any reasonable classifier.

A condenser 104, as shown in FIG. 7, addresses the foregoing difficulty. The condenser **104** condenses information provided by a particular channel into a space of surface parameters that embodies information about that channel. The condenser **104** receives the database **86** and generates a condensed database **106**. In effect, this amounts to projecting a multidimensional matrix that is in a fairly high-dimensional space into a matrix of much less dimensionality.

The condenser **104** carries out any of a variety of database-reduction procedures. Among these are procedures that combine one or more of the database-reduction procedures described herein. These have in common deriving, from a set of data, a surface parameter that embodies at least some of the information embodied in that set.

In some practices, the condenser **104** carries out a first database-reduction procedure. This first database-reduction procedure relies on the observation that each image is ultimately an array that can be combined with other such arrays in a way that yields an object that preserves enough aspects of the information from the arrays that went into it so as to be useful in classifying a sample. For example, tensor addition "⊕" can be used to combine a set of images $M_n^{(k:s:p)}$ along a slice corresponding to one of its indices.

In one specific implementation, the slice corresponds to the index k. In that case, the tensor sum of the images is given by:

$$M_n^{(1:s:p)} \oplus M_n^{(2:s:p)} \oplus M_n^{(3:s:p)} \oplus \ldots M_n^{(K:s:p)}$$

Thus, each element of the condensed database **106** to be used for machine learning becomes the following:

$$D_n^{(1:s:p)} = \{M_n^{(1:s:p)} \oplus M_n^{(2:s:p)} \oplus M_n^{(3:s:p)} \oplus \ldots M_n^{(K:s:p)}\} \qquad (3\text{-}1)$$

This particular example decreases the dimensionality of the database **86** by a factor of K. Therefore, the classifier **100** defines the probability as follows:

$$\text{Prob}_n^{(s:p)(l)} = \text{AI}(M_n^{(1:s:p)} \oplus M_n^{(2:s:p)} \oplus M_n^{(3:s:p)} \oplus \ldots M_n^{(K:s:p)} | C^{(l)})$$

It is also possible to carry out a similar procedure for the remaining indices. Ultimately,

$$\text{Prob}_n^{(l)} = \text{AI}(\oplus\oplus\oplus M_n^{(k:s:p)} | C^{(l)})$$

where "⊕⊕⊕" represents a tensor summation over the indices k,s,p.

In other practices, the condenser **104** instead carries out a second database-reduction procedure. This second database-reduction procedure relies on geometrical or algebraic averaging on each of the indexes k,s,p separately or their combination. Examples of particular ways to carry out the second procedure include the following averaging procedures over all indices k,s,p:

$$\mathrm{Prob}^{(I)}{}_n = \frac{1}{K \times S \times P} \sum_{k,s,p} \mathrm{Prob}^{(k,s,p)(I)}{}_n \quad , \qquad (3\text{-}2)$$

$$\mathrm{Prob}^{(I)}{}_n = \frac{1}{\sqrt[3]{K \times S \times P}} \prod_{k,s,p} \mathrm{Prob}^{(k,s,p)(I)}{}_n \quad , \qquad (3\text{-}3)$$

$$\mathrm{Prob}^{(I)}{}_n = \frac{1}{K \times S \times P} \sum_{k,s,p} (1\text{-}\mathrm{Prob}^{(k,s,p)(I)}{}_n) \quad , \qquad (3\text{-}4)$$

$$\mathrm{Prob}^{(I)}{}_n = \frac{1}{\sqrt[3]{K \times S \times P}} \prod_{k,s,p} (1\text{-}\mathrm{Prob}^{(k,s,p)(I)}{}_n) \quad , \qquad (3\text{-}5)$$

In yet other practices, the condenser **104** instead carries out a third database-reduction procedure. This third database-reduction procedure relies on assigning the highest or lowest probability of the entire series to a particular index. For example, considering scanned-region index s, one can use one of the following relationships:

$$\mathrm{Prob}^{(k,s,p)(I)}{}_n = \underset{s}{Max}\{\mathrm{Prob}^{(k,s,p)(I)}{}_n\} \quad , \qquad (3\text{-}6)$$

$$\mathrm{Prob}^{(k,s,p)(I)}{}_n = \underset{s}{Min}\{\mathrm{Prob}^{(k,s,p)(I)}{}_n\} \quad . \qquad (3\text{-}7)$$

Ultimately, if all indexes are reduced this way

$$\mathrm{Prob}^{(I)}{}_n = \underset{k,s,p}{Max}\{\mathrm{Prob}^{(k,s,p)(I)}{}_n\} \quad \text{or} \qquad (3\text{-}8)$$

$$\mathrm{Prob}^{(I)}{}_n = \underset{k,s,p}{Min}\{\mathrm{Prob}^{(k,s,p)(I)}{}_n\} \quad . \qquad (3\text{-}9)$$

In some practices, the condenser **104** reduces the dimensionality of the database $D_n^{(I,s)}$ by passing each image through a surface-parameter extractor A_m to obtain a surface-parameter set, $P_{mn}^{(k,s)}$. This can be represented formally by:

$$P_{mn}^{(k,s)} = A_m\{M_n^{(k,s,p)}\} \quad (4)$$

where the surface-parameter index m is an integer in $[1,M]$, the channel index k identifies whether the map represents height, adhesion, stiffness, or some other

physical or geometric parameter, the sample index n identifies the sample, the scanned-region index s identifies the particular scanned region with in a sample, and the partition index p identifies the particular partition within a scanned region. This procedure provides a compact way to represent a multidimensional tensor $M_n^{(k,s,p)}$ as a surface-parameter vector $P_{nm}^{(k,s,p)}$.

The surface-parameter vector includes enough residual information concerning the channel from which it was derived to be usable as a basis for classification. However, it is much smaller than the image provided by the channel. As such, a classification procedure that relies on the surface-parameter vector sustains a much lower computational load but without a corresponding loss of accuracy.

A variety of surface parameters can be extracted from a channel. These include roughness average, root mean square, surface skew, surface kurtosis, peak-peak, ten-point height, maximum valley depth, maximum peak height, mean value, mean summit curvature, texture index, root mean square gradient, area root mean square slope, surface area ratio, projected area , surface area, surface bearing index, core fluid retention index, valley fluid retention index, reduced summit height, core roughness depth, reduced valley depth, l-h% height intervals of bearing curve, density of summits, texture direction, texture direction index, dominant radial wave length, radial wave index, mean half wavelength, fractal dimension, correlation length at 20%, correlation length at 37%, texture aspect ratio at 20%, and texture aspect ratio at 37%.

The list of surface parameters may be further extended by introducing the algorithms or mathematical formulas. For example, one can normalize the surface parameters to a surface area of the images, which can be different for different cells, by for example, dividing each parameter by a function of the surface area.

The example described herein relies on three surface parameters: valley fluid retention index ("*Svi*"), the Surfaces Area Ratio ("*Sdr*"), and the Surface Area, ("*S3A*").

The valley fluid retention index is a surface parameter that indicates the existence of large voids in a valley zone. It is defined by:

$$Svi = \frac{V(h_{0.80})}{(M-1)(N-1)\delta x \delta y} / Sq \ , \tag{5}$$

where N is the number of pixels in the x direction, M is the numbers of pixels in the y direction, $V(h_X)$, is a void area over the bearing area ratio curve and under the horizontal line h_X, and S_q is the Root Mean Square (RMS), which is defined by the following expression:

$$Sq = \sqrt{\frac{1}{MN}\sum_{k=0}^{N-1}\sum_{l=0}^{M-1}\left[h(x_k, y_l)\right]^2} \tag{6}$$

The surfaces area ratio ("*Sdr*") is a surface parameter that expresses the increment of the interfacial surface area relative to the area of the projected x, y plane. This surface parameter is defined by:

$$S_{sd} = \frac{\left(\sum_{k=0}^{M-2}\sum_{l=0}^{N-2} A_{kl}\right) - (M-1)(N-1)\delta x \delta y}{(M-1)(N-1)\delta x \delta y} 100\% , \tag{7}$$

where N is the number of pixels in the x direction and M is the numbers of pixels in the y direction.

The Surface Area, ("*S3A*") is defined by:

$$\text{S3A} = \underline{\left(\sum_{k=0}^{M-2}\sum_{l=0}^{N-2} A_{kl}\right) - (M-1)(N-1)\delta x \delta y} \ . \tag{8}$$

To calculate each of the above-mentioned three surface parameters from images provided by the atomic force microscope **8**, each image of a cell was first split

into four partitions, which in this case were quadrants of a square having five-micrometer sides. Thus, each cell yielded four sets of surface parameters, one for each quadrant.

The presence of artifacts in a cell can be addressed in any one of three different ways.

A first way is to have an operator inspect the cells for artifacts and exclude, from further processing, any cell that had one or more such artifacts. This requires human intervention to identify artifacts.

A second way is to provide an artifact-recognition module that is able to recognize an artifact and automatically exclude the cell that contains that artifact. This renders the procedure more operator-independent.

A third way is to use the median value of the parameters for each cell instead of the mean values. The results described herein were virtually unchanged when the median value was used instead of the mean value.

Using the same example of just two classes, the condensed database **106** will look as follows

$$\begin{matrix} D_1^{(1)(k,c,p)} = \{P^{(k,c,p)}{}_1, C^{(1)}\}, D_2^{(1)(k,c,p)} = \{P^{(k,c,p)}{}_2, C^{(1)}\} \quad .. \quad D_{Nbins1}^{(1)(k,c,p)} = \{P^{(k,c,p)}{}_{Nbins1}, C^{(1)}\} \\ D_1^{(2)(k,c,p)} = \{P^{(k,c,p)}{}_1, C^{(2)}\}, D_2^{(2)(k,c,p)} = \{P^{(k,c,p)}{}_2, C^{(2)}\} \quad .. \quad D_{Nbins2}^{(2)(k,c,p)} = \{P^{(k,c,p)}{}_{Nbins2}, C^{(2)}\} \end{matrix} \quad (9)$$

In other embodiments, one can assign additional parameters to help differentiate between different classes even though these parameters are not directly related to the atomic force microscope's images.

For example, when attempting to detect bladder cancer, it is quite possible that one or more samples of urine **88** will not have any cells **90**. A convenient way take into account such a result is to add a new "no cell" parameter that is either true or false. To avoid having to alter the data structure to accommodate such a parameter, a sample with a "no cell" set to "true" receives artificial values for surface parameters that are selected to avoid distorting the statistical results.

As another example, there are other factors that are not related to surface parameters but are nevertheless pertinent to classification. These include characteristics of patients, like age, smoking, and family history, all of which may be relevant to the probability of that patient having bladder cancer. These parameters can be included in a manner similar to the "no cell" parameter so as to avoid having to modify the data structure.

There exist yet other ways to use surface parameters to reduce the size of the database **86**.

One such procedure is that of excluding surface parameters that are sufficiently correlated with each other. Some surface parameters depend strongly on various other surface parameters. Hence, little additional information is provided by including surface parameters that are correlated with each other. These redundant surface parameters can be removed with little penalty.

One way to find the correlation matrix between surface parameters is to generate simulated surfaces, examples of which are shown in FIG. 8. Various sample surfaces imaged with an atomic force microscope **8** can also be used to identify correlation between different surface parameters.

The machine-learning module **84** is agnostic to the nature of its inputs. Thus, although it is shown as operating on an image array, it is perfectly capable of operating on the surface-parameter vector instead. The same machine-learning module **84** is therefore usable to determine the probability that a particular surface-parameter vector belongs to a particular class, i.e., to evaluate $\mathrm{Prob}_n^{(k_x,k_y,p)(l)} = \mathrm{AI}(P_n^{(k_x,k_y,p)}|C^{(l)})$.

Therefore, after having reduced the multidimensional image array $M_n^{(k_x,k_y,p)}$ into a surface-parameter vector $P_{nm}^{(k_x,k_y,p)}$, it becomes possible to substitute the surface-parameter vector $P_{nm}^{(k_x,k_y,p)}$ for the multidimensional image array $M_n^{(k_x,k_y,p)}$ and to then have the machine-learning module **84** learn what surface parameters are important for classification and how to use them to classify cells.

Because certain surface parameters are correlated with each other, it is possible to further reduce the dimensionality. This can be carried out without tensor summation. Instead, such reduction is carried out by direct manipulation of the same parameters from different images.

In addition to the methods that rely on the database-reduction procedures identified above as (3-1) to (3-9), it is also possible to use a classifier **100** that combines different surface parameters of the same kind from the same sample. Formally, this type of classifier **100** can be represented formally as:

$$\text{Prob}_n^{(j)} = \text{AI}(P_n|C^{(j)}) \quad (10)$$

where $P_n = F(P_{nm}^{(k,x,p)})$ and where $F(P_{nm}^{(k,x,p)})$ is a combination of different surface parameters identified by the surface-parameter index m and belonging to the sample identified by the sample index n.

A related classifier **100** is one that combines different surface parameters of the same kind m of the same sample n from the images of the same properties. Such a classifier **100** can be represented formally as:

$$\text{Prob}_n^{(k)(j)} = \text{AI}(P_{nm}^{(k)}|C^{(j)}) \qquad (11)$$

where $P_{nm}^{(k)} = F(P_{nm}^{(k,x,p)})$ and $F(P_{nm}^{(k,x,p)})$ is a combination of different surface parameters identified by the same surface-parameter index m of the sample identified by the sample index n and from the channel identified by the channel index k.

Yet another classifier **100** is one that does not combine all parameters but instead combines surface parameters by only one index. One such classifier **100** assigns one surface parameter to an entire series of partitions p within the same image. Such a classifier **100** is formally represented as:

$$\text{Prob}_n^{(k,x)(j)} = \text{AI}(P_{nm}^{(k,x)}|C^{(j)}) \quad (12)$$

where $P_{nm}^{(k,x)} = \text{F}(P_{nm}^{(k,x,p)})$ and $\text{F}(P_{nm}^{(k,x,p)})$ is a combination of surface parameters, examples of which include a parameter associated with a statistical distribution of $P_{nm}^{(k,x,p)}$ over the partition index. Examples include the average:

$$P_n^{(k;s)}{}_m = \frac{1}{N}\sum_{p=1}^{N} P_n^{(k;x;p)}{}_m \qquad (13)$$

and the median:

$$P_{nm}^{(k,s)} = \text{median}\{P_n^{(k;x;p)}\} \text{ for } p=1\ldots N \qquad (14)$$

When used in connection with detection bladder cancer imaging of multiple cells from each patient, the classifier **100** relies on either the average or the median. However, it is preferable for the classifier **100** to rely on the median rather than the average because the media is less sensitive to artifacts.

In the particular embodiment described herein, the machine-learning module **84** implements any of a variety of machine-learning methods. However, when confronted with multiple parameters, a machine-learning module **84** can easily become over-trained. It is thus useful to use three methods that are least prone to overtraining, namely the Random Forest method, the Extremely Randomized Forest method, and the method of Gradient Boosting Trees.

The Random Forest method and the Extremely Randomized Forest method are bootstrap unsupervised methods. The method of Gradient Boosting Trees is a supervised method of building trees. Variable ranking, classifier training, and validation were carried out using appropriate classifier functions from the SCIKIT-LEARN Python machine-learning package (version 0.17.1).

The Random Forest and Extremely Randomized Forest methods are based on growing many classification trees. Each classification tree predicts some classification. However, the votes of all trees define the final classification. The trees are grown on the training data **87**. In a typical database **86**, 70% of all data is in the training data **87** with the remainder being in the testing data **89.** In the experiments described herein, the split between training data **87** and testing data **89** was random and repeated multiple times to confirm that the classifiers **100** were insensitive to the manner in which the database **86** was split.

Each branching node relies on a randomly chosen subset of the original surface parameters. In the methods described herein, the number of elements in the chosen subset of original surface parameters is the square root of the number of surface parameters originally provided.

The learning process then proceeds by identifying the best split of the tree branches given the randomly chosen subset of surface parameters. The machine-learning module **84** bases the split threshold is based on an estimate of the classification error. Each parameter is assigned to a parameter region with respect to the most commonly occurring class of the training data **87**. In these practices, the machine-learning module **84** defines the classification error as a fraction of the training data **87** in that region that does not belong to the most common class:

$$E = 1 - \max_{k}(p_{mk}) \tag{15}$$

where p_{mk} represents the proportion of training data **87** that is both in the m^{th} region and that also belong to the k^{th} class. However, for a practical use, equation (1) is not sufficiently sensitive to avoid overgrowing the tree. As a result, the machine-learning module **84** relies on two other measures: the Gini index and cross-entropy.

The Gini index, which is a measure of variance across all K classes, is defined as follows:

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{16}$$

The Gini index remains small when all values of p_{mk} remain close to zero or unity. As a result, the Gini index measures an extent to which a particular node contains mostly samples from a single class. This is referred to as the extent of "node purity." Thus, to avoid overgrowing, each tree is grown only until the Gini-index results in complete separation of classes. This occurs when two descendant nodes yield a Gini-index that is less than that of the parent node. There is no pruning of the growing branches in these Random Forest methods.

The cross-entropy, which also provides a metric for node purity, is defined as:

$$D = -\sum_{k=1}^{K} p_{mk} \log(p_{mk}) \quad (17)$$

Like the Gini index, cross-entropy is small when all values of p_{mk} are close to zero. This is indicative of a pure node.

The Gini index also provides a way to obtain an "importance coefficient" that is indicative of the importance of each surface parameter. One such measure comes from adding all values of the decrease of the Gini index at the tree nodes for each of the variables and averaging over all the trees.

The histograms shown in FIG. 9 represent average values for importance coefficients with error bars to show the extent to which they deviate by one-standard-deviation from the mean. These importance coefficients correspond to the various surface parameters that can be derived from a particular channel. Thus, the histograms in the first row represent surface parameters that can be derived from the channel that measures the feature, "height," whereas the surface parameters in the second row represent surface parameters that can be derived from the channel that measures the feature, "adhesion." Note that a mnemonic device has been used to name the features, with all surface parameters that are derivable from the "height" channel beginning with "h" and all surface parameters that are derivable from the "adhesion" channel beginning with "a."

Thus, in the first row, the panel in the first column shows the importance coefficients for those surface parameters that are derived from the "height" channel when the machine-learning module **84** uses the Random Forest Method; the panel in the second column shows the importance coefficients for those surface parameters that are derived from the "height" channel when the machine-learning module **84** uses the Extremely Randomized Forest Method; and the panel in the third column shows the importance coefficients for those surface parameters that are derived from the "height" channel when the machine-learning module **84** uses the Method of Gradient Boosting Trees.

Similarly, in the second row, the panel in the first column shows the importance coefficients for those surface parameters that are derived from the "adhesion" channel when the machine-learning module **84** uses the Random Forest Method; the panel in the second column shows the importance coefficients for those surface parameters that are derived from the "adhesion" channel when the machine-learning module **84** uses the Extremely Randomized Forest Method; and the panel in the third column shows the importance coefficients for those surface parameters that are derived from the "adhesion" channel when the machine-learning module **84** uses the Method of Gradient Boosting Trees.

The histograms in FIG. 9 provide an intelligent way to choose those surface parameters that would be most helpful in correctly classifying a sample. For example, if the machine-learning module **84** were forced to choose only two surface parameters from the channel that measures height, it would probably avoid choosing "h_Sy" and "h_Std" but might instead prefer to choose "h_Ssc" and "h_Sfd."

The importance coefficients in FIG. 9 were arrived at using between a hundred trees and three hundred trees. The maximum number of elements in the chosen subset of original surface parameters was the square root of the number of surface parameters originally provided and the Gini index provided the basis for evaluating classification error. It is apparent from comparing the histograms in the same row that the choice of machine-learning procedure does not make a great deal of difference to the importance of particular surface parameters.

FIG. 10 shows an example of a binary tree from an ensemble of one hundred to three hundred trees used in the bootstrap methods. In the first split, the fourth variable "X[4]" was chosen with a split value of 15.0001. This yielded the Gini index of 0.4992 and split seventy-three samples into two bins having thirty and forty-three samples, respectively.

At the second level split, looking at left hand side node, the sixth variable "X[6]" was chosen with split value of 14.8059, which yielded the Gini index of 0.2778 and split thirty samples (five in class 1 and twenty-five in class 2) into two

bins with twenty seven and three samples, respectively. The split continues until a tree node has the Gini index of zero, thus indicating presence of only one of the two classes.

The method of Extremely Randomized Trees differs from that of the Random Forest in its choice of the split. Instead of computing an optimal parameter and split combination using a Gini index, as was the case for the Random Forest method, a machine-learning module **84** using the method of Extremely Randomized Trees randomly selects each parameter value from the parameter empirical range. To ensure that these random choices eventually converge to a pure node with a zero Gini index, the machine-learning module **84** only chooses the best split among random uniform splits in the set of selected variables for which the current tree is chosen.

In some practices, the machine-learning module **84** implements the method of Gradient Boosting Trees. In this case, the machine-learning module **84** builds a series of trees, each of which converges with respect to some cost function. The machine-learning module **84** builds each subsequent tree to minimize the deviation from the exact prediction, for example by minimizing a mean squared error. In some cases, the machine-learning module **84** relies on the Friedman process for this type of regression. A suitable implementation of this regression process can be carried out using the routine "TREEBOOST" as implemented in the "SCIKIT-LEARN PYTHON" package.

Because the method of Gradient Boosting Trees lacks a criterion for pure nodes, the machine-learning module **84** predefines the size of the tree. Alternatively, the machine-learning module **84** limits the number of individual regressions, thus limiting the maximum depth of a tree.

A difficult that arises is that trees built with predefined sizes can easily be overfitted. To minimize the effect of this difficulty, it is preferable that the machine-learning module **84** impose constraints on such quantities as the number of boosting iterations or that it weaken the iteration rate, for example by using a dimensionless

learning rate parameter. In alternative practices, the machine-learning module **84** limits the minimum number of terminal nodes, or leaves, on a tree.

In the implementations described herein, which relied on the SCIKIT-LEARN PYTHON package, the machine-learning module **84** set the minimum number of leaves to unity and the maximum depth to three. In the application described herein in which bladder cells collected from human subjects were to be classified, the machine-learning module **84** throttled back on its ability to learn by deliberating selecting an unusually low learning rate of 0.01. The resulting slow learning procedure decreases variance that resulted from having a small number of human subjects, and hence a small number of samples.

In creating the training data **87** and the testing data **89**, it is important to avoid dividing the sets $\{M^{(k;1,p)}, M^{(k;2,p)}, \ldots M^{(k;S,p)}\}$ between the training data **87** and testing data **89**. The procedure disclosed in FIG. 11 avoids this.

In the particular implementation of classifying bladder cells **90**, each patient provided several cells, with the image of each cell **90** being divided into four partitions. A human observer visually inspected the partitions in an effort to spot artifacts, two of which can be seen in FIG. 12. If an artifact was found to be present in a partition, then whoever inspected the image would flag that partition as one that is to be ignored.

This process can become tedious when many cells **90** are involved. One can automate this process by using the classifier **100** shown in equation (10) and taking the median of the four partitions. This significantly dilutes the contribution of the artifact.

The machine-learning module **84** randomly splits the database **86** so that S% of its data is in the training data **87** and 100-S% is in the testing data **98.** Experiments were carried out with S set to 50%, 60%, and 70%. The machine-learning module **84** split the database **86** in such a way as to keep data from the same individual entirely in either the training data **87** or the testing data **98** to avoid artificial over-training that

may otherwise result from correlation between different cells **90** of the same individual.

The machine-learning module **84** then causes the condenser **104** to further reduce the number of surface parameters to be relied upon for classification. In some practices, the condenser **104** does so by ranking surface parameters within a particular channel based on their respective Gini indices and keeping some number M_p of the best parameters for that channel. In some practices, the best parameters are selected based on their ability to their segregation power and their low correlation with other surface parameters. For example, by changing the inter-parameter correlation threshold, it becomes possible to change the number of surface parameters that will be relied upon for classification.

FIG. 13 shows how changing the threshold value of the correlation coefficient affects the number of surface parameters selected using the Random Forest Method, with the leftmost panel corresponding to the surface parameters available from the height channel and the middle panel corresponding to the surface parameters available from the adhesion channel. As is apparent from the change to the vertical scale, the rightmost panel represents the combination of the height channel and the adhesion channel. Although FIG. 13 is specific to the Random Forest Method, the other methods have similar curves.

Once the trees have been trained, it is appropriate to test their ability to classify correctly on the testing data **98** or alternatively, to use them to classify unknown samples. The classification process includes obtaining the result of tree voting and using that result as a basis for a probability indicative of what class a sample belongs to. This result is then compared with a classifier threshold that is set based on what error is tolerable. This classifier threshold is typically made to vary as part of building a receiver operating characteristic.

In one experiment, samples of urine **88** were collected from twenty-five cancer-afflicted patients and forty-three cancer-free patients. Of the cancer-afflicted patients, fourteen were low grade and eleven were high grade as defined by TURBT.

The cancer-free patients were either healthy or had had cancer in the past. Using an optical microscope that was coupled to the atomic force microscope **8**, a human observer randomly selected round objects that appeared to be cells.

The database was further reduced by using the data-reduction process referred to in equation (14). The resulting probably generator **100** was therefore $P_{nm}^{(k,s)}$=median$\{P_{nm}^{(k,s;p)}\}$ where p is an integer between 1 and 4 inclusive to correspond with the four partitions of each image. The resulting condensed database has two classes and can be formally represented as:

$$\begin{matrix} D_1^{(1,s)} = \{P_1^{(k,s)}{}_m, C^{(1)}\}, D_2^{(1,s)} = \{P_2^{(k,s)}{}_m, C^{(1)}\} & .. & D_{Subset1}^{(1,s)} = \{P_{Subset1}^{(k,s)}{}_m, C^{(1)}\} \\ D_1^{(2,s)} = \{P_1^{(k,s)}{}_m, C^{(2)}\}, D_2^{(2,s)} = \{P_2^{(k,s)}{}_m, C^{(2)}\} & .. & D_{Subset2}^{(2,s)} = \{P_{Subset2}^{(k,s)}{}_m, C^{(2)}\} \end{matrix} \quad (18)$$

At least five cells were imaged per patient. For the sake of simplicity only two properties were considered: height and adhesion.

FIG. 14 shows the hierarchy of importance of the surface parameters for height and adhesion properties calculated within the Random Forest method. The figure shows the averages of the importance coefficients together with an error bar indicating one standard deviation about the average. The database **86** was randomly split into training data **87** and testing data **89** a thousand times.

The mapped properties for height and adhesion were combined through tensor addition, which is basically the data-reduction method (3-1) adapted for vectors of surface parameters). The relevant tensor addition operation is represented by:

$$P_{nm}^{(1,s)} \oplus P_{nm}^{(2,s)}$$

As was the case in FIG. 9, each surface parameter in FIG. 14 has, as its name, the standard name of the surface parameter but prepending by a letter indicating the mapped property from which it was derived. For example, "a_Sds" means the "Sds" parameter derived from an image of adhesion property.

A suitable statistical performance metric for the Random Forest method comes from inspecting the receiver operating characteristic and the confusion matrix. The

receiver operating characteristic permits defining range of sensitivity and specificity. The range of sensitivity corresponds to "accuracy" when classifying a cell as coming from a cancer-afflicted patient, whereas specificity corresponds to "accuracy" when the cell is classified as from a cancer-free person. The receiver operating characteristic makes it possible to use the receiver operating characteristic to define a range of specificity and a range of sensitivity, as follows:

$$\begin{aligned} \text{sensitivity} &= TP/(TP+FN); \\ \text{specificity} &= TN/(TN+FP); \\ \text{accuracy} &= (TN+TP)/(TP+FN+TN+FP), \end{aligned} \qquad (19)$$

where TN, TP, FP, FN stand for true negative, true positive, false positive, and false negative, respectively.

FIG. 15 shows three different curves, each of which shows the accuracy achieved by considering different numbers of surface parameters, wherein the surface parameters were chosen based choosing different self-correlation thresholds and importance coefficients as described above.

Each of the three different curves in FIG. 15 was arrived at through a thousand random splits between training data **87** and testing data **89**. The curves differ in the allocation of data to each set. A first curve corresponds to 70% of the data being allocated to the training data **87** and 30% being allocated to the testing data **89**. A second curve corresponds to only 60% of the data being allocated to training data **87** and 40% being allocated to the testing data **89**. And a third curve corresponds correspond to an even split between training data **87** and testing data **89**.

It is apparent from inspection of FIG. 15 that there is virtually no dependence on a particular threshold split. This indicates robustness of the procedure carried out by the machine-learning module **84**.

FIG. 16 shows a family of receiver operating characteristics. The individual receiver operating characteristic in the family of characteristics shown in FIG. 16

arose from two hundred different random splits of the database **86** into training data **87** and testing data **89**.

Each receiver operating characteristic shows sensitivity and specificity for different thresholds when attempting to classify between two classes. The diagonal line that bisects the plot in FIG. 16 amounts to a classifier that classifies by flipping a coin. Thus, the closer a receiver operating characteristic comes to the diagonal line shown in FIG. 16, the poorer its classifier is at classifying. The fact that the curves are clustered far from this diagonal line with little variation between individual curves suggests both the effectiveness of the classifier and its insensitivity to the specific choice of training data **87** and testing data **89**.

In constructing a receiver operating characteristic, the threshold that defines whether a particular probability value corresponds to one class or the other is a free parameter. The choice of this parameter governs both specificity and sensitivity. For each receiver operating characteristic, there exists a point that corresponds to the minimum error in classifying a sample that should have been in the first class into the second class and vice versa. This is shown in FIG. 21 for each of the three machine-learning methods used when using a single channel.

Each row in the table shown in FIG. 21 is characterized by a particular number of collected cells (N) and a smaller number (M) that was used as a threshold for diagnosis. For each row, there were two channels considered: height and adhesion. For each of the three machine-learning methods used, the table shows the averaged AUC and accuracy for a thousand random splits of the database into training data and testing data with 70% of the database being allocated to the training data. The accuracy is that associated with the smallest error in classification. Each row in FIG. 21 also shows sensitivity and specificity.

In principle, the sensitivity and specificity can also be defined around a balanced point in which sensitivity and specificity are equal. Because of a limited number of human subjects, it is difficult to define precisely where this balanced point would be. Thus, in FIG. 21, the requirement for equality was relaxed and a balance

range was defined in which the magnitude of the difference between sensitivity and specificity had to be less than a selected value, which for FIG. 21 was 5%.

Only ten surface parameters were used to calculate the receiver operating characteristic. As was apparent from FIG. 15, there is a point of diminishing returns at which adding more surface parameters does not significantly improve accuracy. According to FIG. 15, it is apparently sufficient to use only eight to ten judiciously chosen surface parameters to achieve a relatively high accuracy of 80%. The top ten surface parameters were considered to characterize the statistical behavior of the receiver operating characteristic and the confusion matrix, including the specificity, sensitivity, and accuracy of the classifier **100**.

The process of classifying a cell as having come from a cancer-free patient or a cancer-afflicted patient relies on averaging the probability obtained for that cell over all repetitions of the procedure used to acquire that probability. This is formally represented as:

$$\mathrm{Prob}^{(i)}{}_s = \frac{1}{S}\sum_s \mathrm{Prob}^{(s)(i)}{}_s, \text{ where } \mathrm{Prob}^{(s)(i)}{}_s = \mathrm{AI}(P_s^{(0,s)}{}_m \oplus P_s^{(2,s)}{}_m \mid C^{(i)}), \quad (20)$$

where the classifier *AI* was developed using the machine learning methods developed on the training database **87**. According to this procedure, and assuming class 1 represents a cancer cell, a cell is identified as having come from a cancer-afflicted patient if $\mathrm{Prob}_m^{(i)}$ exceeds a particular threshold, which can be obtained from the receiver operating characteristic.

In an effort to confirm the veracity of the data shown in FIGS 18 and 19, a control experiment was carried out with the same procedure as was used for FIGS. 19 and 20 but with the samples to be classified having been split evenly between cancer cells and healthy cells. FIGS. 17 and 18 show the result of a thousand random choices of classification. It is apparent that the accuracy has dropped to 53%±10%, which is consistent with expectations. This suggests the reliability of the data shown in FIGS. 19 and 20 as well as the classifier's resistance to overtraining, which is a common problem that arises when a machine-learning method is made to cope with too many parameters.

An alternative method of classification relies on more than one cell to establish a diagnosis of a patient. This avoids a lack of robustness based on a high sampling error. Moreover, this avoids error that arises because one cannot be sure that a cell found in urine **88** is actually from the bladder itself. Other parts of the urinary tract are perfectly capable of shedding cells. In addition, urine **88** can contain an assortment of other cells, such as exfoliated epithelial cells from other parts of urinary tract. One such classification method includes diagnosing a patient with cancer if the number of cells *M* classified as having come from a cancer-afflicted patient out of the total number of cells classified *N* is greater or equal to a predefined value. This is a generalization of the previously discussed case in which *N*=*M*=1.

The probability of having cancer based on probabilities for *N* cells can be assigned using algorithms (3-2) - (3-9) or (10) - (14). As a preferable procedure to define the probability of classifying the *N* tested cells as coming from a cancer patient (Class 1) is as follows:

$$\mathrm{Prob}^{(1)}{}_{n}=\underset{s=1..N}{Max}\{\mathrm{Prob}^{(s)(1)}{}_{n}\} \text{ where } \mathrm{Prob}^{(s)(1)}{}_{n}=\mathrm{AI}(P_{n}^{(1,s)}{}_{m}\oplus P_{n}^{(2,s)}{}_{m}\mid C^{(1)}), \quad (21)$$

where the classifier *AI* is developed from the training database **87.**

FIGS. 19 and 20 show accuracy and receiver operating characteristics similar robustness to those in FIGS. 15 and 16 but for the case of N=5 and M=2. One can see that the accuracy of such method can reach 94%. The randomization test described above shows 50±22% for the area under receiver operating characteristic curves (the result of a thousand random choices of diagnosis sets). These imply the lack of overtraining.

The results of calculation of the confusion matrix for multiple N and M are shown in Figure 20 table exampled for two single channels (height and adhesion). The robustness of combined channels is better compared to the diagnostic based on single channels.

The procedure described above can also be applied to classify cancer free patients. In such a case, the probabilities discussed above are the probabilities that the cell belongs to a cancer free patient.

Having described the invention and a preferred embodiment thereof, what is claimed as new and secured by letters patent is:

CLAIMS

1. A method comprising, using an atomic-force microscope to acquire a set of images associated with surfaces, combining said images, and, using a machine-learning method applied to said combined images, classifying said surfaces.

2. The method of claim 1, further comprising processing said images to obtain surface parameters and using machine learning to classify said surfaces based at least in part on said surface parameters.

3. The method of claim 1, wherein acquiring said set of images comprises using a multi-channel atomic-force microscope in ringing mode, wherein each channel of said atomic-force microscope provides information indicative of a corresponding surface property of said surfaces.

4. The method of claim 1, further comprising selecting said surfaces to be surfaces of cells collected from urine of a subject and classifying said cells as indicative of cancer or not indicative of cancer.

5. The method of claim 1, wherein using said atomic-force microscope comprises using a multi-channel atomic force microscope, wherein each channel corresponds to a surface property of said surfaces.

6. The method of claim 1, wherein using said atomic-force microscope comprises using said atomic-force microscope in sub-resonant tapping mode.

7. The method of claim 1, wherein using an atomic-force microscope comprises acquiring multiple channels of information, each of which corresponds to a different surface property of said surface, said method further comprising condensing information provided by said channels and constructing, from said condensed information, a condensed database.

8. The method of claim 7, further comprising forming a first database based on said information provided by said channels, wherein constructing said condensed database comprise projecting said first database into a subspace of

dimensionality lower than that of said first database, said projection defining said condensed database, said condensed database having a dimensionality that is less than that of said first database.

9. The method of claim **7**, further comprising forming a first database based on said information provided by said channels, said first database having indices, said method further comprising deriving a condensed database from said first database, said condensed database having fewer indices than said first database.

10. The method of claim **9**, wherein deriving said condensed database comprises carrying out tensor addition to generate tensor sums that combine information from said first database along one or more slices corresponding to one or more indices of said first database and forming said condensed database using said tensor sums.

11. The method of claim **9**, wherein deriving a condensed database from said first database comprises defining a subset of values from said first database, each of said values being representative of a corresponding element in said first database, deriving a condensed value from said values in said subset of values, and representing said corresponding elements from said first database with said condensed value, wherein deriving said condensed value comprises summing said values in said subset of values.

12. The method of claim **11**, wherein summing said values comprises carrying out tensor addition to generate tensor sums that combine values from said first database along one or more slices corresponding to corresponding indices of said first database and forming a condensed database using said tensor sums.

13. The method of claim **9**, wherein deriving a condensed database from said first database comprises defining a subset of values from said first database, each of said values being representative of a corresponding element in said first database, deriving a condensed value from said values in said subset of values, and representing said corresponding elements from said first database with

said condensed value, wherein deriving said condensed value comprises averaging said values in said subset of values.

14. The method of claim **13**, wherein averaging said values comprises obtaining an arithmetic average.

15. The method of claim **13**, wherein averaging said values comprises obtaining a geometric average.

16. The method of claim **9**, wherein deriving a condensed database from said first database comprises defining a subset of values from said first database, each of said values being representative of a corresponding element in said first database, deriving a condensed value from said values in said subset of values, and representing said corresponding elements from said first database with said condensed value, wherein said condensed value is one of a maximum or a minimum of said values in said subset of values.

17. The method of claim **9**, wherein deriving a condensed database from said first database comprises defining a subset of values from said first database, each of said values being representative of a corresponding element in said first database, deriving a condensed value from said values in said subset of values, and representing said corresponding elements from said first database with said condensed value, wherein deriving said condensed value comprises passing information from said first database through a surface-parameter extractor to obtain a surface-parameter set.

18. The method of claim **17**, further comprising normalizing said surface parameters representative of said surface-parameter set to be independent of surface areas of images from which they were derived.

19. The method of claim **17**, further comprising dividing said surface parameter by another parameter of the same dimension.

20. The method of claim **1**, said method further comprising automatically detecting that an image of a sample has an artifact and automatically excluding said image from being used for classifying said surfaces.

21. The method of claim **1**, further comprising partitioning an image of a sample into partitions, obtaining surface parameters for each partition, and defining a surface parameter of said cell as being the median of said surface parameters for each partition.

22. The method of claim **1**, further comprising processing said images to obtain surface parameters and using machine learning to classify said surfaces based at least in part on said surface parameters and from externally-derived parameters.

23. The method of claim **22**, wherein said surfaces are surfaces of bodies that are derived from collected samples, at least one of said samples being a body-free sample that has no bodies, said method further comprising selecting said externally-derived parameters to include data indicative of an absence of bodies from said body-free sample.

24. The method of claim **23**, further comprising assigning an artificial surface parameter to said body-free sample.

25. The method of claim **22**, wherein said surfaces are surfaces of cells derived from samples obtained from a patient, said method further comprising selecting said externally-derived parameters to include data indicative of a probability that said patient has a particular disease.

26. The method of claim **25**, wherein said data indicative of said probability includes said patient's age.

27. The method of claim **25**, wherein said data indicative of said probability includes said patient's smoking habits.

28. The method of claim **25**, wherein said data indicative of said probability includes said patient's family history.

29. The method of claim **2**, said method further comprising defining a subset of said surface parameters and generating a database based on said subset, wherein defining said subset of surface parameters comprises determining a correlation between said surface parameters, comparing said correlation with a threshold to identify a set of correlated parameters, and including a subset of said set of correlated parameters in said subset of surface parameters.

30. The method of claim **2**, said method further comprising defining a subset of said surface parameters and generating a database based on said subset, wherein defining said subset of surface parameters comprises determining a correlation matrix between said surface parameters and wherein determining said correlation matrix comprises generating simulated surfaces.

31. The method of claim **2**, said method further comprising defining a subset of said surface parameters and generating a database based on said subset, wherein defining said subset of surface parameters comprises combining different surface parameters of the same kind from the same sample.

32. The method of claim **1**, wherein using a machine-learning method comprises using the Random Forest Method.

33. The method of claim **1**, wherein using a machine-learning method comprises using the Extremely Randomized Forest Method.

34. The method of claim **1**, wherein using a machine-learning method comprises using the method of Gradient Boosting Trees.

35. The method of claim **1**, wherein using a machine-learning method comprises using a neural network.

36. The method of claim 1, wherein using a machine-learning method comprises using at least two methods selected from the group consisting of the Gradient

Boosting Trees, the Extremely Randomized Forest Method, and the Random Forest Method.

37. The method of claim **1**, wherein using a machine-learning method comprises using a method of decision trees.

38. The method of claim **1**, wherein said surfaces are surfaces of a first plurality of cells from a patient, wherein a second plurality of said cells has been classified as having come from a cancer-afflicted patient and a third plurality of said cells has been classified as having come from a cancer-free patient, said method further comprising diagnosing said patient with cancer if a ratio of said second plurality to said first plurality exceeds a predetermined threshold.

39. The method of claim **1**, wherein said atomic-force microscope comprises a cantilever and a probe disposed at a distal tip of said cantilever, where said cantilever has a resonant frequency and wherein using said using said atomic-force microscope comprises causing a distance between said probe and said surface to oscillate at a frequency that is less than said resonant frequency.

40. The method of claim **1**, wherein using said atomic-force microscope comprises using a microscope that has been configured to output multiple channels of information corresponding to different physical properties of said sample surface.

41. The method of claim **1**, further comprising processing said images to obtain surface parameters and using machine learning to classify said surfaces based at least in part on said surface parameters and from externally-derived parameters, wherein said surfaces are surfaces of cells derived from samples obtained from a patient, at least one of said samples being a cell-free sample that has no cells from said patient, said method further comprising selecting said externally-derived parameters to include data indicative of an absence of cells from said cell-free sample.

42. The method of claim **41**, further comprising assigning an artificial surface parameter to said cell-free sample.

43. The method of claim **1**, wherein said images are images of cells, wherein said method further comprises automatically detecting that an image of a cell has an artifact and automatically excluding said image from being used for classifying said surfaces.

44. The method of claim **1**, wherein said images are images of cells and wherein said method further comprises partitioning an image of a simple into partitions, obtaining surface parameters for each partition, and defining a surface parameter of said cell as being the median of said surface parameters for each partition.

45. An apparatus comprising an atomic force microscope for acquiring images associated with surfaces and a processing system that receives signals from said atomic force microscope representative of said images and combines said images, said processing system comprising a machine-learning module and a classifier that classifies an unknown sample after having learned a basis for classification from said machine-learning module.

46. The apparatus of claim **45**, wherein said processing system is configured to process said images to obtain surface parameters and to use said machine-learning module to classify said surfaces based at least in part on said surface parameters.

47. The apparatus of claim **46**, wherein said atomic-force microscope comprises a multi-channel atomic force microscope, each channel of which corresponds to a surface property of said surfaces.

48. The apparatus of claim **47**, wherein said processing system comprises a condenser that condenses information provided by said channels and constructs, from said condensed information, a condensed database.

49. The apparatus of claim **48**, further comprising a classifier that classifies an unknown sample based on said condensed database.

50. The apparatus of claim **48**, wherein said condenser is configured to construct said condensed database by projecting said first database into a subspace of dimensionality lower than that of said first database, said projection defining said condensed database, said condensed database having a dimensionality that is less than that of said first database.

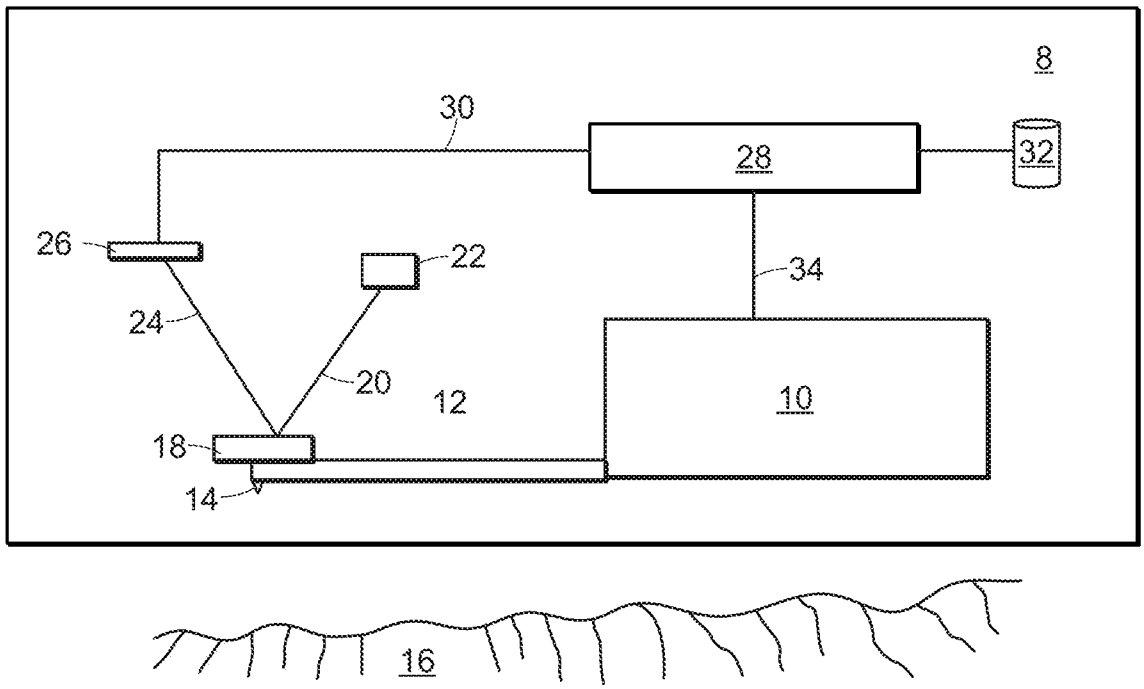

8
30
28
32
26
22
34
24
20
12
10
18
14
16

FIG. 1

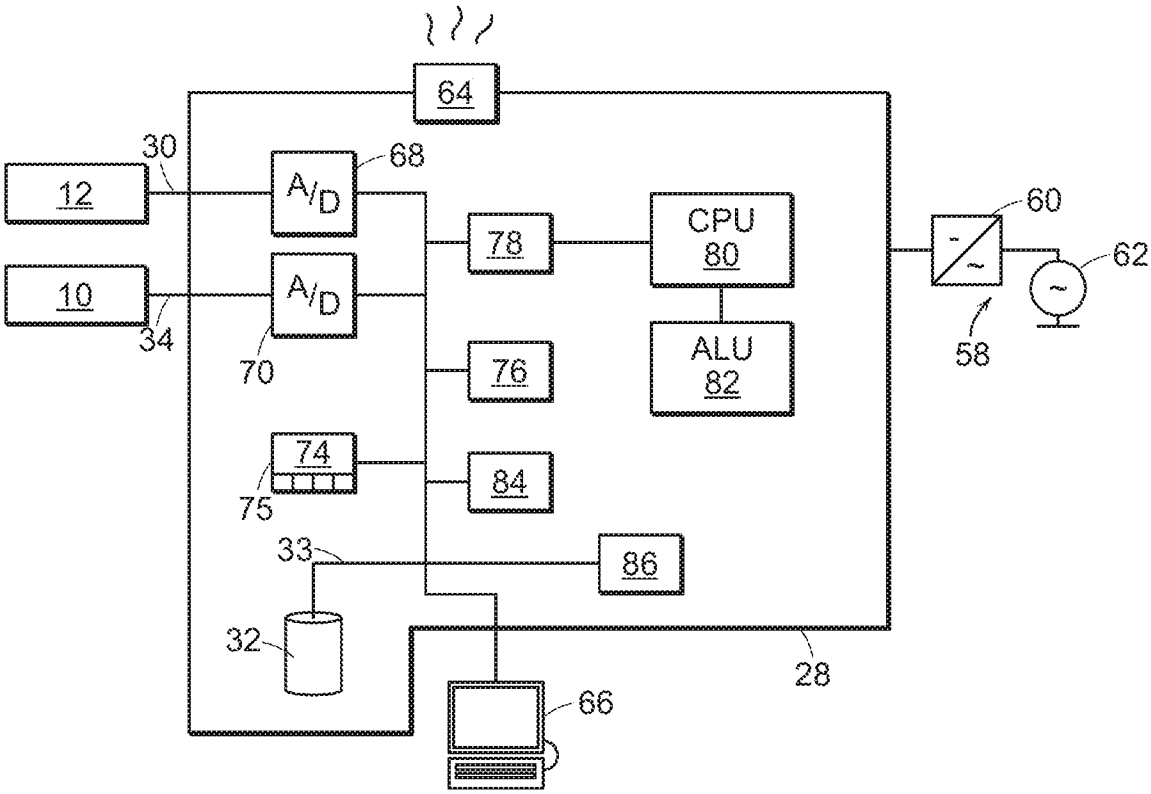

64
30
68
12
A/D
78
CPU
80
60
62
10
A/D
34
70
ALU
82
58
76
74
75
84
33
86
32
28
66

FIG. 2

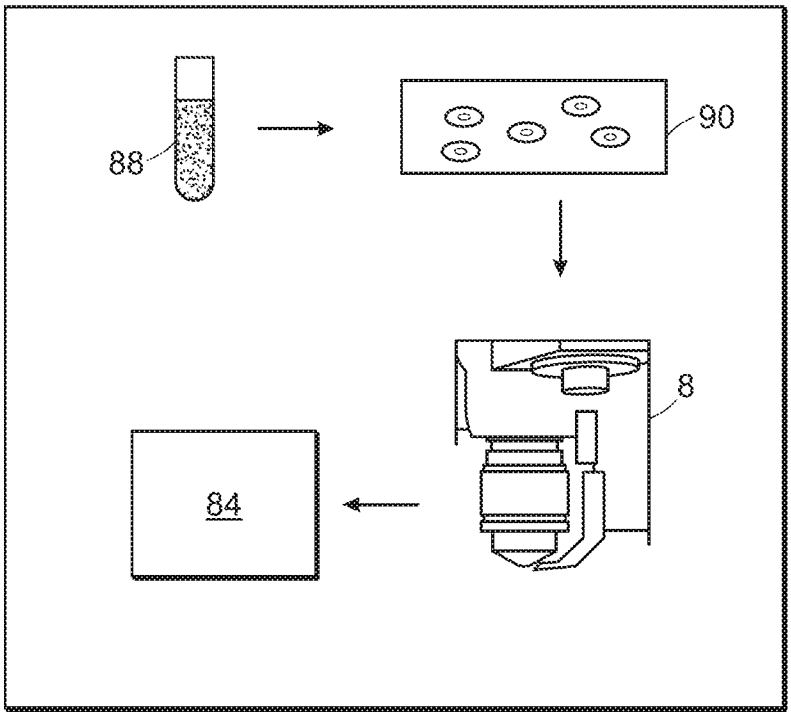


FIG. 3

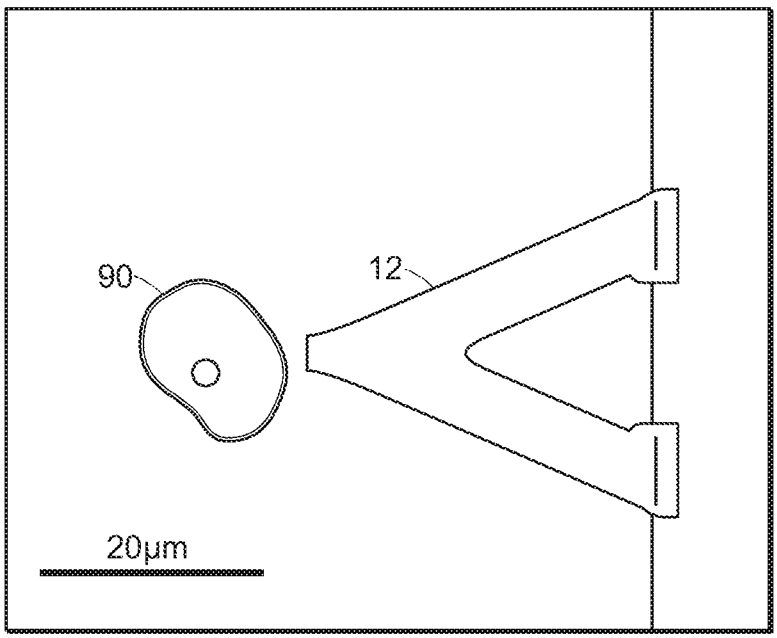


FIG. 4

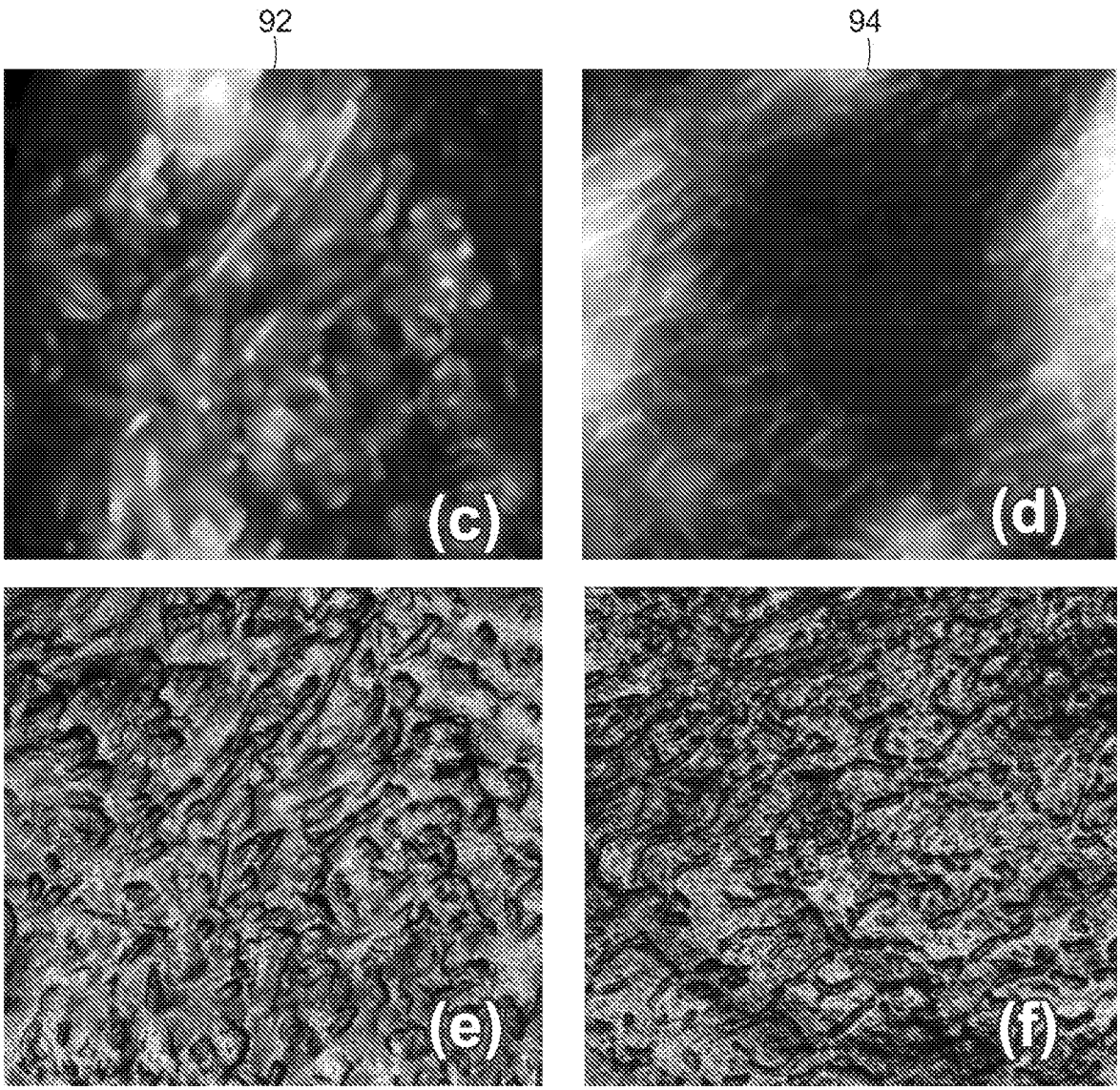


FIG. 5

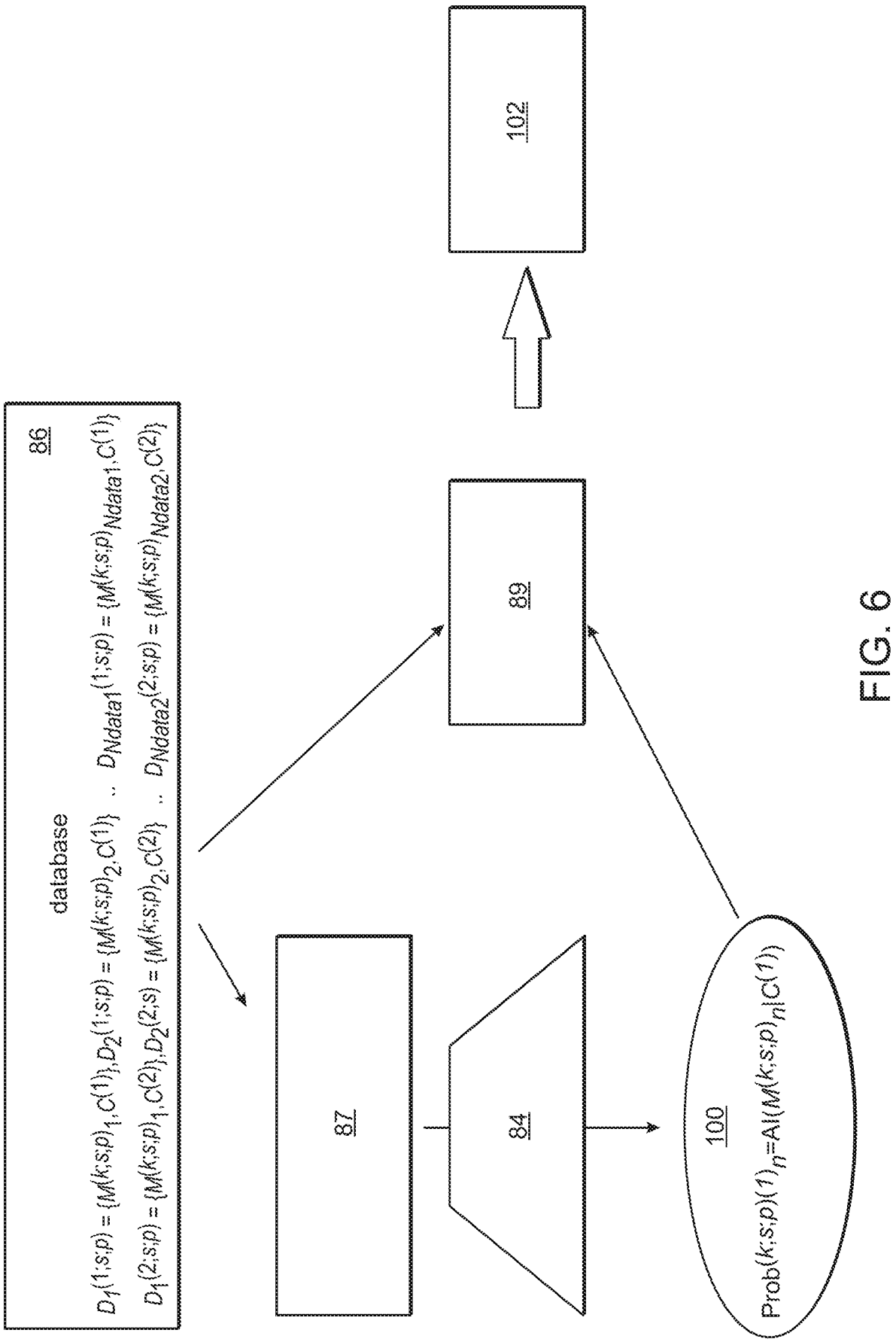

86
database
$D_1(1;s;p) = \{M(k;s;p)_1, C(1)\}, D_2(1;s;p) = \{M(k;s;p)_2, C(1)\}$ .. $D_{Ndata1}(1;s;p) = \{M(k;s;p)_{Ndata1}, C(1)\}$
$D_1(2;s;p) = \{M(k;s;p)_1, C(2)\}, D_2(2;s) = \{M(k;s;p)_2, C(2)\}$ .. $D_{Ndata2}(2;s;p) = \{M(k;s;p)_{Ndata2}, C(2)\}$
87
84
100
$Prob(k;s;p)(1)_n = AI(M(k;s;p)_n | C(1))$
89
102

FIG. 6

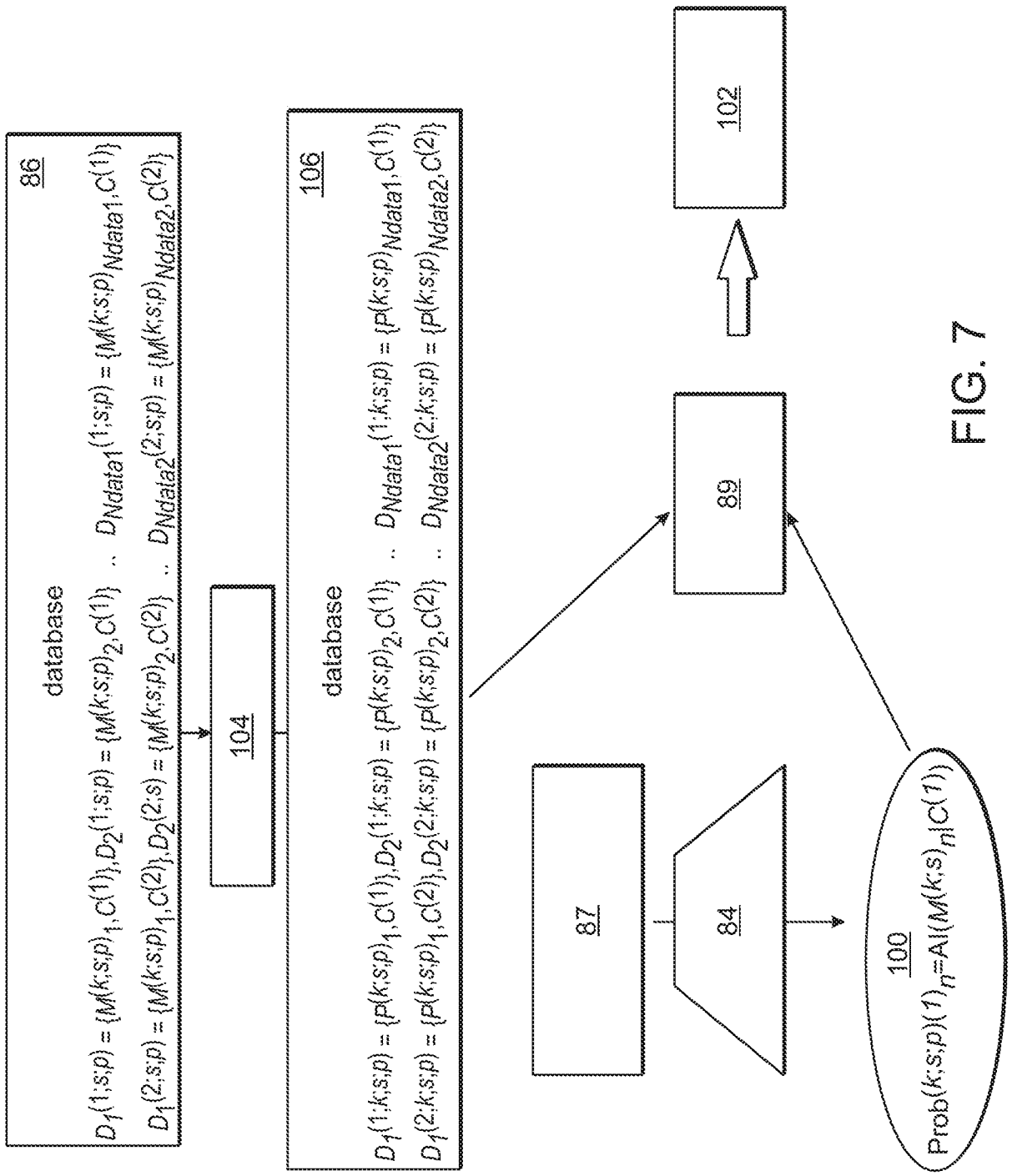

86
database
$D_1(1;s;p) = \{M(k;s;p)_1, C(1)\}, D_2(1;s;p) = \{M(k;s;p)_2, C(1)\}$ .. $D_{Ndata1}(1;s;p) = \{M(k;s;p)_{Ndata1}, C(1)\}$
$D_1(2;s;p) = \{M(k;s;p)_1, C(2)\}, D_2(2;s) = \{M(k;s;p)_2, C(2)\}$ .. $D_{Ndata2}(2;s;p) = \{M(k;s;p)_{Ndata2}, C(2)\}$
104
106
database
$D_1(1;k;s;p) = \{P(k;s;p)_1, C(1)\}, D_2(1;k;s;p) = \{P(k;s;p)_2, C(1)\}$ .. $D_{Ndata1}(1;k;s;p) = \{P(k;s;p)_{Ndata1}, C(1)\}$
$D_1(2;k;s;p) = \{P(k;s;p)_1, C(2)\}, D_2(2;k;s;p) = \{P(k;s;p)_2, C(2)\}$ .. $D_{Ndata2}(2;k;s;p) = \{P(k;s;p)_{Ndata2}, C(2)\}$
87
84
89
102
100
$Prob(k;s;p)(1)_n = AI(M(k;s)_n | C(1))$

FIG. 7

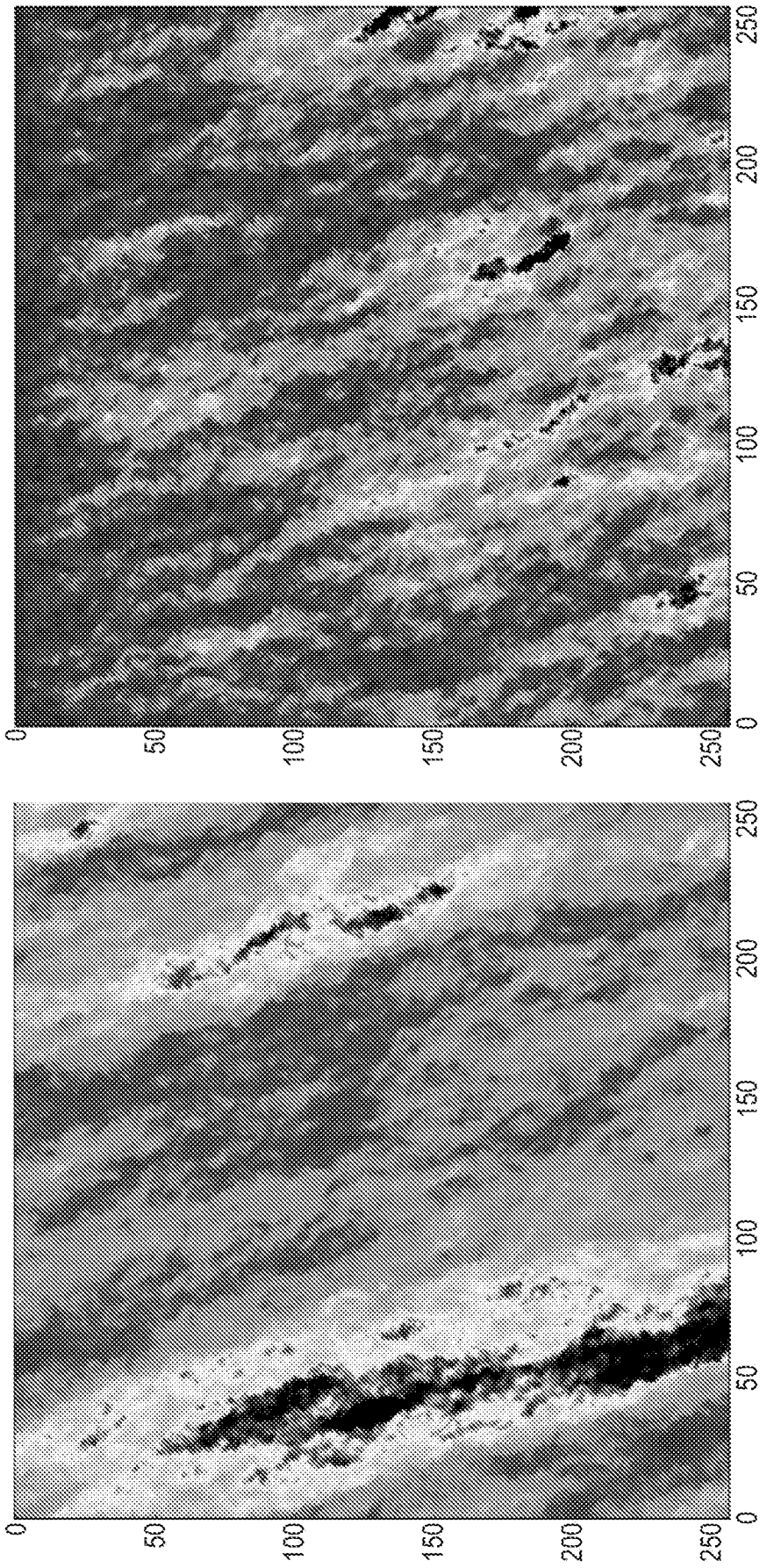


FIG. 8

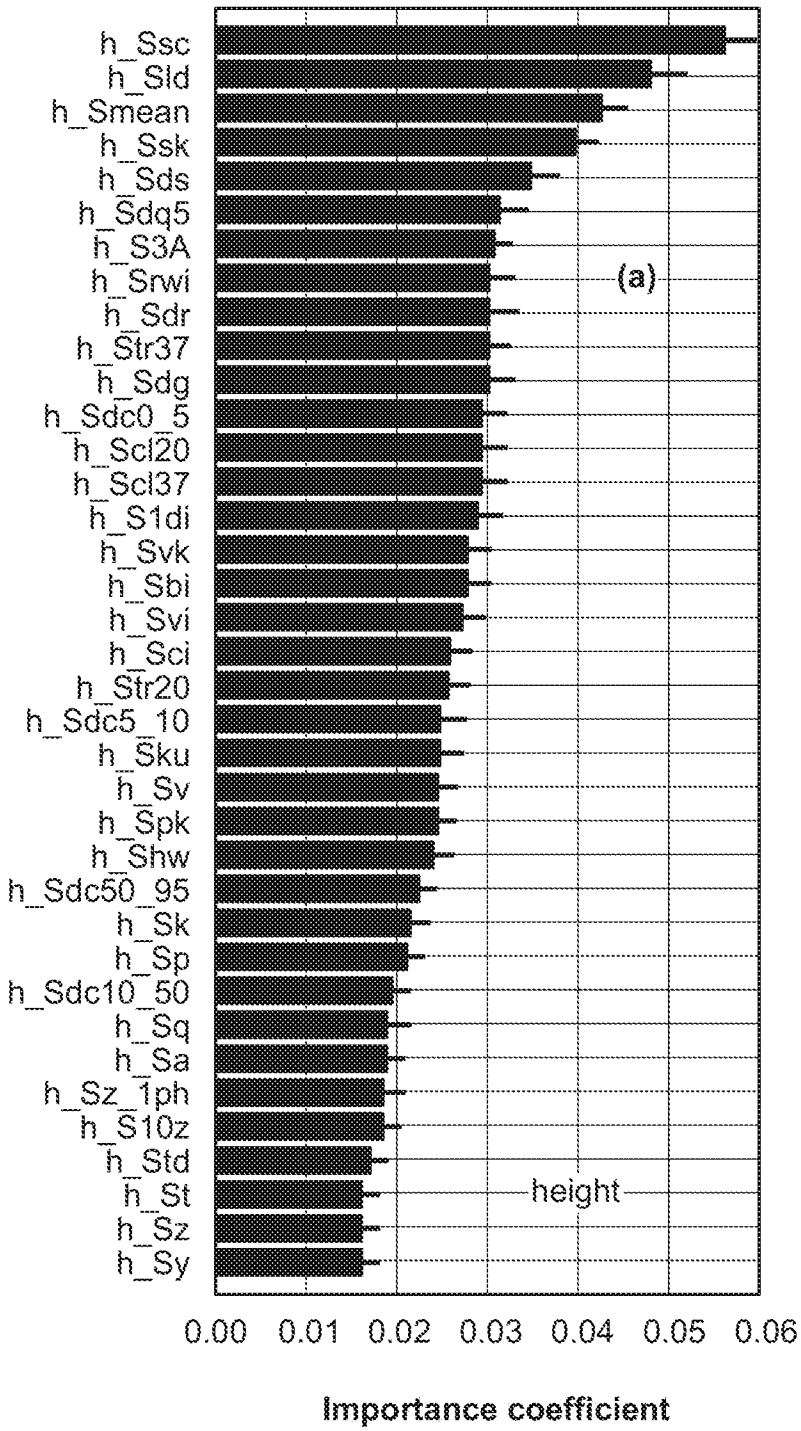

h_Ssc
h_Sld
h_Smean
h_Ssk
h_Sds
h_Sdq5
h_S3A
h_Srwi
h_Sdr
h_Str37
h_Sdg
h_Sdc0_5
h_Scl20
h_Scl37
h_S1di
h_Svk
h_Sbi
h_Svi
h_Sci
h_Str20
h_Sdc5_10
h_Sku
h_Sv
h_Spk
h_Shw
h_Sdc50_95
h_Sk
h_Sp
h_Sdc10_50
h_Sq
h_Sa
h_Sz_1ph
h_S10z
h_Std
h_St
h_Sz
h_Sy
(a)
height
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Importance coefficient

FIG. 9

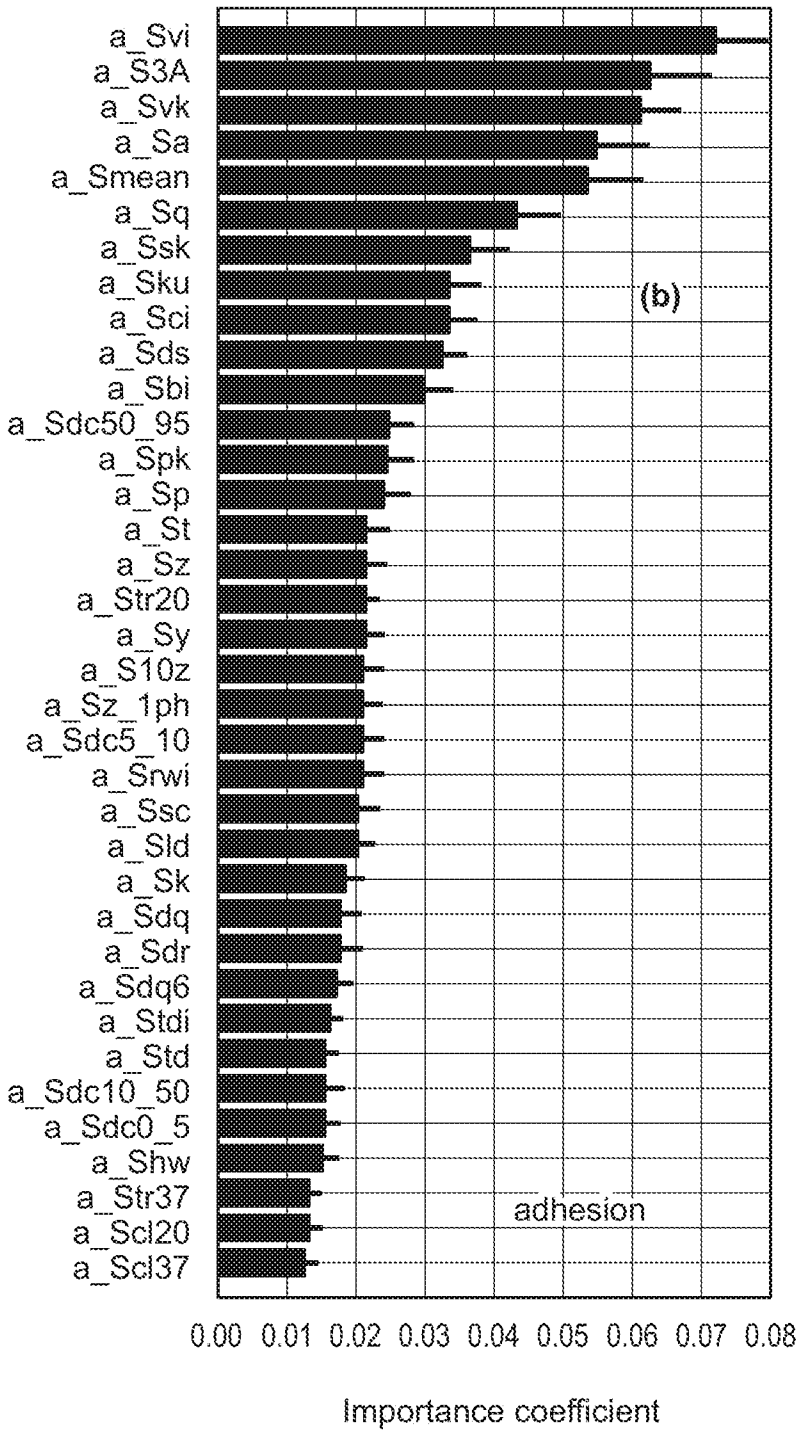

a_Svi
a_S3A
a_Svk
a_Sa
a_Smean
a_Sq
a_Ssk
a_Sku
a_Sci
a_Sds
a_Sbi
a_Sdc50_95
a_Spk
a_Sp
a_St
a_Sz
a_Str20
a_Sy
a_S10z
a_Sz_1ph
a_Sdc5_10
a_Srwi
a_Ssc
a_Sld
a_Sk
a_Sdq
a_Sdr
a_Sdq6
a_Stdi
a_Std
a_Sdc10_50
a_Sdc0_5
a_Shw
a_Str37
a_Scl20
a_Scl37
(b)
adhesion
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
Importance coefficient

FIG. 9 (cont.)

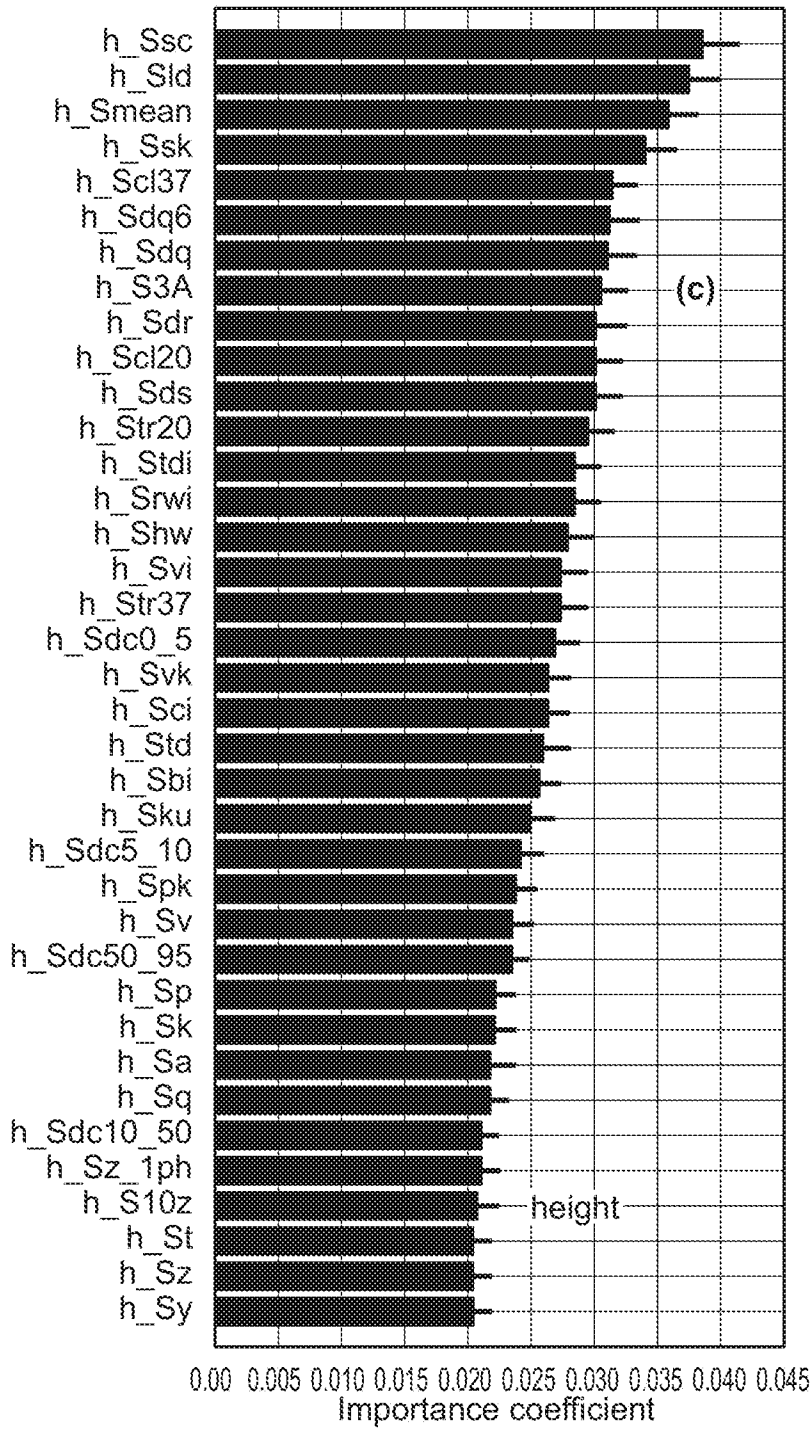

h_Ssc
h_Sld
h_Smean
h_Ssk
h_Scl37
h_Sdq6
h_Sdq
h_S3A
h_Sdr
h_Scl20
h_Sds
h_Str20
h_Stdi
h_Srwi
h_Shw
h_Svi
h_Str37
h_Sdc0_5
h_Svk
h_Sci
h_Std
h_Sbi
h_Sku
h_Sdc5_10
h_Spk
h_Sv
h_Sdc50_95
h_Sp
h_Sk
h_Sa
h_Sq
h_Sdc10_50
h_Sz_1ph
h_S10z
h_St
h_Sz
h_Sy
(c)
height
0.00 0.005 0.010 0.015 0.020 0.025 0.030 0.035 0.040 0.045
Importance coefficient

FIG. 9 (cont.)

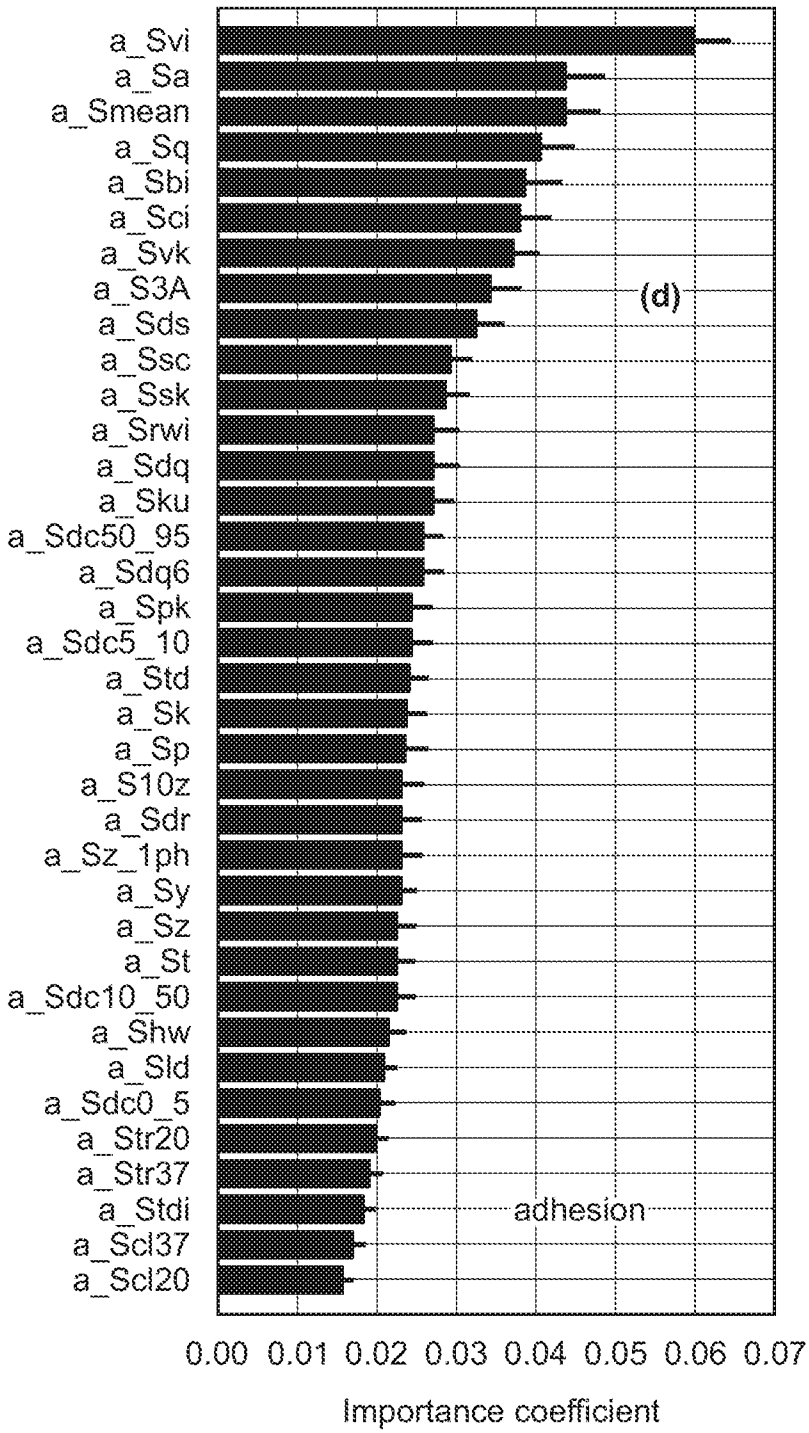

a_Svi
a_Sa
a_Smean
a_Sq
a_Sbi
a_Sci
a_Svk
a_S3A
a_Sds
a_Ssc
a_Ssk
a_Srwi
a_Sdq
a_Sku
a_Sdc50_95
a_Sdq6
a_Spk
a_Sdc5_10
a_Std
a_Sk
a_Sp
a_S10z
a_Sdr
a_Sz_1ph
a_Sy
a_Sz
a_St
a_Sdc10_50
a_Shw
a_Sld
a_Sdc0_5
a_Str20
a_Str37
a_Stdi
a_Scl37
a_Scl20
(d)
adhesion
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07
Importance coefficient

FIG. 9 (cont.)

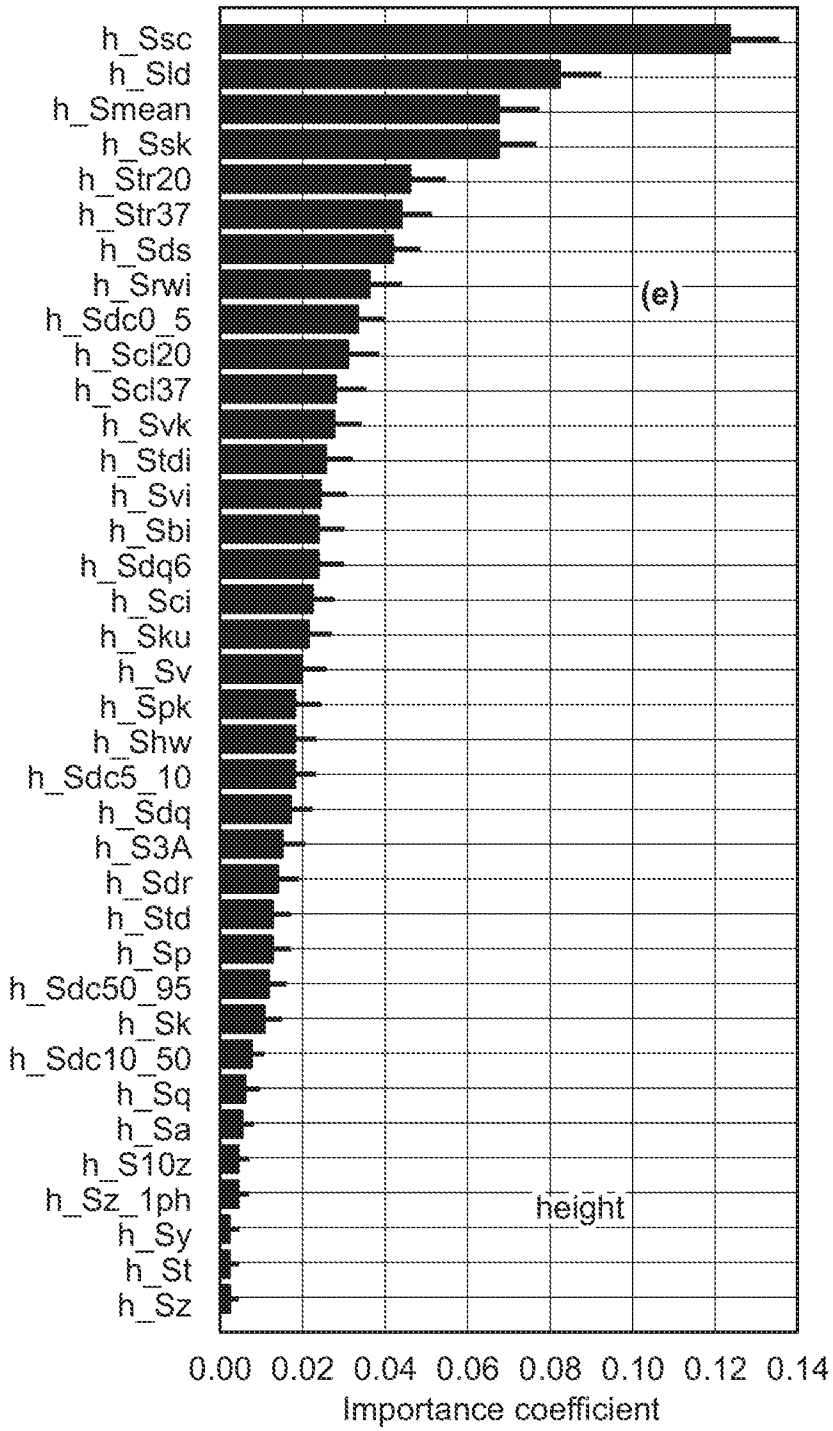

h_Ssc
h_Sld
h_Smean
h_Ssk
h_Str20
h_Str37
h_Sds
h_Srwi
h_Sdc0_5
h_Scl20
h_Scl37
h_Svk
h_Stdi
h_Svi
h_Sbi
h_Sdq6
h_Sci
h_Sku
h_Sv
h_Spk
h_Shw
h_Sdc5_10
h_Sdq
h_S3A
h_Sdr
h_Std
h_Sp
h_Sdc50_95
h_Sk
h_Sdc10_50
h_Sq
h_Sa
h_S10z
h_Sz_1ph
h_Sy
h_St
h_Sz
(e)
height
0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14
Importance coefficient

FIG. 9 (cont.)

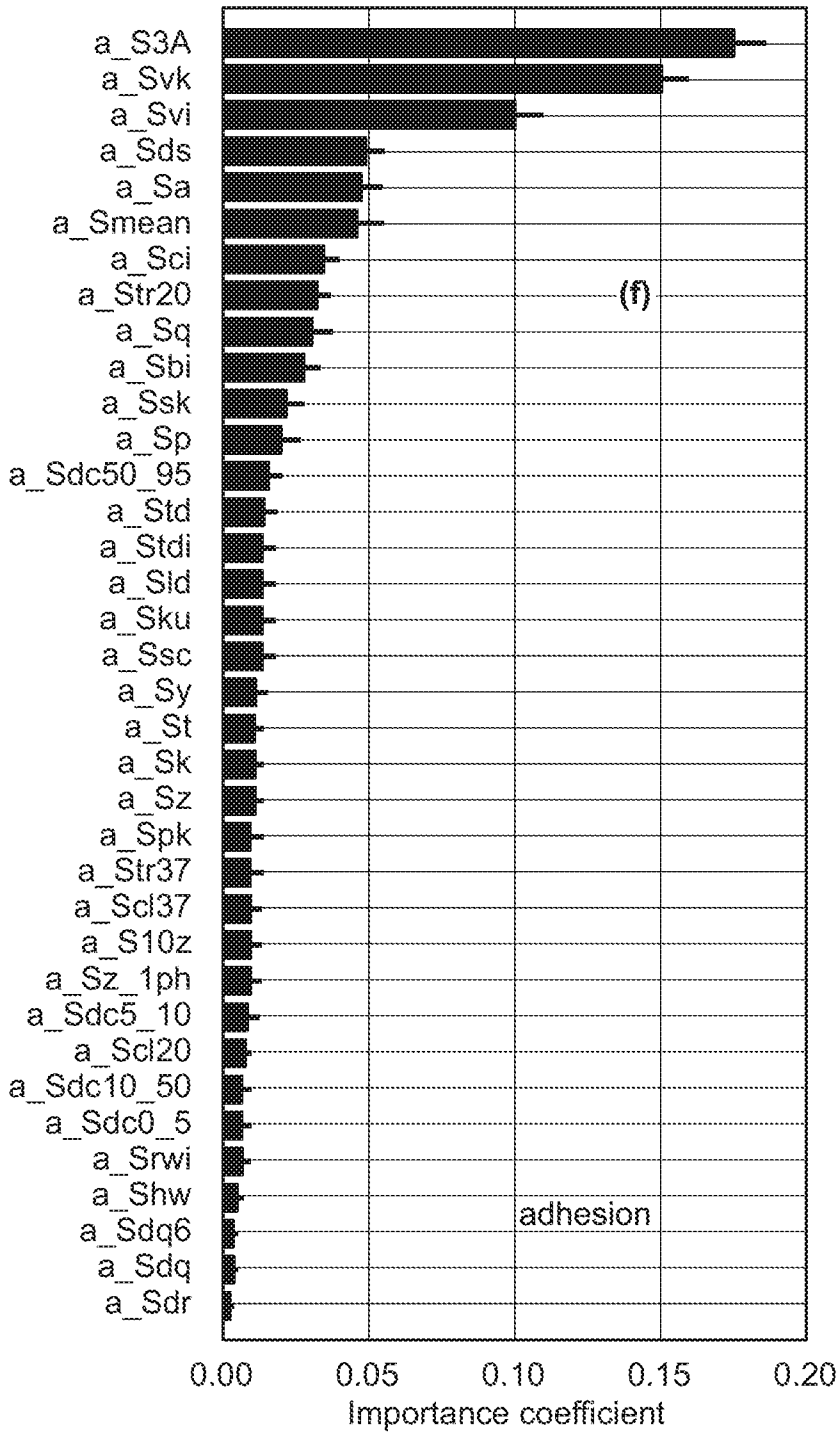

a_S3A
a_Svk
a_Svi
a_Sds
a_Sa
a_Smean
a_Sci
a_Str20
a_Sq
a_Sbi
a_Ssk
a_Sp
a_Sdc50_95
a_Std
a_Stdi
a_Sld
a_Sku
a_Ssc
a_Sy
a_St
a_Sk
a_Sz
a_Spk
a_Str37
a_Scl37
a_S10z
a_Sz_1ph
a_Sdc5_10
a_Scl20
a_Sdc10_50
a_Sdc0_5
a_Srwi
a_Shw
a_Sdq6
a_Sdq
a_Sdr
(f)
adhesion
0.00
0.05
0.10
0.15
0.20
Importance coefficient

FIG. 9 (cont.)

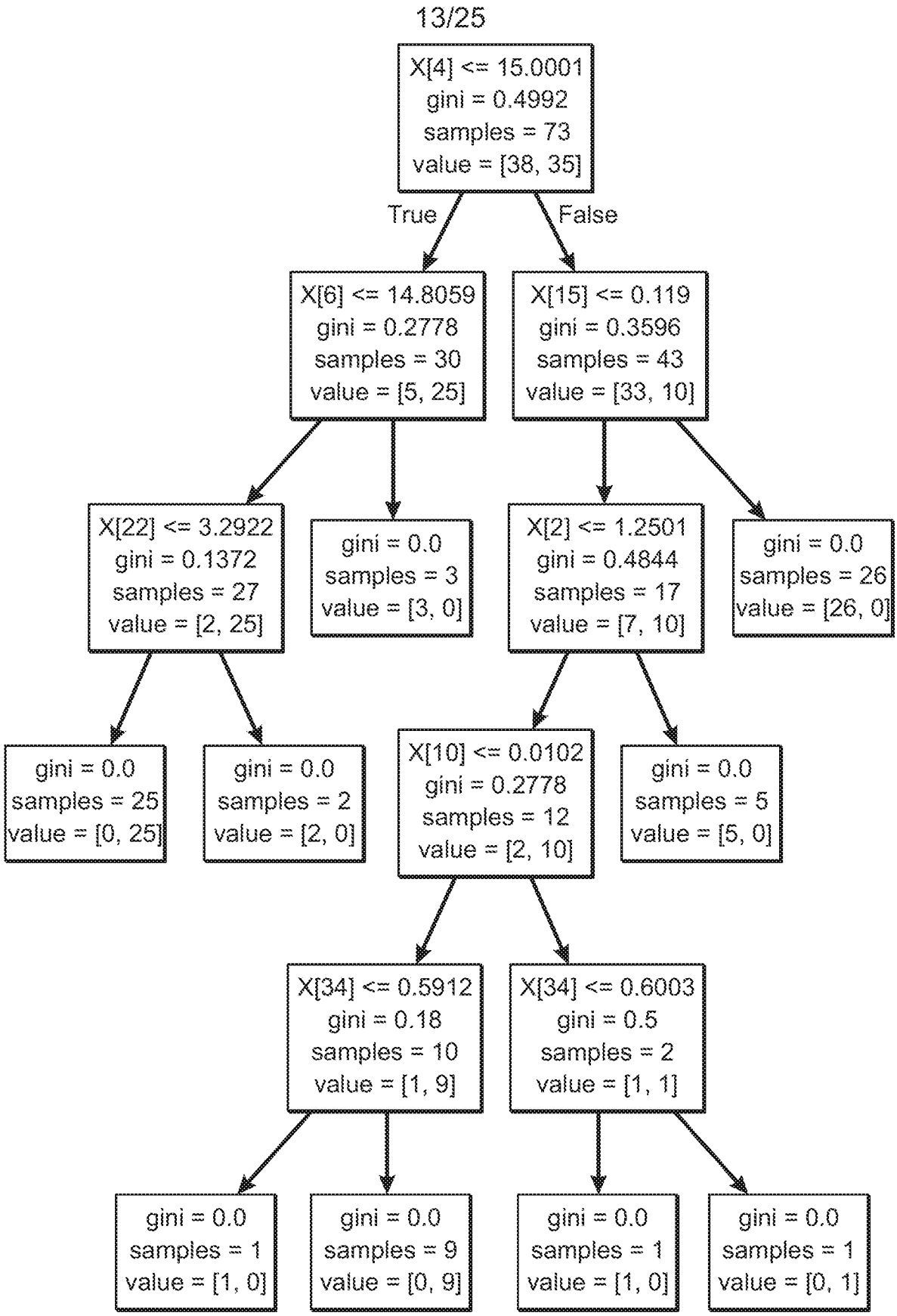

X[4] <= 15.0001
gini = 0.4992
samples = 73
value = [38, 35]
True
False
X[6] <= 14.8059
gini = 0.2778
samples = 30
value = [5, 25]
X[15] <= 0.119
gini = 0.3596
samples = 43
value = [33, 10]
X[22] <= 3.2922
gini = 0.1372
samples = 27
value = [2, 25]
gini = 0.0
samples = 3
value = [3, 0]
X[2] <= 1.2501
gini = 0.4844
samples = 17
value = [7, 10]
gini = 0.0
samples = 26
value = [26, 0]
gini = 0.0
samples = 25
value = [0, 25]
gini = 0.0
samples = 2
value = [2, 0]
X[10] <= 0.0102
gini = 0.2778
samples = 12
value = [2, 10]
gini = 0.0
samples = 5
value = [5, 0]
X[34] <= 0.5912
gini = 0.18
samples = 10
value = [1, 9]
X[34] <= 0.6003
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 9
value = [0, 9]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]

FIG. 10

Algorithm Cancer Detection using Cell Surface Parameters

function make_dataset() **Inputs:** CSV file, statistics parameter [mean or median]

Steps:

1) Load data from CSV file
2) Reject variables where all samples have the same value
3) Reject samples where data acquisition artifacts are present
4) Store data-frame with preprocessed data
5) Create data-frame with non-cancer samples with no cells
6) Merge preprocessed and non-cancer sample data-frames
7) Group samples by patient ID
8) Get mean or median values of all variables per cell, from cell quadrants, for all patient IDs and store in data-frame

Return: Cell samples data-frame

function split_data_train_test() **Inputs:** data-frame, train-test sample ratio

Steps:

1) Split data-frame into training and testing samples, based on patient IDs, using selectable train-test sample ratio
2) Store training and testing samples data-frames

Return: training and testing samples data-frames

FIG. 11

function make_classifier_model() **Inputs:** model name, model parameters

Steps:

1) Select one of the following classification models: random forest, extra-trees, or gradient boosted trees

2) Specify selected model's parameters and instantiate the model

Return: Classification model object

function classification_performance_metrics() **Inputs:** classification model object, cell samples data-frame, M number of cells, N number of cells, R number of iterations

Steps:

for *n-th run in R number of iterations:*

run ***make_classifier_model()***

run ***split_data_train_test()***

1) Select M number of cells per patient ID to be used in measuring classification performance metrics for testing data-frame

2) Select N number of cells per patient ID which need to be classified as cancer to classify the whole patient ID as cancer

3) Using all cells available per patient ID, make all possible combinations of M cells for all patient IDs in the testing samples and store in data-frame

4) Calculate model classification performance metrics: ROC curve, ROC AUC score, accuracy, and confusion matrix

Return: ROC curves, ROC AUC scores, accuracy scores, confusion matrices

FIG. 11 (cont.)

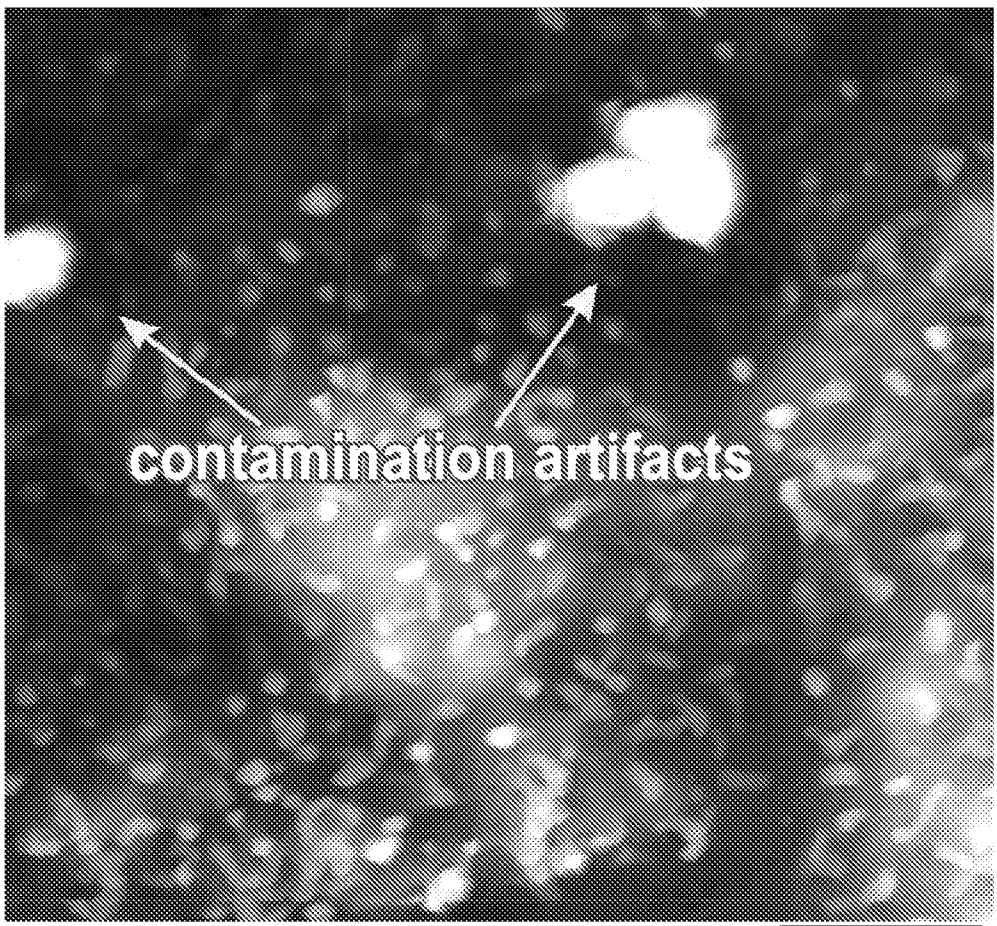


FIG. 12

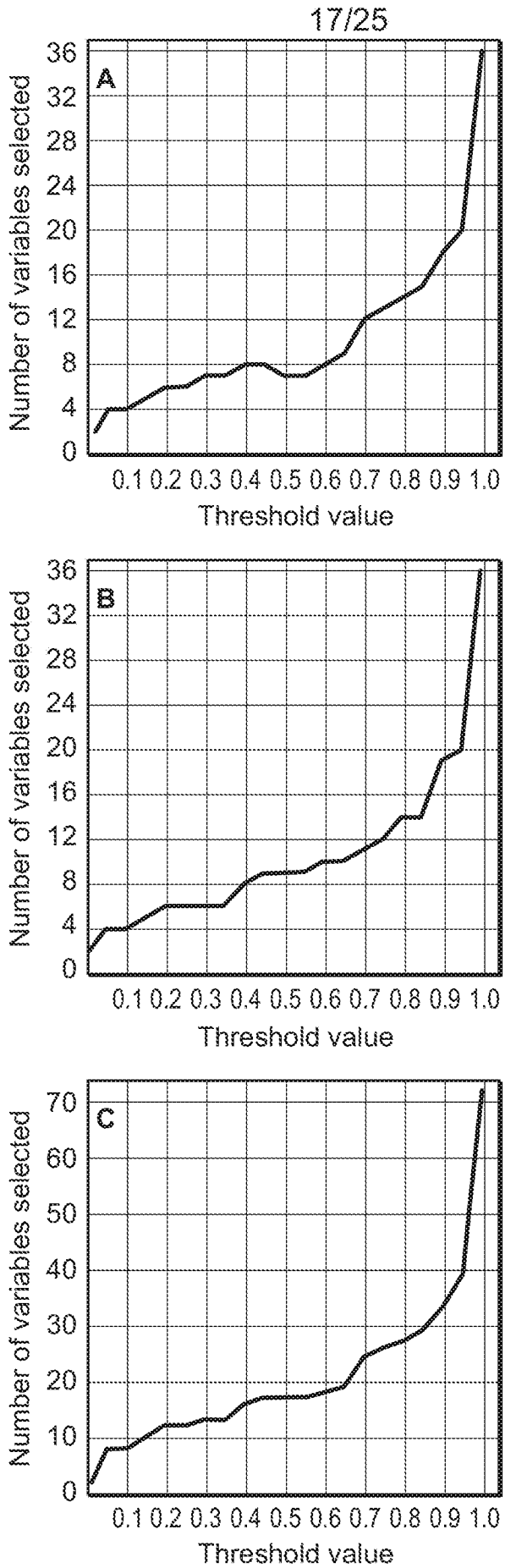
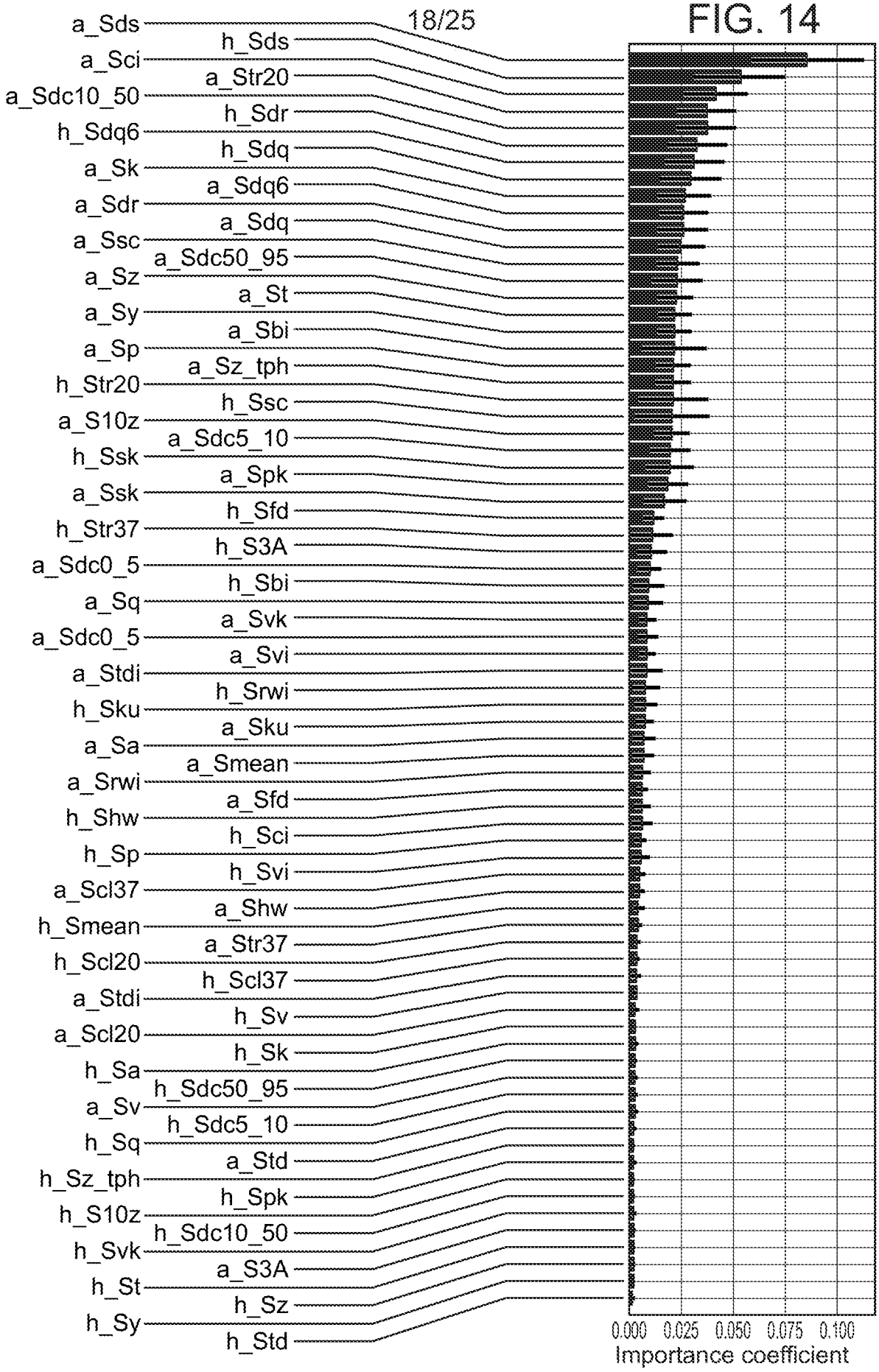

A
Number of variables selected
0
4
8
12
16
20
24
28
32
36
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Threshold value
B
Number of variables selected
0
4
8
12
16
20
24
28
32
36
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Threshold value
C
Number of variables selected
0
10
20
30
40
50
60
70
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Threshold value

FIG. 13

FIG. 14


a_Sds
h_Sds
a_Sci
a_Str20
a_Sdc10_50
h_Sdr
h_Sdq6
h_Sdq
a_Sk
a_Sdq6
a_Sdr
a_Sdq
a_Ssc
a_Sdc50_95
a_Sz
a_St
a_Sy
a_Sbi
a_Sp
a_Sz_tph
h_Str20
h_Ssc
a_S10z
a_Sdc5_10
h_Ssk
a_Spk
a_Ssk
h_Sfd
h_Str37
h_S3A
a_Sdc0_5
h_Sbi
a_Sq
a_Svk
a_Sdc0_5
a_Svi
a_Stdi
h_Srwi
h_Sku
a_Sku
a_Sa
a_Smean
a_Srwi
a_Sfd
h_Shw
h_Sci
h_Sp
h_Svi
a_Scl37
a_Shw
h_Smean
a_Str37
h_Scl20
h_Scl37
a_Stdi
h_Sv
a_Scl20
h_Sk
h_Sa
h_Sdc50_95
a_Sv
h_Sdc5_10
h_Sq
a_Std
h_Sz_tph
h_Spk
h_S10z
h_Sdc10_50
h_Svk
a_S3A
h_St
h_Sz
h_Sy
h_Std
0.000
0.025
0.050
0.075
0.100
Importance coefficient

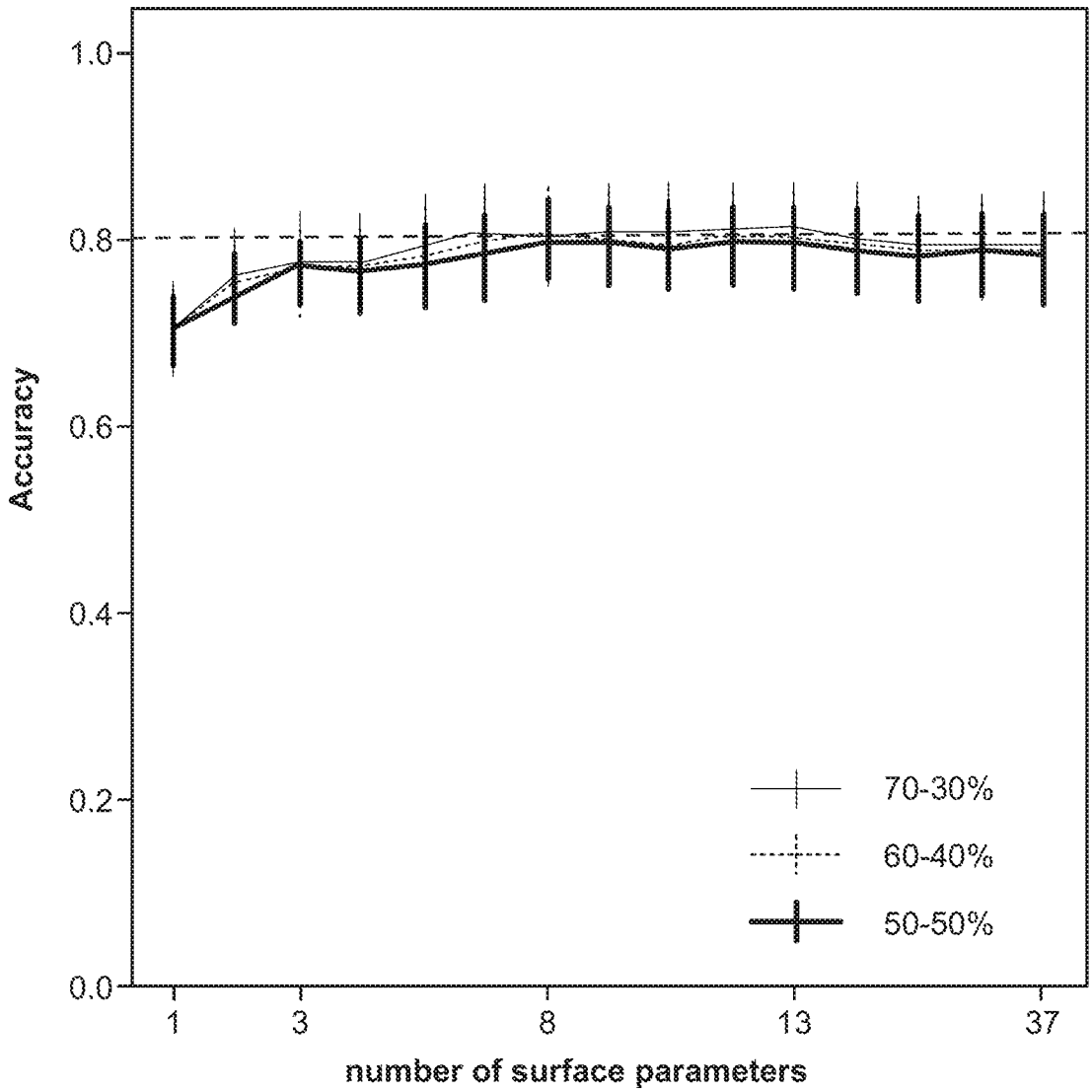

1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
1
3
8
13
37
number of surface parameters
70-30%
60-40%
50-50%

FIG. 15

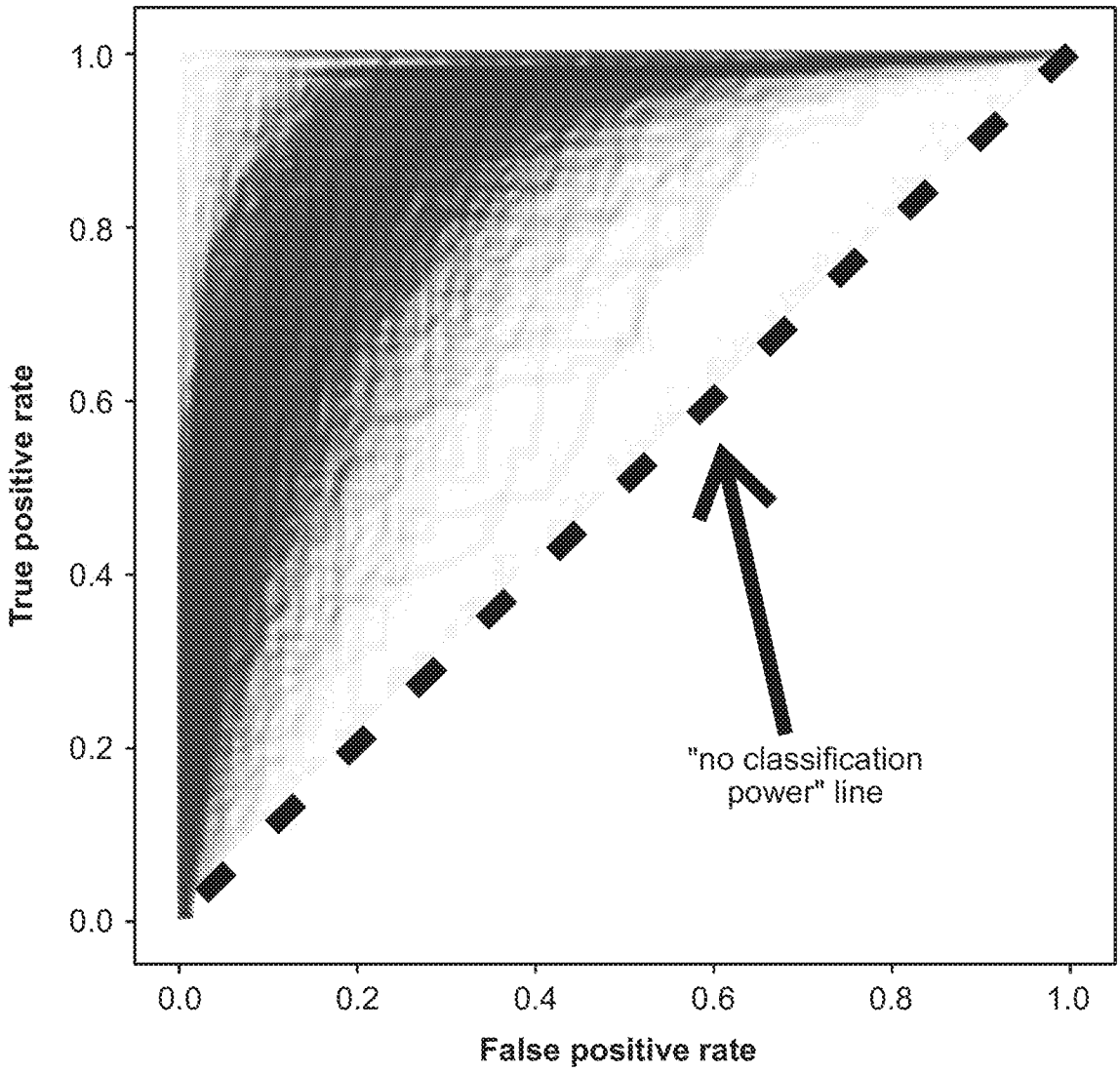

1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
"no classification power" line

FIG. 16

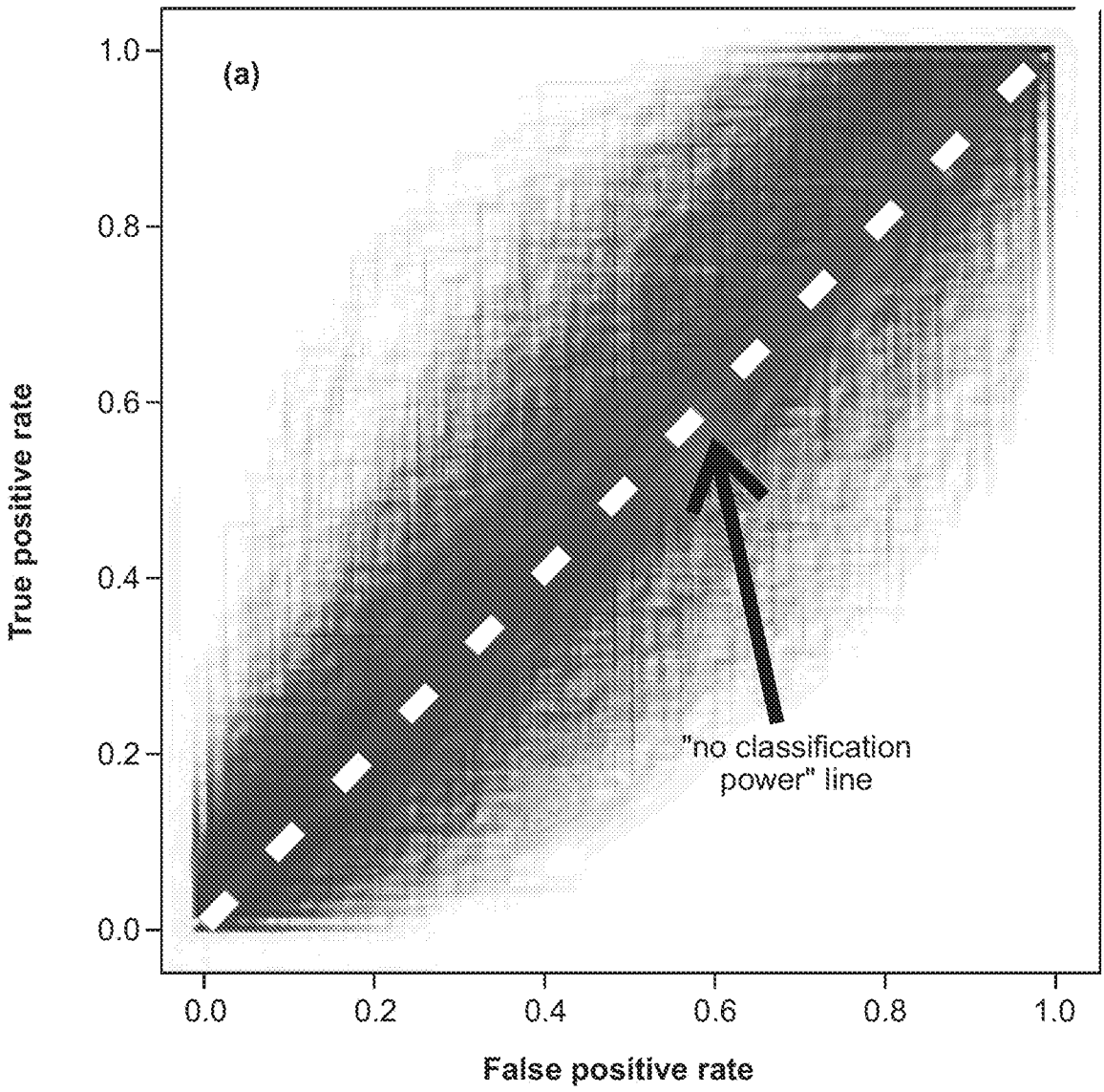

(a)
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
"no classification power" line
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

FIG. 17

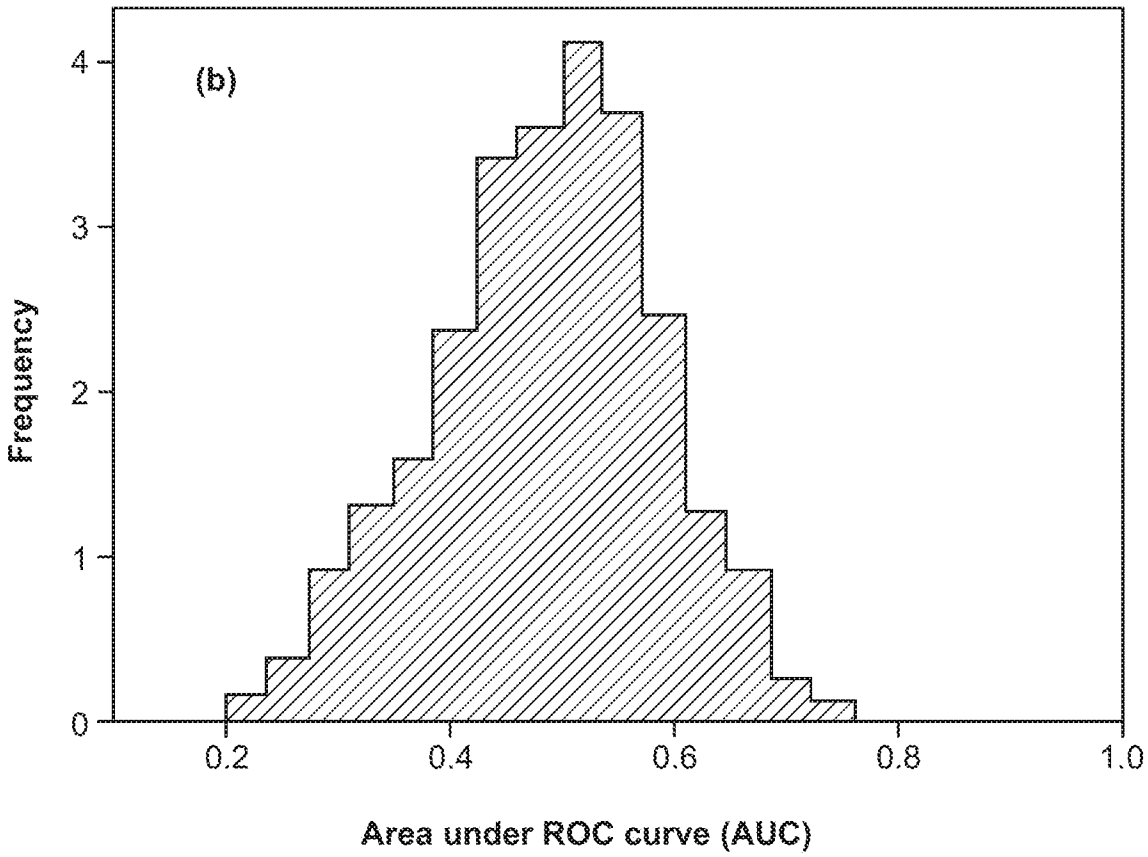


FIG. 18

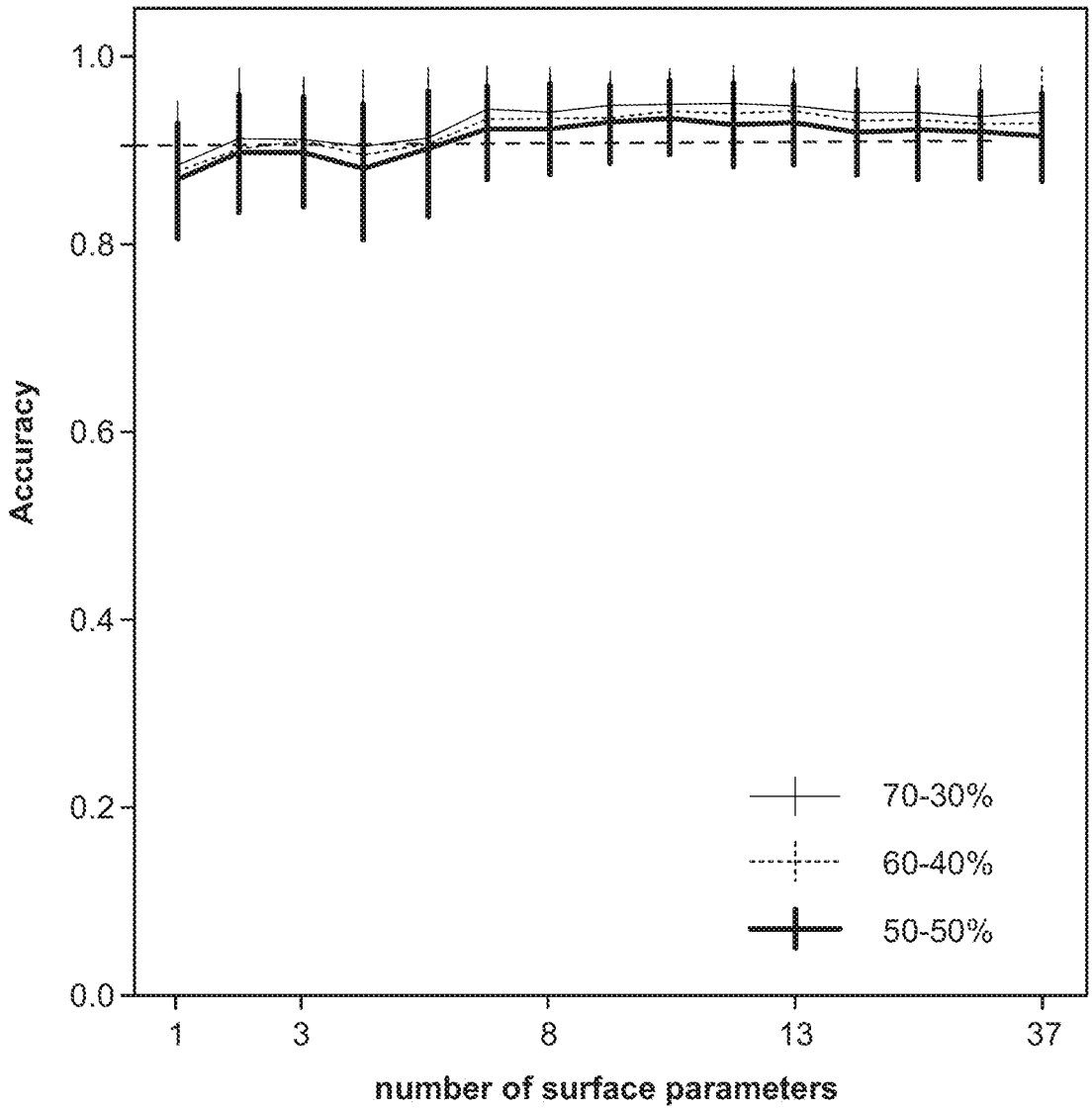

1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
1
3
8
13
37
number of surface parameters
70-30%
60-40%
50-50%

FIG. 19

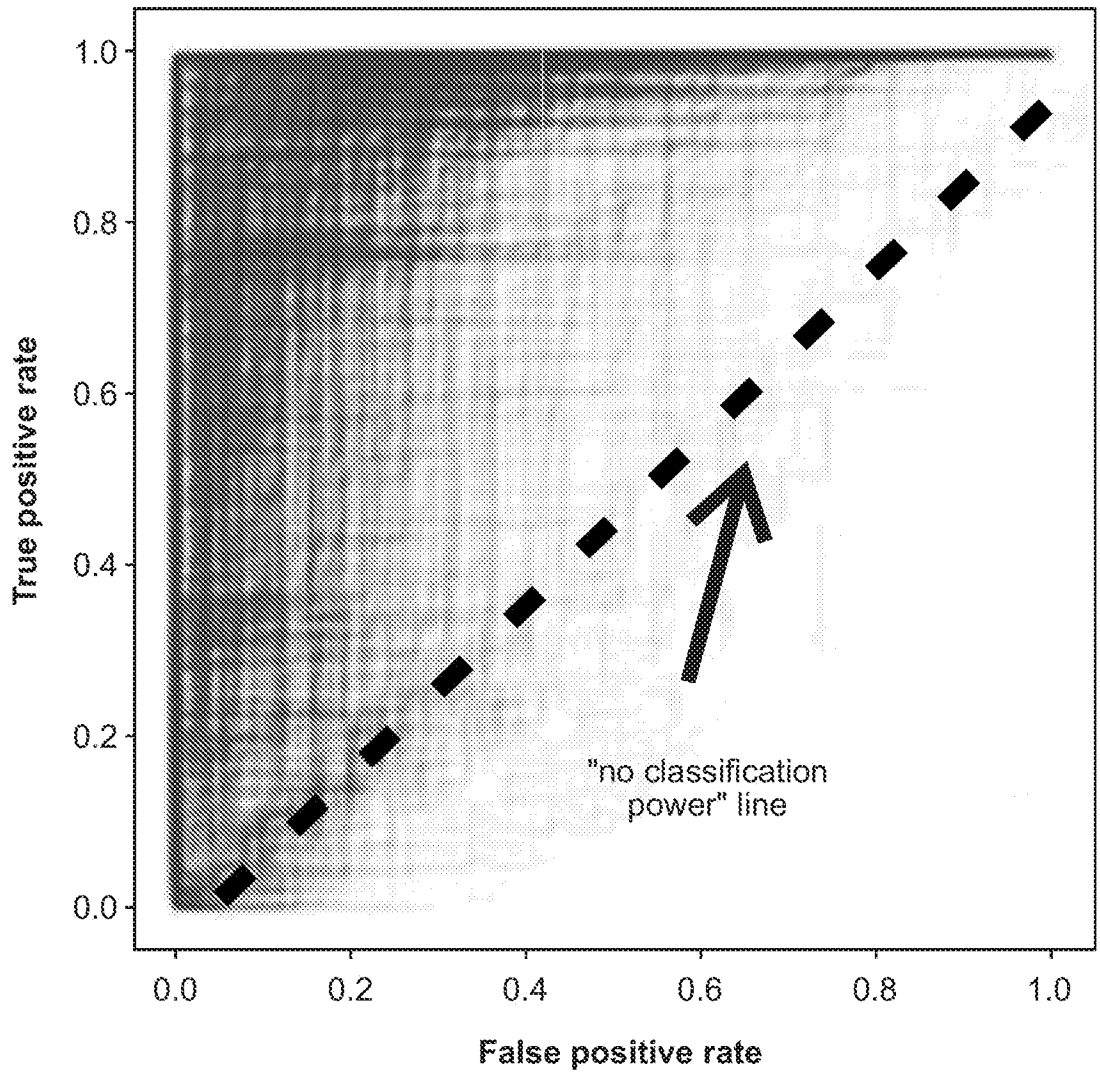

1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
"no classification power" line

FIG. 20

	data	Random forest			Extremely Randomized Forest			Gradient Boosting Trees		
		AUC/ Accuracy	Sense/spec		AUC/ Accuracy	Sense/spec		AUC/ Accuracy	Sense/spec	
N=1	height	75/73	50/84	76/69	77/74	53/84	77/70	74/73	46/86	75/68
M=1	adh.	88/83	68/90	84/77	88/83	69/90	85/78	87/82	69/89	84/77
N=2	height	75/77	40/91	76/69	77/78	44/91	77/71	75/77	40/92	75/69
M=1	adh.	89/87	71/93	86/80	89/86	70/93	87/80	89/87	71/93	86/80
N=2	height	77/78	43/92	76/69	78/79	48/91	78/71	75/78	38/93	74/68
M=2	adh.	89/86	69/93	86/80	90/87	70/94	86/79	89/87	69/94	85/79
N=3	height	73/81	34/95	75/66	74/81	35/95	76/67	73/81	35/95	74/66
M=1	adh.	90/89	70/95	88/81	89/88	69/94	86/79	89/90	70/96	87/80
N=3	height	75/82	36/95	75/68	78/82	42/94	77/69	73/81	34/96	73/66
M=2	adh.	91/90	73/96	87/80	90/90	72/96	87/80	89/90	69/96	85/78
N=3	height	75/82	37/95	76/67	78/83	42/95	77/69	72/81	32/96	73/66
M=3	adh.	90/90	71/96	87/80	90/90	70/96	87/76	88/89	65/96	85/77
N=4	height	69/84	31/97	71/61	70/84	33/97	73/64	69/84	34/97	73/64
M=1	adh.	89/91	75/95	88/80	88/90	71/96	87/79	90/92	74/96	88/80
N=4	height	72/84	35/97	73/64	74/85	37/96	76/67	71/84	37/96	72/63
M=2	adh.	91/92	77/97	88/80	90/92	74/96	88/80	90/92	73/97	87/79
N=4	height	72/84	34/97	74/64	76/85	41/96	76/67	70/84	32/97	71/62
M=3	adh.	89/92	72/97	88/79	90/92	72/97	87/79	87/91	64/97	84/75
N=4	height	72/84	32/97	74/64	75/85	38/96	76/66	70/84	32/97	72/62
M=4	adh.	89/91	71/96	88/79	89/91	71/96	87/78	86/90	64/97	84/75
N=5	height	66/86	33/98	69/57	66/85	36/97	72/60	67/86	35/98	70/59
M=1	adh.	88/93	77/96	87/78	87/92	75/96	87/78	89/94	77/97	88/79
N=5	height	69/86	34/97	72/61	70/86	39/96	76/64	68/86	36/97	70/61
M=2	adh.	91/94	81/98	91/82	90/94	78/97	89/80	91/94	78/98	88/80
N=5	height	68/86	31/98	72/61	72/87	39/97	76/65	67/86	34/97	69/58
M=3	adh.	90/93	77/97	88/78	90/93	77/97	88/78	88/93	70/98	85/75
N=5	height	67/86	34/97	72/60	72/87	39/97	76/64	65/85	30/97	69/57
M=4	adh.	89/93	75/97	87/77	88/93	73/97	88/77	85/92	61/98	84/74
N=5	height	68/86	30/97	72/59	74/87	39/97	76/63	67/85	32/97	71/59
M=5	adh.	88/93	74/97	88/77	88/93	73/97	88/77	84/91	62/97	84/73

FIG. 21

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US19/60225

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G06F 19/20; G06T 7/00, 7/136, 7/143, 7/174, 7/30, 7/32, 7/33, 7/35 (2020.01)

CPC - G01N 1/30, 33/574, 33/57407; G01Q 10/045, 20/04, 60/38; G06F 3/04847; G06K 9/00127, 9/0014, 9/00147; G06T 7/0012, 7/0016

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y --- A	US 2015/0036889 A1 (CRIXLabs, Inc.) 05 February 2015, abstract, Fig. 1, 5, para. [0011], [0014], [0021]-[0032], [0034], [0037], [0040]-[0048]	1, 2, 20, 21, 29-32, 34-37 & 43-46 --- 3-6, 33, 38, 39, 40, 47 --- 7-19, 22-28, 41, 42, 48-50
Y	WO 2018/134377 A1 (Centre National de la Recherche Scientifique, et al.) 16 July 2018, see machine translation; pg. 2 second para., pg. 4 last para., pg. 6 third para., pg. 7 second para., pg. 13 second full para through pg. 14 last full para., pg. 15 last para.	3, 5, 6, 39, 40, 47
Y --- A	US 2015/0160225 A1 (SomaLogic, Inc.) 11 June 2015, para. [0090], [0104]-[0107], [0110], [0111], [0188], [0221], [0226], [0246], [0253]-[0263], [0367]	4, 38 --- 7-19, 22-28, 41, 42, 48-50
Y	US 9,904,866 B1 (Noble, I.) 27 February 2018, col. 9 ln. 46-65	33
A	WO 2017/216123 A1 (Nanolive SA) 21 December 2017, entire document	1-50

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
09 January 2020 (09.01.2020)	24 JAN 2020
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas Telephone No. PCT Helpdesk: 571-272-4300