

SYSTEMS AND METHODS FOR COMPREHENSIVE AND STANDARDIZED IMMUNE SYSTEM PHENOTYPING AND AUTOMATED CELL CLASSIFICATION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the priority benefit of United States Provisional Patent Application Serial No. 63/386,476, filed December 7, 2022, the contents of which are incorporated herein by reference in their entirety.

FIELD

[0002] This disclosure is related to methods and systems for the generation of standardized and comprehensive immune profiles for mammalian (*e.g.*, human) subjects.

BACKGROUND

[0003] Immune profiling, *i.e.*, the analysis of a subject's immune health at the serological or cellular level at a given point in time, can aid in diagnosing immune-related diseases and disorders (*e.g.*, allergies, overreactive immune responses (as in asthma and Crohn's disease (inflammatory bowel disease)), or autoimmune diseases (such as autoimmune polyglandular syndrome and some facets of diabetes; see, for example, "Diseases of the Immune System", National Center for Biotechnology Information (US), in "Genes and Disease" [Internet], 1998-, Bethesda (MD)). Immune profiling can also be used to, *e.g.*, identify an individual's specific response to infectious diseases (*e.g.*, viral, bacterial, fungal, or parasite infections), to monitor the response of patients, *e.g.*, cancer patients, to treatment (see, *e.g.*, Lyons, *et al.* (2017), "Immune Cell Profiling in Cancer: Molecular Approaches to Cell-Specific Identification", Precision Oncology 1, 26), and potentially, to predict healthcare outcomes.

[0004] Conventional methods for generating an immune profile include, *e.g.*, enzyme-linked immunosorbent assays (ELISAs), immunoblotting techniques, and flow cytometry-based techniques comprising the use of panels of fluorescently-labeled antibodies directed to a variety of cell surface receptors and manual gating of the flow cytometry data. However these techniques are often laborious and time-consuming, and are not easily scalable to a level that enables the processing of hundreds or thousands of sample.

[0005] More recently, automated techniques have been applied to flow cytometry-based methods. Several programmatic approaches exist to finding high dimensional clusters of

labeled immune cells and identifying them accordingly using manual expert guidance. Off-the-shelf clustering algorithms are often used for this purpose, including tSNE, FlowSOM, UMAP, *etc.* Other purpose-built algorithms also exist for defining labeled cell clusters in high-dimensional immunespace, *e.g.*, flowType and Phenograph. Additionally, companies like Cytapex and Dotmatics OMIQ produce automated cell gating using similar methods. However, despite the existence of established immune profiling platforms and approaches, the immunophenotyping panels used in various experiments tend to differ, requiring a user to configure existing automated gating methods on a per-experiment basis. As a result, direct comparability between experiments is reduced or eliminated. Hence there is a need for improved methods and systems that are capable of providing standardized and comprehensive immune profiles for subjects of interest (*e.g.*, patients) that will facilitate biomedical research and enable improved healthcare outcomes.

BRIEF SUMMARY

[0006] Disclosed herein are methods and systems for processing a sample, such as a blood sample, and generating a standardized and comprehensive immune profile for the subject. The disclosed methods and systems combine full spectrum flow cytometry (FSFC)-based analysis of cells within the sample with an ensemble machine learning-based approach to filtering and classifying individual immune cells into a plurality of distinct immune cell sub-populations. The key advantages of the disclosed methods and systems are enabled by standardizing the immune profiling platform (including the immunophenotyping panels used for fluorescently-labeling cells) and implementing automated data processing. This in turn enables the processing of highly complex FSFC data to generate translatable immune profiles in clinically-relevant timeframes (*i.e.*, hours).

[0007] Disclosed herein are methods for generating an immune profile for a subject, the methods comprising: contacting at least a first aliquot of a sample from the subject with at least a first immunophenotyping panel to fluorescently-label cells contained within the sample; processing the fluorescently-labeled cells using a full spectrum flow cytometer to generate fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells from the sample; providing at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-

labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

[0008] In some embodiments, the ensemble machine learning model is organized in a cascading hierarchical tree structure comprising a plurality of nodes, and wherein each node comprises an individual machine learning model. In some embodiments, each individual machine learning model comprises one input data set and from one to eight output data sets corresponding to branches of the cascading hierarchical tree structure. In some embodiments, each individual machine learning model comprises a neural network model. In some embodiments, each individual machine learning model comprises a gradient descent boosted tree model. In some embodiments, the plurality of nodes comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 nodes. In some embodiments, a number of individual machine learning models in the ensemble machine learning model is equal to a number of distinct immune cell sub-populations in the plurality of distinct immune cell sub-populations. In some embodiments, the design of the cascading hierarchical tree structure is based at least in part on an expert analysis of manually-gated fluorescence intensity data, or data derived therefrom, for one or more control samples.

[0009] In some embodiments, individual cells are classified independently of all other cells in the plurality of fluorescently-labeled cells. In some embodiments, individual cells are classified recursively with all other cells of the plurality of fluorescently-labeled cells.

[0010] In some embodiments, the ensemble machine learning model is trained using one or more labeled training data sets generated by an expert through manual gating of fluorescence intensity data, or data derived therefrom, for one or more control samples. In some embodiments, the individual machine learning models in the ensemble machine learning model are trained individually using the one or more labeled training data sets. In some embodiments, during training, an individual model's predictions are used to validate the individual model but do not propagate forward through the ensemble machine learning model, thereby eliminating error propagation during training. In some embodiments, the individual machine learning models in the ensemble machine learning model are trained collectively using a recursive training method. In some embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are the same for every node in the cascading hierarchical tree structure. In some

embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are different for a subset of nodes in the cascading hierarchical tree structure. In some embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are determined by performing a random grid search of the value ranges for the one or more hyperparameters.

[0011] In some embodiments, the method further comprises performing a mathematical transformation of the fluorescence intensity data, or data derived therefrom, prior to using the transformed fluorescence intensity data as input for the ensemble machine learning model.

[0012] In some embodiments, the fluorescence intensity data, or data derived therefrom, comprises fluorescence intensity data for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, or 40 fluorescence detection channels. In some embodiments, the fluorescence intensity data, or data derived therefrom, further comprises forward scatter height data, forward scatter area data, side scatter height data, side scatter area data, autofluorescence data, or any combination thereof.

[0013] In some embodiments, the sample comprises a blood sample, a buffy coat sample, or a cell suspension.

[0014] In some embodiments, the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with antigen-presenting cells (APCs). In some embodiments, the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to IGM, CD5, CD62L, CD294, CD69, CD38, PD1, CD11C, CD3, CD8, HLADR, CD24, CD337, CD123, CD141, CD1C, CD4, TACI, CD319, CD335, PDL1, CD10, CD45, CD16, IGD, CD40, CD19_TCRGD, CD43, CD14, CD138, CD15, CD56, CD86, CD303, CD27, or any combination thereof. In some embodiments, the panel of fluorescently-labeled antibodies further comprises fluorescently-labeled antibodies directed to cell surface markers that are indicative of live cells, dead cells, or both.

[0015] In some embodiments, the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with T cells. In some embodiments, the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to TIGIT, CD5, CD28, CXCR5, CD39, TIM3, CD38, PD1, TCRVA7_2_TCRVD1, CD95, CD3, CD8, HLADR, CD31, CCR4, CCR6, CCR7, CD57,

ICOS, CD4, KLRG1, TCRVA24_JA18, CD122, CD103, CXCR3, TCRVD2, CD45, CCR10, CD16, CD25, CD161, CD19_TCRGD, LAG3, CD14, CD45RO, CD56, CD127, CD45RA, CD27, or any combination thereof.

[0016] In some embodiments, the plurality of distinct immune cell sub-populations comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 distinct immune cell sub-populations. In some embodiments, the plurality of distinct immune cell sub-populations includes white blood cells (WBC), Eosinophils, Eosinophil/CD5+, Neutrophils, Neutrophils/big, Neutrophils/CD5+, Neutrophils/small, B-cells, B-cells/CD5- CD27-, Monocytes/CD56+, Monocytes/CD56-, NK-cells, Dendritic cells (DC), T-cells, iNKT cells, gamma delta T-cells (total GD), Vd1 cells, Vd2 cells, Vdx cells, Mucosal-associated invariant T (MAIT) cells, TEMRA cells, CD4 naïve cells, T helper cells, CD4 effector memory cells, Treg cells, or any combination thereof.

[0017] In some embodiments, the total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample is output as part of an immune profile in less than 24 hours, 12 hours, 10 hours, 8 hours, 6 hours, or 4 hours.

[0018] In some embodiments, the immune profile is used to diagnose an immune-related disease or disorder, monitor progression of an immune-related disease or disorder, or monitor a response to treatment of an immune-related disease or disorder in the subject.

[0019] Disclosed herein are computer-implemented methods for generating an immune profile for a subject, the methods comprising: receiving fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from the subject; providing at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

[0020] In some embodiments, the ensemble machine learning model is organized in a cascading hierarchical tree structure comprising a plurality of nodes, and wherein each node comprises an individual machine learning model. In some embodiments, each individual

machine learning model comprises one input data set and from one to eight output data sets corresponding to branches of the cascading hierarchical tree structure. In some embodiments, each individual machine learning model comprises a neural network model. In some embodiments, each individual machine learning model comprises a gradient descent boosted tree model. In some embodiments, the plurality of nodes comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 nodes. In some embodiments, a number of individual machine learning models in the ensemble machine learning model is equal to a number of distinct immune cell sub-populations in the plurality of distinct immune cell sub-populations. In some embodiments, the design of the cascading hierarchical tree structure is based at least in part on an expert analysis of manually-gated fluorescence intensity data, or data derived therefrom, for one or more control samples.

[0021] In some embodiments, individual cells are classified independently of all other cells in the plurality of fluorescently-labeled cells. In some embodiments, individual cells are classified recursively with all other cells of the plurality of fluorescently-labeled cells.

[0022] In some embodiments, the ensemble machine learning model is trained using one or more labeled training data sets generated by an expert through manual gating of fluorescence intensity data, or data derived therefrom, for one or more control samples. In some embodiments, the individual machine learning models in the ensemble machine learning model are trained individually using the one or more labeled training data sets. In some embodiments, during training, an individual model's predictions are used to validate the individual model but do not propagate forward through the ensemble machine learning model, thereby eliminating error propagation during training. In some embodiments, the individual machine learning models in the ensemble machine learning model are trained collectively using a recursive training method. In some embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are the same for every node in the cascading hierarchical tree structure. In some embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are different for a subset of nodes in the cascading hierarchical tree structure. In some embodiments, the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are determined by performing a random grid search of the value ranges for the one or more hyperparameters.

[0023] In some embodiments, the method further comprises performing a mathematical transformation of the fluorescence intensity data, or data derived therefrom, prior to using the transformed fluorescence intensity data as input for the ensemble machine learning model.

[0024] In some embodiments, the fluorescence intensity data, or data derived therefrom, comprises fluorescence intensity data for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, or 40 fluorescence detection channels. In some embodiments, the fluorescence intensity data, or data derived therefrom, further comprises forward scatter height data, forward scatter area data, side scatter height data, side scatter area data, autofluorescence data, or any combination thereof.

[0025] In some embodiments, the plurality of distinct immune cell sub-populations comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 distinct immune cell sub-populations. In some embodiments, the plurality of distinct immune cell sub-populations includes white blood cells (WBC), Eosinophils, Eosinophil/CD5+, Neutrophils, Neutrophils/big, Neutrophils/CD5+, Neutrophils/small, B-cells, B-cells/CD5- CD27-, Monocytes/CD56+, Monocytes/CD56-, NK-cells, Dendritic cells (DC), T-cells, iNKT cells, gamma delta T-cells (total GD), Vd1 cells, Vd2 cells, Vdx cells, Mucosal-associated invariant T (MAIT) cells, TEMRA cells, CD4 naïve cells, T helper cells, CD4 effector memory cells, Treg cells, or any combination thereof.

[0026] In some embodiments, the total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample is output as part of an immune profile in less than 24 hours, 12 hours, 10 hours, 8 hours, 6 hours, or 4 hours.

[0027] In some embodiments, the immune profile is used to diagnose an immune-related disease or disorder, monitor progression of an immune-related disease or disorder, or monitor a response to treatment of an immune-related disease or disorder in the subject.

[0028] Disclosed herein are systems comprising: one or more processors; and a memory communicatively coupled to the one or more processors and configured to store instructions that, when executed by the one or more processors, cause the system to: receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the

fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject. In some embodiments, the system further comprises a full spectrum flow cytometer.

[0029] Disclosed herein are non-transitory computer-readable storage media storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of a system, cause the system to: receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

INCORPORATION BY REFERENCE

[0030] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference in their entirety to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference in its entirety. In the event of a conflict between a term herein and a term in an incorporated reference, the term herein controls.

BRIEF DESCRIPTION OF THE DRAWINGS

[0031] Various aspects of the disclosed methods, devices, and systems are set forth with particularity in the appended claims. A better understanding of the features and advantages of the disclosed methods, devices, and systems will be obtained by reference to the following detailed description of illustrative embodiments and the accompanying drawings, of which:

[0032] **FIG. 1** provides a non-limiting example of a process flow diagram for a method of processing blood sample and generating an immune profile, according to one embodiment as described herein.

[0033] **FIG. 2** provides a non-limiting example of a process flow diagram for a method of training an ensemble machine learning (ML) model, according to one embodiment as described herein.

[0034] **FIG. 3** provides an exemplary illustration of machine learning model-based prediction and immune profile generation, according to one embodiment of the methods and systems described herein.

[0035] **FIG. 4** illustrates an exemplary computing system, in accordance with some embodiments of the methods and systems described herein.

[0036] **FIG. 5** provides a non-limiting schematic illustration of a manual gating process for processing full spectrum flow cytometry data.

[0037] **FIG. 6** provides a non-limiting example of a simplified gating hierarchy used to build a neural network for immune cell classification.

[0038] **FIG. 7** provides a simplified schematic illustration of a neural network.

[0039] **FIG. 8** provides a non-limiting example of test data generated by a trained neural network classifier.

[0040] **FIG. 9** provides another non-limiting example of test data generated by a trained neural network classifier.

[0041] **FIG. 10** provides a non-limiting example of validation data generated by a trained neural network classifier.

[0042] **FIG. 11** provides another non-limiting example of validation data generated by a trained neural network classifier.

[0043] **FIG. 12** provides a non-limiting example of the data flow through cascading predictions from node to node, *i.e.*, from parent ML model output to child/children ML model inputs in an ensemble machine learning model.

[0044] **FIG. 13A** provides a non-limiting example of data for T helper 17 (T17) cell frequency versus Vitamin D level for different age groups accounting for sex as a covariate.

[0045] FIG. 13B provides a non-limiting example of data for T helper 2 (T2) cell frequency versus Vitamin D level for different age groups accounting for sex as a covariate.

DETAILED DESCRIPTION

[0046] Disclosed herein are methods and systems for processing a sample, such as a blood sample, and generating a standardized and comprehensive immune profile for the subject. The disclosed methods and systems combine full spectrum flow cytometry (FSFC)-based analysis of cells within the sample with an ensemble machine learning-based approach to filtering and classifying individual immune cells into a plurality of distinct immune cell sub-populations. The key advantages of the disclosed methods and systems are enabled by standardizing the immune profiling platform (including the immunophenotyping panels used for fluorescently-labeling cells) and implementing automated data processing. This in turn enables the processing of highly complex FSFC data to generate translatable immune profiles in clinically-relevant timeframes (*i.e.*, hours).

[0047] In some instances, for example, methods for generating an immune profile for a subject are described, the methods comprising: contacting at least a first aliquot of a sample from the subject with at least a first immunophenotyping panel to fluorescently-label cells contained within the sample; processing the fluorescently-labeled cells using a full spectrum flow cytometer to generate fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells from the sample; providing at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

[0048] Also described are systems comprising: one or more processors; and a memory communicatively coupled to the one or more processors and configured to store instructions that, when executed by the one or more processors, cause the system to: receive fluorescence intensity data, or data derived therefrom, acquired using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset

of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

[0049] Also described are non-transitory computer-readable storage media storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of a system, cause the system to: receive fluorescence intensity data, or data derived therefrom, acquired using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

Definitions

[0050] Unless otherwise defined, all of the technical terms used herein have the same meaning as commonly understood by one of ordinary skill in the art in the field to which this disclosure belongs.

[0051] As used in this specification and the appended claims, the singular forms “a”, “an”, and “the” include plural references unless the context clearly indicates otherwise. Any reference to “or” herein is intended to encompass “and/or” unless otherwise stated, and encompasses any and all possible combinations of one or more of the associated listed items.

[0052] As used herein, the terms “includes,” “including,” “comprises,” and/or “comprising” specify the presence of stated features, integers, steps, operations, elements, components, and/or units but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, units, and/or groups thereof.

[0053] Throughout this application, various parameter values may be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity, and should not be construed as an inflexible limitation on the scope of the disclosure. Accordingly, the description of a range should be considered to have specifically disclosed all possible subranges as well as individual numerical values within that range, irrespective of whether a specific numerical value or specific sub-range is expressly stated. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6, *etc.*, as well as individual numbers within that range, for example, 1, 1.4, 2, 3, 3.6, 4, 5, 5.8, and 6. This applies regardless of the breadth of the range.

[0054] Numbers may be expressed herein as being “about” a particular value. Similarly, ranges may be expressed herein as from “about” one particular value and/or to “about” another particular value. The terms “about” and “approximately” shall generally mean an acceptable degree of error or variation for a given value or range of values, such as, for example, a degree of error or variation that is within 20 percent (%), within 15%, within 10%, or within 5% of a given value or range of values.

[0055] It should be recognized that use of ordinal terms such as “first” and “second” in the description of methods and systems disclosed herein does not by itself connote any priority, order of importance of one system component over another, or the temporal order in which acts of a method are performed, but are used merely as labels to distinguish, for example, one system component having a certain name from another system component having the same name but for the use of the ordinal term to distinguish the two system components.

[0056] Additionally, various implementations of the methods and systems set forth herein may be described in terms of exemplary block diagrams, process flow charts, and other illustrations. As will become apparent to one of ordinary skill in the art after reading this document, the various implementations set forth herein can be implemented without confinement to the illustrated examples. For example, block diagrams and their accompanying description should not be construed as mandating a particular architecture or configuration. Similarly, in exemplary process flow charts, some blocks are optionally combined, the order of some blocks is optionally changed, and some blocks are optionally omitted. In some implementations, additional steps may be performed in combination with the exemplary processes. Accordingly, the methods and systems as described and illustrated

in greater detail below are exemplary by nature and, as such, should not be viewed as limiting.

[0057] As used herein, the terms “full spectrum flow cytometry” and “full spectrum flow cytometer” refer to a technique and instrument, respectively, for performing flow cytometry where the instrument is configured to capture the full emission spectrum of fluorescent molecules using arrays of highly sensitive light detectors, thereby enabling the capture of highly multiplexed fluorescence intensity data sets.

[0058] As used herein, the term “immunophenotyping panel” refers to a panel of antibodies (*e.g.*, fluorescently-labeled antibodies) that are used to identify cells based on the types of antigens or markers (*e.g.*, cell surface receptor proteins) present on the surface of the cells.

[0059] The section headings used herein are for organizational purposes only and are not to be construed as limiting the subject matter described. The description is presented to enable one of ordinary skill in the art to make and use the invention, and is provided in the context of a patent application and its requirements.

Methods for Immune System Phenotyping and Automated Cell Classification

[0060] As described above, certain programmatic approaches for finding high dimensional clusters of labeled immune cells and purpose-built algorithms for defining high-dimensional clusters in immunespace exist. However, these approaches and algorithms have many shortcomings as described, for example, in the several FlowCAP challenges publications (see, *e.g.*, Aghaeepour, *et al.* (2013), “Critical Assessment of Automated Flow Cytometry Data Analysis Techniques”, *Nature Methods* 10(3):228-239).

[0061] As further described above, certain automated cell gating approaches for flow cytometry applications are known. However, these prior approaches rely on identification of patterns in the aggregate flow cytometry event distributions and on clustering-first approaches, and do not leverage an ensemble of machine-learning models that classify individual cell identities.

[0062] Overall, current immunophenotyping platforms (including high-throughput approaches) have several limitation, including the following two issues: (1) phenotyping platforms are constructed for a specific purpose and are by design different enough so as to

confound data comparison between different experiments and across different immunophenotyping panels, and (2) the manual labor required to establish counts and/or frequencies of immune cells in the distinct immune cell sub-populations identified by an immunophenotyping panel is prohibitive to scaling the analysis to more than a few thousand samples. For existing platforms and approaches, the immunophenotyping panels used in different experiments are not comparable, and existing automated gating methods must be configured on a per-experiment basis that is only applicable for that experiment. The results of the immune profiling experiment are thus not generalizable

[0063] Accordingly, there is a need for improved immunophenotyping platforms and methods.

[0064] Disclosed herein are immunophenotyping platforms (also referred to as immune profiling platforms) and methods that may address one or more of the above-identified shortcomings and needs. The immunophenotyping methods disclosed herein comprise performing high throughput, full spectrum flow cytometry in the context of an immunophenotyping processing pipeline comprising several stages: sample processing, data generation, raw data analysis, and results analysis. The disclosed immunophenotyping platforms and methods, combined with the novel application of machine learning algorithms as disclosed herein, may provide an expandable, high-throughput, and automated method that addresses the above-identified shortcomings and needs.

[0065] The immunophenotyping platforms disclosed herein provide the ability to process a blood sample from an individual (or any other single-cell suspension of immune cells extracted from a tissue sample) and generate an immune signature for the individual using full spectrum flow cytometry. Although described primarily in the context of immune profiling, the disclosed immunophenotyping platforms and methods can also be used more generally to generate cell type profiles based on FSFC analysis of any blood sample (or other single cell suspension) for which a suitable panel of fluorescently-labeled antibodies directed to an appropriate set of distinguishing cell surface antigens can be assembled.

[0066] In some instances, whole blood samples may be processed using FSFC to generate a fluorescence profile for each cell (*e.g.*, each immune cell) within the sample. In some instances, blood samples may be processed to extract immune cells, and the immune cells may then be processed using FSFC to generation a fluorescent profile for each immune cell

extracted from the sample. These profiles may then be used to generate an immune profile for the individual. This can be a highly standardized process that produces directly comparable results for analyses performed in different labs or at different times. In some instances, these results are used to train the machine learning models used to predict what category of immune cell type (*i.e.*, to which distinct immune cell sub-population) each cell detected belongs. Together, the disclosed immunophenotyping platform and machine learning training and prediction frameworks enable the generation of clinically-relevant reference ranges for immune cell subtypes and support diagnostic decisions by leveraging the standardization of sample processing and machine learning-based automation of flow cytometry data processing in a scalable and less biased approach.

[0067] The immunophenotyping platforms and methods disclosed herein differ from existing platforms and methods in that other platforms do not provide comprehensive immune system level cell phenotyping at high sample processing throughput. Existing platforms are focused on a specific solution to a specific scientific question. The methods disclosed herein allow an integrated, automated, and standardized approach to producing high resolution immune profiles from blood samples at unprecedented speed and scale. The standardization and high sample throughput nature of the platform (*e.g.*, up to 100, 150, 200, or 250 samples processed per FSFC instrument per 8 hour work day) enables establishing biological reference ranges for cell types or subtypes (*e.g.*, immune cell subtypes) for human populations which will lead to better patient stratification and improved clinical diagnostic applications.

[0068] The platform and methods disclosed herein leverage full spectrum flow cytometry (FSFC) coupled with machine learning algorithms to analyze cells labeled using a standardized set of immune system status antibody panels. This represents the first comprehensive and standardized immunophenotyping platform with automatic cell classification. Firstly, all biological samples processed by the FSFC instrument are treated using a highly standardized and tightly controlled protocol that is subject to quality control (QC) standards. Secondly, the training and application of bespoke machine learning models to analyze the resulting flow cytometry data enables generation of immune profiles for blood samples on a timeline measured in hours.

[0069] Previous attempts at classifying immune cell identity are based on a clustering approach that is similar to that used in manual gating. The ML-based data analysis pipeline disclosed herein differs from previous approaches in that it examines each cell's fluorescent

signature and classifies it based on an ensemble of gradient boosting machine learning models. Hence, identification of immune cell clusters and their hierarchical positioning are a byproduct of individual cell identification and classification by application of the trained ensemble model that comprises several thousand individual machine learning models, and are not based on identification of clusters first, and then measuring the correspondence of individual cells to the identified clusters.

[0070] FIG. 1 provides a non-limiting example of a process flow diagram for an immune profiling method 100 as disclosed herein. As shown, blood samples (or in some instances, buffy coat samples, peripheral blood mononuclear cell (PBMC) samples, or cell suspensions, *etc.*) are received at the laboratory facility, the sample is prepared 102 (*e.g.*, comprising performing one or more of a dilution step, a centrifugation step, a staining step (using one or more fluorescently-labeled antibody panels) and/or a wash step) and analyzed on the full spectrum flow cytometer (FSFC) 104 resulting in the creation of a flow cytometry standard FCS file (comprising flow cytometry data) for the sample that is stored in a database 106. The FCS file may comprise, for example, fluorescence intensity data for one or more fluorescence detection channels (*e.g.*, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 40, 50, or more than 50 fluorescence detection channels), as well as data derived therefrom (*e.g.*, forward scatter height data, forward scatter area data, side scatter height data, side scatter area data, autofluorescence data, or any combination thereof). In some instances, the number of fluorescence detection channels available may be determined by, for example, a combination of the detection hardware available as part of the flow cytometry instrument (*e.g.*, comprising 5, 10, 20, 25, 50, 75, 100, 125, 150, 175, 200, or more than 200 detectors) and the number of spectrally-distinct fluorophores (*e.g.*, 5, 10, 20, 25, 30, 35, 40, 45, 50, 60, or more than 60 spectrally-distinct fluorophores).

[0071] During a training phase, the data for several FCS files may be manually gated (*e.g.*, by an immunologist or other expert) to generate labeled training data sets for one or more samples (*e.g.*, control samples), and machine learning models (*e.g.*, the individual models in an ensemble machine learning model) are trained 108 using the one or more labeled training data sets. These training sets are full instances of the gating hierarchy implemented on the FSFC output for a sample (*e.g.*, a blood sample). The output of the gating process is a set of industry standard FlowJo workspace files describing the immunophenotyping classification hierarchy associated with their respective samples. The gating data encoded in these files is

then extracted and fed through a cascading ensemble machine learning hierarchy which follows the same gating procedure. Within the hierarchy, each node represents an execution pipeline which trains and persists a singular ML model that predicts events specific to the corresponding biological cell type for that position in the hierarchy. These models are trained using the designated fluorescent channel data and gate location parameters provided by domain experts.

[0072] In some instances, the disclosed immunophenotyping and automated data analysis platform may use a single machine learning model to classify all cells within a cell populations. This can work well for some cell populations, but for populations where the true number of positive classes (cell subtypes) is small (*e.g.*, less than about 100), the class imbalance can overwhelm the ability of a single ML model to reliably classify cell detection events. Weighting the smaller class detection events produces extremes in the case of misclassification (especially apparent as the class imbalance grows) that cause the model performance to degrade. In some instances, gradient boosted machine learning approaches (*e.g.*, gradient boosted ensemble models) may provide superior performance.

[0073] The trained machine learning models 110 are then used to produce predictions of cell type or subtype (*e.g.*, immune cell sub-population) for individual cell detection events and to determine cell counts (or frequencies) for each of a plurality of distinct cell types or subtypes which are then combined into an immune profile 112.

[0074] As noted above, processing of samples (*e.g.*, blood samples) 102 may comprise one or more steps, including a staining step that comprises contacting cells within at least a first aliquot of the sample with at least a first panel (*i.e.*, an immunophenotyping panel or FSFC panel) of fluorescently-labeled antibodies directed to a set of specific cell surface antigens (*e.g.*, cell surface proteins) that collectively enable discrimination between the cell types or cell subtypes of interest. Sample processing may also include immunophenotyping panel design. A sample processing platform may comprise contacting each of one or more sample aliquots (*e.g.*, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more than 10 sample aliquots) with one or more FSFC panels (*e.g.*, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or more than 10 FSFC panels).

[0075] For example, in some instances, each of two sample aliquots may be stained with a different FSFC panel, one focusing on the antigen-presenting cell (APC) arm of the immune system (A panel), which comprises antibodies directed to 36 different cell surface proteins,

and the other focusing on the adaptive arm of the immune system (T panel), which comprises antibodies directed to 41 different cell surface proteins. In some instances, the panels may also include cell viability staining to distinguish between live cells and dead cells. In some instances, the panels may also comprise an autofluorescence measurement as a “marker”. Specific examples of the cell surface proteins and additional markers that may be included in these panels are listed in **Table 1**.

Table 1. Non-limiting examples of cell surface receptor proteins and other markers for distinguishing between immune cell sub-populations.

APC panel markers (A panel)	T panel markers (T panel)
IGM, LIVE_DEAD_BLUE, CD5, CD62L, CD294, CD69, CD38, PD1, CD11C, CD3, CD8, HLADR, CD24, CD337, CD123, CD141, Autofluorescence 1, CD1C, CD4, TACI, CD319, CD335, PDL1, CD10, CD45, CD16, IGD, CD40, CD19_TCRGD, CD43, CD14, CD138, CD15, CD56, CD86, CD303, CD27	TIGIT, CD5, CD28, CXCR5, CD39, TIM3, CD38, PD1, TCRVA7_2_TCRVD1, CD95, CD3, CD8, HLADR, CD31, CCR4, CCR6, CCR7, Autofluorescence 1, CD57, ICOS, CD4, KLRG1, TCRVA24_JA18, CD122, CD103, CXCR3, TCRVD2, CD45, CCR10, CD16, CD25, CD161, CD19_TCRGD, LAG3, CD14, CD45RO, CD56, CD127, CD45RA, CD27

[0076] These panels are custom to the immunophenotyping platform disclosed herein, and are designed to be a maximally comprehensive view of the state of the sample donor’s immune system. The panels include markers for determining immune cell type, immune system activation, lineage (*e.g.*, the main marker(s) that are commonly used to define a certain cell population prior to further subsetting the cell type; examples include, but are not limited to, CD3 to define total T cells, and CD56 and CD16 to define natural killer cells), and exhaustion (cells that express markers associated with “cell exhaustion” (*e.g.*, PD-1, TIGIT) can no longer proliferate and lose their functionalities as a result of chronic stimulation/prolonged activation of immune response). The immunophenotyping platform also defines a common hierarchy (*a.k.a.*, a gating tree) that is used to process the fluorescence profile data for each detected cell and determine which and how many cells belong to each measured population (*e.g.*, immune cell sub-population) in the hierarchy. This comprises

200+ gates for the APC panel and 2000+ gates for the T cell panel for the current configuration of the platform.

[0077] As part of panel design, it can be advantageous to define the maximum number of different fluorescent dyes that can be used within a single panel (which is related to the number of different fluorescence detection channels available in the FSFC instrument) to maximally cover the available spectrum whilst also providing clearly distinguishable signals to distinguish cell surface markers from each other. Multiple iterations may be leveraged to determine panel design parameters, and biological constraints may be leveraged to multiplex dye usage (*e.g.*, Gamma Delta (GD) T cell receptors (TCR GD) may not be expressed on B cells, therefore the same fluorophore conjugated to anti-CD19 and also anti-TCR GD antibodies may be used, for identifying B cells and GD T cells respectively).

[0078] Furthermore, for panel design, it can be advantageous to minimize the sample (*e.g.*, blood) volume required to process a sample. In some instances, the process described herein uses whole blood directly instead of isolating peripheral blood mononuclear cells. This may allow for the determination of granulocyte counts and frequencies.

[0079] Referring back to **FIG. 1**, data generation may refer to the largely automated data generation pipeline that comprises performing full spectrum flow cytometry 104 on the prepared blood sample to produce raw immunofluorescence intensities for each detected cell (*e.g.*, immune cell), and register the output into database 106. In some instances, *e.g.*, where two immunophenotyping panels are used to stain aliquots of the blood sample (such as the A panel and T panel described above), this process produces two data sets, *e.g.*, a data set corresponding to each of the two immunophenotyping panels. In some instances, automated data transformation (*e.g.*, a mathematical transformation of the fluorescence intensity data, or data derived therefrom, to minimize the impact of outlier data points) may be triggered by FCS file deposition onto the FSFC instrument hard drive. The transformed data may then be used as input for a machine learning model configured to classify individual cells according to cell type or sub-type.

[0080] Any of a variety of supervised machine learning models may be used to implement the methods and systems described herein. Examples include, but are not limited to, neural networks (*e.g.*, deep neural networks), decision trees, and random forests.

[0081] In some instances, the disclosed methods and systems may be implemented using an ensemble machine learning model comprising a plurality of individual machine learning models. For example, in some instances, the disclosed methods and systems may be implemented using an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations.

[0082] In some instances, the ensemble machine learning model may be organized in, for example, a cascading hierarchical tree structure comprising a plurality of nodes, where each node comprises an individual machine learning model. In some instances, each individual machine learning model comprises one input data set and up to eight output data sets (*e.g.*, 1, 2, 3, 4, 5, 6, 7, or 8 output data sets) corresponding to branches of the cascading hierarchical tree structure. In some instances, each individual machine learning model may comprise a neural network model. In some instances, each individual machine learning model may comprise a gradient descent boosted tree model. In some instances, the plurality of nodes (*i.e.*, the number of individual machine learning models in the ensemble) comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 nodes. In some instances, the number of individual machine learning models in the ensemble machine learning model is equal to a number of distinct immune cell sub-populations in the plurality of distinct immune cell sub-populations.

[0083] In some instances, the design of the cascading hierarchical tree structure may be based at least in part on an expert analysis of manually-gated fluorescence intensity data, or data derived therefrom, for one or more samples, *e.g.*, control samples as described elsewhere herein. In some instances, individual cells may be classified by the machine learning model (*e.g.*, an ensemble machine learning model) independently of all other cells in the plurality of fluorescently-labeled cells contained within a sample. In some instances, individual cells may be classified recursively with all other cells of the plurality of fluorescently-labeled cells within a sample.

[0084] As noted above, the machine learning models (*e.g.*, the individual models in an ensemble machine learning model) are trained during a training phase 108 using labeled training data sets generated using one or more FCS files for one or more samples (*e.g.*, control samples) that are manually gated, *e.g.*, by an immunologist or other expert. In some

instances, the one or more samples may comprise whole blood samples. In some instances, the one or more control samples may comprise, *e.g.*, cell suspensions comprising purified or partially purified cells of a single or only a few cell types or subtypes. Training and predictions may be periodically updated when advances in model architecture are made, more appropriate hyperparameters are identified, or new data is available.

[0085] In some instances, the individual machine learning models in the ensemble machine learning model may be trained individually using one or more labeled training data sets. For example, during training, an individual model's predictions may be used to validate the individual model but may not be propagated forward through the ensemble machine learning model, in order to minimize or eliminate error propagation during the training process.

[0086] In some instances, the individual machine learning models in the ensemble machine learning model may be trained collectively, *e.g.*, using a recursive training method.

[0087] In some instances, the training of the ensemble machine learning model (or the individual models contained therein) may be controlled by one or more hyperparameter values that are the same for every node (individual model) in the cascading hierarchical tree structure. In some instances, the training of the ensemble machine learning model may be controlled by one or more hyperparameter values that are different for every node (individual model), or different for a subset of the nodes, in the cascading hierarchical tree structure. In some instances, the training of the ensemble machine learning model may be controlled by one or more hyperparameter values that are determined by performing a random grid search of the value ranges for the one or more hyperparameters.

[0088] Referring back to **FIG. 1**, data analysis may refer to the automated application of a trained machine learning model (*e.g.*, an ensemble machine learning model comprising 2200+ individual machine learning models) 110 to the raw data files generated by the FSFC instrument and stored in database 106. The cascading hierarchy of the ensemble machine learning model processes all or a portion of the fluorescent intensity readings (or data derived therefrom) from each cell and predicts the identity of the cell (*i.e.*, the cell type or sub-type (also referred to herein as the cell sub-population) to which it belongs) as determined by the gating hierarchy defined by the immunophenotyping panel design.

[0089] Each ML model's input may use the same set of cell surface markers that an expert would use to analyze the data at that level of the hierarchy. For instance, to determine events

that are identified as Neutrophils, an expert would look at a 2-dimensional plot with side scatter area on one axis and the fluorescent intensity of the CD16 marker on the second axis. The machine learning model at that node in the hierarchy may use the same two inputs. These inputs represent the encoded values of a proxy of the size and the amount of the marker (measured as the intensity). These input channels are optionally selected to mimic as closely as possible what the expert uses without including noise from the other channels which can overwhelm the signal provided. Biological expertise guides the selection of each channel for each immune population subset in the ML ensemble. Each ML model's input may also be the complete set of fluorescent channels available in the panel.

[0090] The ensemble ML models work together to produce accurate predictions of cell types for the entire gating hierarchy. During prediction, the ML models are arranged programmatically so that the top level model forwards its predictions to the next layer of models, *e.g.* the white blood cell (WBC) model uses two proximate size parameters (side scatter area vs. side scatter height) to determine if a cell detection event is likely due to the cell being a WBC. All of the events that are classified as WBC events may then be further predicted using one or two ML models and classified into side scatter high (SSChi) and side scatter low (SSClo) event types using the forward scatter (FSC) and side scatter (SSC) channels. Cell detection events that are predicted as SSChi are then further characterized to determine if they are positive for expression of the CD15 cell surface marker. Each step of this process is performed by a specific ML model trained for this explicit purpose. This process is followed for each node in the gating hierarchy.

[0091] The machine learning platform 110 produces cell type or subtype predictions at each leaf (*i.e.*, final node having no child nodes) in the defined gating hierarchy independently using one or multiple machine learning models, which may have the same or different architectures, working together in an ensemble. The predictions from these models can either be recursively cascaded through the hierarchy from more general to more specific (*e.g.*, from parent to child) or each cell may be classified independently of all other cells and all other gating hierarchy decisions.

[0092] The trained machine learning model(s) may be configured to classify individual cells as belonging to one of a plurality of distinct cell sub-populations, *e.g.*, immune cell sub-populations. In some instances, the plurality of distinct cell sub-populations (*e.g.*, distinct

immune cell sub-populations) may comprise at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 distinct immune cell sub-populations.

[0093] Non-limiting examples of immune cell sub-populations (or sub-types) that may be identified using, *e.g.*, the A panel and T panel sets of fluorescently-tagged antibodies described above are illustrated in **Table 2**.

Table 2. Non-limiting examples of immune cell sub-populations identified using A panel and T panel antibodies.

APC panel (A panel) populations	T panel populations
WBC Eosinophil Eosinophil/CD5+ Neutrophil Neutrophil/Big Neutrophil/CD5+ Neutrophil/small B-cell B-cell/CD5- CD27- Monocyte/CD56+ Monocyte/CD56- NK-cell DC (and 180+ additional sub-populations)	T-cell iNKT Total GD Vd1 Vd2 Vdx MAIT TEMRA CD4 NAÏVE T_HELPER CD4 Effector Memory Treg (and 2000+ additional sub-populations)

[0094] **FIG. 2** provides a non-limiting example of a flow diagram of a machine learning training process 200. Domain experts may produce and store manually labeled training data 206 from selected FCS files. Machine learning specialists may define machine learning architecture, training algorithms, and hyperparameters 202 to generate ML scripts. The ML scripts may be then executed to train the machine learning model(s) 204 on a selection of manually labeled data 206, and subsequently stored in database 208. The output models may be validated using a hold-out data set 212 that has also been manually labeled by an expert. The results of the validation may be manually reviewed by data science experts and/or biologists 210. Models failing validation may be tuned and retrained.

[0095] The training data used to train the machine learning models may be manually produced using a subset of expertly gated exemplar immunophenotyping outputs. These manually gated cells may be used to train the machine learning model to be able to identify

similar cells based on the fluorescence profile characterized by the spectral flow cytometer for a given cell type or subtype.

[0096] In some instances, the use of training data generated by multiple experts may remove bias from the gating process, so replacing the automated gating process with manual gating may produce different (and potentially erroneous) results.

[0097] As noted above, machine learning model predictions may be validated against a hold-out set of manually gated files generated by domain experts using prediction metrics which are relative to the manually-gated hold-out set. Examples of prediction metrics include, but are not limited to, comparison of the means of cell population distributions from the predicted set to within 5%, 10%, 20%, 30% of those for the hold-out set; comparison of standard deviations of cell population distributions from the predicted set to within 5%, 10%, 20% of that for the hold-out set; and/or correlations between cell population ratios for the predicted set to cell population ratios for the hold-out set of greater than 85%, 90%, 95%, or 98%. A correlation of 100% means that the ML model predicted exactly the same cell population ratios as were determined by manual gating.

[0098] In some instances, the threshold for determining validity of the ML model predictions can be modulated by thresholding the standard deviation differences using a standard deviation difference threshold of, *e.g.*, less than 5%, 10%, 15%, or 20%.

[0099] In some instances, the threshold for determining validity of the ML model predictions can be modulated by thresholding the correlation using a correlation threshold of greater than 80%, 85%, 90%, 95%, or 98%.

[0100] The models that pass validation may be accepted as good predictors and stored in database 208. Remaining models may be cycled through one or more additional rounds of training by data science experts and/or biologists 210 with more training data, changes of algorithm, or hyperparameter tuning at each round until they pass validation.

[0101] In some instances, the machine learning model architectures may comprise, for example:

- a gradient descent boosted tree machine learning algorithm with up to a thousand trees;

- a deep neural network with typically between 2 and 4 hidden layers and up to 250 nodes per layer;
- a convolutional neural network with; and/or
- an autoencoder with, for example, up to 7 layers (*e.g.*, 5 convolutions, 2 pooling).

[0102] The model hyperparameters are tunable per a specific application of model training.

[0103] Trained and validated models may be used to predict immune cell sub-population counts and frequencies on all appropriate samples. The results of the predictions on each cell sub-population within a given sample may be aggregated to produce a systems view of the immune state of the individual from which the sample was collected, which may form all or part of an immune profile.

[0104] **FIG. 3** provides an exemplary illustration of a process 300 for machine learning model-based prediction and immune profile generation. Machine learning models that pass validation (stored in database 302) are applied to raw FCS files 306 to produce immune cell population count predictions 304. These predictions are collected into an immune profile (stored in database 308) and made available for further analysis 310, including, *e.g.*, queries of aggregated immune profiles (*e.g.*, aggregated according to an individual or patient's demographic or clinical data including, but not limited to, age sex, race, family history, disease diagnosis, *etc.*) stored in database 312.

[0105] Identification of the most appropriate machine learning algorithm and metadata parameters may be advantageous to help to produce high quality predictions on immune data. The system may be configured to use the most appropriate model architecture, and the model architecture may continue to be iterated. The training of the model(s) may also be updated periodically or continuously as new training data becomes available.

[0106] In some instances, an ensemble machine learning model comprising more than 2000 individual machine learning models may be used to perform automated analysis on a sample. Training and managing large numbers of machine learning models such as this may be accomplished by programmatic infrastructure. The platforms and methods disclosed herein may use a software platform to manage the application of trained machine learning models to raw flow cytometry data to produce cell count predictions.

[0107] Advantages of the platforms and methods disclosed herein may be provided by standardizing the immunophenotyping platform and enabling machine learning-based data processing automation. This enables the disclosed platforms and methods to go from highly complex FSFC data to translatable immune profiles for use in clinically-relevant timeframes (*e.g.*, hours rather than days). In some instances, an immune profile for a sample may be generate in less than 24 hours, 12 hours, 10 hours, 8 hours, 6 hours, or 4 hours.

[0108] The output of the prediction process may include an immune profile. This profile may comprise the immune cell counts and/or sub-population frequencies of all measured cell sub-populations. These immune profiles represent a snapshot of the donor's immune system at the time of sample collection, and several longitudinal samples can be compiled to demonstrate an immune trajectory. Furthermore, several different donor immune signatures may be compiled to discover statistically significant population-level immune signatures. Display of these different use cases may be dependent on specific project requirements.

[0109] The methodology described herein may have advantages as compared to using different panels, different algorithms, and/or different automated gating approaches. Advantages may include generating more accurate and effective cell clustering, cell type or sub-type predictions, and immune profiles. The accuracy of the disclosed machine learning models may be evaluated for example, based on precision (*e.g.*, the percentage of cells classified as being of a given cell type that are actually of that type), recall (*e.g.*, the percentage of cells in a data set that are correctly classified as belonging to a given cell type), and F1 score (*i.e.*, an accuracy metric that combines the precision and recall of a model to assess how many times the model made a correct prediction across the entire data set).

Applications

[0110] The disclosed methods and systems for generating immune profiles may be used for a variety of biomedical research and clinical diagnostic applications. Examples include, but are not limited to, diagnosis of immune-related diseases and disorders, diagnosis of autoimmune diseases and cancer, diagnosis of and/or identification of individual-specific response to infectious diseases, the prediction of response to treatment either prior to or during treatment, and the characterization of donors and products in cell therapy manufacturing.

Systems for Immune System Phenotyping and Automated Cell Classification

[0111] Also disclosed herein are systems designed to implement any of the disclosed methods for generating an immune profile for a sample from a subject. The systems may comprise, *e.g.*, one or more processors, and a memory unit communicatively coupled to the one or more processors and configured to store instructions that, when executed by the one or more processors, cause the system to: receive fluorescence intensity data, or data derived therefrom, acquired using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject. In some instances, the system may further comprise a full spectrum flow cytometer (FSFC) instrument.

[0112] Similarly, non-transitory computer-readable storage media are disclosed that may comprise instructions for operating a system configured to perform any of the disclosed methods for generating an immune profile for a sample from a subject. For example, non-transitory computer-readable storage media storing one or more programs are described, the one or more programs comprising instructions, which when executed by one or more processors of a system, cause the system to: receive fluorescence intensity data, or data derived therefrom, acquired using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject; provide at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

Computer Processors & Computing Systems

[0113] FIG. 4 illustrates an exemplary computing system, in accordance with some implementations. Computing system 400 can be a component of a system for generating an immune profile for a sample from a subject.

[0114] Computing system **400** can include a host computer connected to a network. Computing system **400** can be a client computer or a server. As shown in **FIG. 4**, computing system **400** can comprise any suitable type of microprocessor-based device, such as a personal computer; workstation; server; or handheld computing device, such as a phone or tablet. The computer can include, for example, one or more of processor **410**, input device **420**, output device **430**, memory storage **440**, and communication device **460**.

[0115] Input device **420** can be any suitable device that provides input, such as a touch screen or monitor, keyboard, mouse, or voice-recognition device. Output device **430** can be any suitable device that provides output, such as a touch screen, monitor, printer, disk drive, or speaker.

[0116] Memory storage **440** can be any suitable device that provides storage, such as an electrical, magnetic, or optical memory, including a RAM, cache, hard drive, CD-ROM drive, tape drive, or removable storage disk. Communication device **460** can include any suitable device capable of transmitting and receiving signals over a network, such as a network interface chip or card. The components of the computer can be connected in any suitable manner, such as via a physical bus or wirelessly. Memory storage **440** can be a non-transitory computer-readable storage medium comprising one or more programs, which, when executed by one or more processors, such as processor **410**, cause the one or more processors to execute any of the methods described herein.

[0117] Software **450**, which can be stored in memory storage **440** and executed by processor **410**, can include, for example, the programming that embodies the functionality of the present disclosure (*e.g.*, as embodied in the methods, systems, computers, servers, and/or devices as described above). In some embodiments, software **450** can be implemented and executed on a combination of servers such as application servers and database servers.

[0118] Software **450** can also be stored and/or transported within any computer-readable storage medium for use by or in connection with an instruction execution system, apparatus, or device, such as those described above, that can fetch and execute instructions associated with the software from the instruction execution system, apparatus, or device. In the context of this disclosure, a computer-readable storage medium can be any medium, such as storage **440**, that can contain or store programming for use by or in connection with an instruction execution system, apparatus, or device.

[0119] Software **450** can also be propagated within any transport medium for use by or in connection with an instruction execution system, apparatus, or device, such as those described above, that can fetch and execute instructions associated with the software from the instruction execution system, apparatus, or device. In the context of this disclosure, a transport medium can be any medium that can communicate, propagate, or transport programming for use by or in connection with an instruction execution system, apparatus, or device. The transport-readable medium can include, but is not limited to, an electronic, magnetic, optical, electromagnetic, or infrared wired or wireless propagation medium.

[0120] Computing system **400** may be connected to a network, which can be any suitable type of interconnected communication system. The network can implement any suitable communications protocol and can be secured by any suitable security protocol. The network can comprise network links of any suitable arrangement that can implement the transmission and reception of network signals, such as wireless network connections, T1 or T3 lines, cable networks, DSL, or telephone lines.

[0121] Computing system **400** can implement any operating system suitable for operating on the network. Software **450** can be written in any suitable programming language, such as C, C++, Java, or Python. In various embodiments, application software embodying the functionality of the present disclosure can be deployed in different configurations, such as in a client/server arrangement or through a Web browser as a Web-based application or Web service, for example.

EXAMPLES

[0122] The following examples are included for illustrative purposes only and are not intended to limit the scope of the present disclosure.

Example 1 – Manual gating of full spectrum flow cytometry data.

[0123] One of the underlying principles of flow cytometry data analysis is “gating,” which is the sequential identification and refinement of a cellular population of interest based on labeling a panel of cell type-specific molecules (also known as markers) using, *e.g.*, fluorescently-labeled antibodies, that are detected by fluorescence (Verschoor, *et al.* (2015), “An Introduction to Automated Flow Cytometry Gating Tools and Their Implementation”, *Frontiers in Immunology* Vol. 6, Article 380). **FIG. 5** provides a non-limiting schematic

illustration of a manual gating process for processing full spectrum flow cytometry data. Fluorescence intensity data, or data derived therefrom (*e.g.*, forward scatter data, side scatter data, live cell/dead cell staining, autofluorescence, *etc.*), is generated by the FSFC instrument and reviewed by an expert practitioner. Cell types or sub-types are identified by selecting sections or subsections of the plotted fluorescence data (*e.g.*, fluorescence intensity data plotted as heatmaps in the individual panels of **FIG. 5**) and refining it further using additional criteria, where each additional criterion applied to the analysis constitutes a “gate”. The number of cell detection events circumscribed within each gate is equal to the number of cells identified for that cell population or sub-population. Exemplary gating criteria, as illustrated in **FIG. 5**, include SSC-A (side scatter – area), FSC-A (forward scatter – area), SSC-H (side scatter – height), FSC-H (forward scatter – height), CD45 (fluorescence signal arising from a fluorescently-labeled anti-CD45 monoclonal antibody), *etc.* The numbers associated with specified cell types in some of the panels indicate the percentage of cells from the previous gate that satisfied the current gating criteria.

Example 2 – Training of a neural network model for immune cell type classification.

[0124] This example provides an illustration of using pre-defined gates from an expert to train a machine learning model (*e.g.*, a neural network) that accurately predicts an immune cell type. The initial assumptions underlying the development of the model included: (i) that FCS files include data for hundreds of thousands of cell detection events, (ii) each cell detection event represents a training opportunity, and (iii) a cell’s type, sub-types, and status can be distinctly identified by its inclusion in a set of labelled gates.

[0125] **FIG. 6** provides a non-limiting example of a simplified gating hierarchy used to build a neural network for immune cell classification. Each node of the gating hierarchy represents a different immune cell type, sub-type or cell status. Starting from the root, side scatter high (SSChi) and side scatter low (SSClow) signals may be used, for example, to discriminate between white blood cells (eosinophils and neutrophils) and other immune cells. Fluorescent signals associated with labeling of appropriate cell surface markers may then be used to discriminate between eosinophils and neutrophils. As indicated in the gating hierarchy illustrated in **FIG. 6**, detection of fluorescent signals associated with labeling of other cell surface markers (*e.g.*, CD19, CD3T, GDT, CD4, CD8, *etc.*) may be used to classify immune cells into a number of distinct immune cell sub-types.

[0126] **Table 3** provides a non-limiting example of a panel of fluorescently-labeled monoclonal antibodies that may be used to detect cell surface markers (adapted from Mahnke, *et al.* (2012), "OMIP-013: Differentiation of Human T-Cells", *Cytometry Part A* 81A: 935-936).

Table 3. Non-limiting examples of fluorescently-labeled monoclonal antibodies.

Cell Surface Marker	Monoclonal Antibody Clone	Fluorochrome	Purpose
CD3	SK7	APC-H7	Lineage
CD4	M-T477	QD605	
CD8	RPA-T8	QD585	
CCR7	150503	Ax680	Memory/ Differentiation
CD27	O323	FITC	
CD28	CD28.2	PE-Cy5	
CD31	WM59	PE-Cy7	
CD45RA	HI100	APC	
CD57	NK-1	QD705	
CD95	DX2	PE	
CD127	A019D5	BV421	
CD244	C1.7	PE-Cy5.5	
Dead cells	---	AqBlu	Dump

APC, allophycocyanin; H7, highlight 750; QD, quantum dot; Ax, Alexa; FITC, fluorescein isothiocyanate; PE, R-phycoerythrin; Cy, cyanine; BV, brilliant violet; AqBlu, live/dead fixable aqua dead cell stain.

[0127] **FIG. 7** provides a simplified schematic illustration of a neural network comprising an input layer comprising two or more nodes (or “perceptrons”), at least one hidden layer (comprising four or more nodes), and an output layer (comprising a single node in this non-limiting example). In general, the neural network may comprise any total number of layers, and any number of hidden layers, where the hidden layers function as trainable feature extractors that allow mapping of a set of input data to a preferred output value or set of output values. Each layer of the neural network comprises a number of nodes (or perceptrons). A

node receives input that comes either directly from the input data (*e.g.*, flow cytometry data, or data derived therefrom) or the output of nodes in previous layers, and performs a specific operation, *e.g.*, a summation operation. In some cases, a connection from an input to a node is associated with a weight (or weighting factor). In some cases, the node may, for example, sum up the products of all pairs of inputs from a previous layer and their associated weights. In some cases, the weighted sum is offset with a bias, *b*. In some cases, the output of a node may be gated using a threshold or activation function, *f*, which may be a linear or non-linear function. The activation function may be, for example, a rectified linear unit (ReLU) activation function or other function such as a saturating hyperbolic tangent, identity, binary step, logistic, arcTan, softsign, parameteric rectified linear unit, exponential linear unit, softPlus, bent identity, softExponential, Sinusoid, Sine, Gaussian, or sigmoid function, or any combination thereof.

[0128] The weighting factors, bias values, and threshold values, or other computational parameters of the neural network, can be "taught" or "learned" in a training phase using one or more sets of training data. For example, the parameters may be trained using the input data from a training data set and a gradient descent or backward propagation method so that the output value(s) (*e.g.*, immune cell classifications) that the neural network predicts are consistent with the examples included in the training data set. The adjustable parameters of the model may be obtained using, *e.g.*, a back propagation neural network training process that may or may not be performed using the same hardware as that used for performing immune cell classification.

[0129] A neural network such as that illustrated schematically in **FIG. 7** was trained using labeled training data split into training (typically 80% of the total) and test (typically 20% of the total) data sets. For each cell detection event in the training data set, data was matched to the appropriate input nodes, the neural network model performing an initial classification, the output was compared to the actual (known) output for the detection event, internal model weights were adjusted, and the cycle was repeated. The trained model was then tested by processing the reserved test data and comparing the model predictions for cell type to the actual (known) cell types. The performance of the model was evaluated using metrics such as precision and recall.

[0130] **FIG. 8** provides a non-limiting example of test data generated by a trained neural network classifier. Support, indicated in the right-most column, is the number of cells

classified by manual gating as belonging to the indicated cell type. The performance of the model was evaluated by calculating the precision, recall, and F1-score as described elsewhere herein for 16 immune cell types. Micro average: metrics calculated globally by counting the total number of true positives, false negatives and false positives. Macro average: metrics calculated for each label (cell type) to determine their unweighted mean (this does not take label imbalance into account). Weighted average: metrics calculated for each label to determine their average weighted by support (the number of true instances for each label). This weighted average alters the ‘macro’ calculation to account for label imbalance, and can result in an F-score that is not between precision and recall. Sample average: metrics calculated for each instance to determine their average (this calculation is only meaningful for multilabel classification where this differs from `accuracy_score`). As can be seen, the weighted average for precision and recall were 0.94 and 0.92, respectively, in this study.

[0131] FIG. 9 provides a non-limiting example of test data generated by a trained neural network classifier after training the model on a much larger training data set. As can be seen, the weighted average for precision and recall were 0.98 and 0.98, respectively, in this study.

[0132] FIG. 10 provides a non-limiting example of validation data (*i.e.*, model prediction data for input data not previously provided to the model) for a trained neural network classifier. The weighted average for precision and recall were 0.99 and 0.98, respectively, in this validation run.

[0133] FIG. 11 provides another non-limiting example of validation data generated by a trained neural network classifier. The weighted average for precision and recall were 0.98 and 0.91, respectively, in this validation run.

Example 3 – Application of an ensemble machine learning model to automated cell counting.

[0134] An FCS file consists of metadata about the fluorophores and associated markers used when performing the flow cytometry experiment. It also includes data corresponding to each fluorescent channel for every event detected by the flow cytometer. The conventional approach to extract counts and frequencies of immune cell populations from a sample (such as determining the percentage of Neutrophils identified out of all white blood cells detected) from an FCS file involves using specialized software (*e.g.*, FlowJo, BD Biosciences, Ashland OR). This manual method entails opening the file within the software, visually examining the events across multiple bivariate plots, and identifying cell clusters that conform with well-

established biological phenotypes. It also requires leveraging the association between fluorescent dyes and immune cell surface proteins as defined by the immunophenotyping panel design. Subsequently, enclosing polygons are manually created to delineate clusters of immune cells in a two-dimensional space, aligning with background immunology expertise in looking at cell surface marker distribution. When all immune cell subsets of interest have been identified using this method (2000+ in the case of the IMU (IMU Biosciences, London, UK) phenotyping panel), a signature of cell population ratios can be extracted.

[0135] To increase the speed and accuracy of this process of cell counting and classification, automated cell classification can be applied. Given a standard gating hierarchy, a set of FCS files, and a set of reference configurations of each gate stored in manually gated workspace files, an ensemble machine learning model can be trained to determine a cell population signature. The ensemble machine learning model comprises a set of individual ML models, where each ML model of the set is trained to predict if an event falls inside or outside of a configured gate. Once the set of models is trained, a signature can be constructed without manual intervention as follows.

[0136] For each event in the FCS file, the machine learning ensemble is invoked. Every node in the ensemble may consist, for example, of one machine learning model configured to receive input from each fluorescent channel, where the machine learning model has an internal state that was learned during the training stage, and outputs a set of events that are predicted to be enclosed by the defined gate represented within the ML model. These ML-filtered events are then cascaded into one or several predicted output categories corresponding to the number of children gates at the next level of the gating hierarchy. The set of machine learning models in the ensemble are organized in a directed acyclic graph structure, with the root node being the first filter in the gating hierarchy. To begin the process of predicting an immune signature, the complete set of events in the FCS file is passed as input into the root machine learning model. Events that conform to the gating definition contained in the ML model are retained and provided as input to the next model (or models) in the hierarchy. Every node in the hierarchy represents an immune cell type or status and when the process is complete, all the counts and frequencies of all nodes can be extracted to comprise the complete immune signature for that data file.

[0137] **FIG. 12** provides a non-limiting example of the data flow through cascading predictions from node to node, *i.e.*, from parent ML model output to child/children ML model

inputs. In this example, fluorescence intensities are input to the root node (white blood cell (WBC)) model, and individual detection events are classified into two output categories (WBC subtypes). The fluorescence intensity data for the latter are then input into the next set of models in the gating hierarchy, *e.g.*, either an Eosinophil ML model and Neutrophil ML model, or a B-cell ML model and T-cell ML model. As indicated in the illustration, the fluorescence intensity data for a subset of T-cell gated events may then be input into, *e.g.*, an INKT cell ML model or a γ D cell ML-model.

Example 4 – Application of immune signatures to highlight biologically meaningful insight.

[0138] To demonstrate the effectiveness of the disclosed immunophenotype profiling method to identify and correct for changes in general health lifestyle and health factors of a donor sample, an analysis was performed to determine to what extent vitamin D fluctuations impacted parameters of the immune system. A dataset of 609 donors was assembled (58% female, age 38.7 ± 12.5 years), fresh blood collected and analyzed using the FCSC approach defined herein. In addition, a comprehensive general health blood test (including determination of a lipid profile, vitamin levels, and ~50 other biomarkers) was performed on the same set of donors at the same experimental timepoint. The automated cell classification method produced immune signatures for all donors. These signatures were then evaluated using linear regression models to detect significant changes in immune cell populations based on vitamin D level, with sex and age group (<30years: young, ≥ 30 <60years: middle, ≥ 60 years old) as covariates. Following the application of multiple testing correction, a total of 60 immune populations were identified as affected by variations of Vitamin D level in both sexes and across age groups.

[0139] Recovering the significant interaction and trajectory of Vitamin D with Th2 cells, implicated in allergic responses (Georas, *et al.* (20005), “T-helper cell type-2 regulation in allergic disease”, *Eur Respir J.* 26(6):1119-37), and Th17 cells, implicated in autoimmune diseases and infection response pathways, (Zambrano-Zaragoza, *et al.* (2014), “Th17 cells in autoimmune and infectious diseases”, *Int J Inflam.* 2014:651503) provides a reference range for the baseline phenotype of these two crucial populations in a healthy cohort, which could be used as a monitoring and diagnostic metric for patients undergoing treatment for any of these conditions.

[0140] FIGS. 13A-B provide non-limiting example plots showing the trajectory of T helper 2 (FIG. 13B) and T helper 17 (FIG. 13A) cell populations decreasing as Vitamin D levels increase in 609 health individuals across age groups accounting for sex as a covariate.

EXEMPLARY EMBODIMENTS

[0141] Exemplary methods, systems, and computer-readable storage media are set out in the following list:

1. A method for generating an immune profile for a subject, the method comprising:
 - contacting at least a first aliquot of a sample from the subject with at least a first immunophenotyping panel to fluorescently-label cells contained within the sample;
 - processing the fluorescently-labeled cells using a full spectrum flow cytometer to generate fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells from the sample;
 - providing at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and
 - outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.
2. The method of embodiment 1, wherein the ensemble machine learning model is organized in a cascading hierarchical tree structure comprising a plurality of nodes, and wherein each node comprises an individual machine learning model.
3. The method of embodiment 2, wherein each individual machine learning model comprises one input data set and from one to eight output data sets corresponding to branches of the cascading hierarchical tree structure.
4. The method of embodiment 2 or embodiment 3, wherein each individual machine learning model comprises a neural network model.
5. The method of embodiment 2 or embodiment 3, wherein each individual machine learning model comprises a gradient descent boosted tree model.

6. The method of any one of embodiments 2 to 5, wherein the plurality of nodes comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 nodes.
7. The method of any one of embodiments 2 to 6, wherein a number of individual machine learning models in the ensemble machine learning model is equal to a number of distinct immune cell sub-populations in the plurality of distinct immune cell sub-populations.
8. The method of any one of embodiments 2 to 7, wherein the design of the cascading hierarchical tree structure is based at least in part on an expert analysis of manually-gated fluorescence intensity data, or data derived therefrom, for one or more control samples.
9. The method of any one of embodiments 1 to 8, wherein individual cells are classified independently of all other cells in the plurality of fluorescently-labeled cells.
10. The method of any one of embodiments 1 to 8, wherein individual cells are classified recursively with all other cells of the plurality of fluorescently-labeled cells.
11. The method of any one of embodiments 1 to 10, wherein the ensemble machine learning model is trained using one or more labeled training data sets generated by an expert through manual gating of fluorescence intensity data, or data derived therefrom, for one or more control samples.
12. The method of embodiment 11, wherein the individual machine learning models in the ensemble machine learning model are trained individually using the one or more labeled training data sets.
13. The method of embodiment 11 or embodiment 12, wherein, during training, an individual model's predictions are used to validate the individual model but do not propagate forward through the ensemble machine learning model, thereby eliminating error propagation during training.
14. The method of embodiment 11, wherein the individual machine learning models in the ensemble machine learning model are trained collectively using a recursive training method.
15. The method of any one of embodiments 11 to 14, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are the same for every node in the cascading hierarchical tree structure.

16. The method of any one of embodiments 11 to 14, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are different for a subset of nodes in the cascading hierarchical tree structure.
17. The method of any one of embodiments 11 to 16, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are determined by performing a random grid search of the value ranges for the one or more hyperparameters.
18. The method of any one of embodiments 1 to 17, further comprising performing a mathematical transformation of the fluorescence intensity data, or data derived therefrom, prior to using the transformed fluorescence intensity data as input for the ensemble machine learning model.
19. The method of any one of embodiments 1 to 18, wherein the fluorescence intensity data, or data derived therefrom, comprises fluorescence intensity data for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, or 40 fluorescence detection channels.
20. The method of embodiment 19, wherein the fluorescence intensity data, or data derived therefrom, further comprises forward scatter height data, forward scatter area data, side scatter height data, side scatter area data, autofluorescence data, or any combination thereof.
21. The method of any one of embodiments 1 to 20, wherein the sample comprises a blood sample, a buffy coat sample, or a cell suspension.
22. The method of any one of embodiment 1 to 21, wherein the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with antigen-presenting cells (APCs).
23. The method of embodiment 22, wherein the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to IGM, CD5, CD62L, CD294, CD69, CD38, PD1, CD11C, CD3, CD8, HLADR, CD24, CD337, CD123, CD141, CD1C, CD4, TACI, CD319, CD335, PDL1, CD10, CD45, CD16, IGD, CD40, CD19_TCRGD, CD43, CD14, CD138, CD15, CD56, CD86, CD303, CD27, or any combination thereof.

24. The method of embodiment 23, wherein the panel of fluorescently-labeled antibodies further comprises fluorescently-labeled antibodies directed to cell surface markers that are indicative of live cells, dead cells, or both.
25. The method of any one of embodiments 1 to 24, wherein the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with T cells.
26. The method of embodiment 25, wherein the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to TIGIT, CD5, CD28, CXCR5, CD39, TIM3, CD38, PD1, TCRVA7_2_TCRVD1, CD95, CD3, CD8, HLADR, CD31, CCR4, CCR6, CCR7, CD57, ICOS, CD4, KLRG1, TCRVA24_JA18, CD122, CD103, CXCR3, TCRVD2, CD45, CCR10, CD16, CD25, CD161, CD19_TCRGD, LAG3, CD14, CD45RO, CD56, CD127, CD45RA, CD27, or any combination thereof.
27. The method of any one of embodiments 1 to 26, wherein the plurality of distinct immune cell sub-populations comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 distinct immune cell sub-populations.
28. The method of any one of embodiments 1 to 27, wherein the plurality of distinct immune cell sub-populations includes white blood cells (WBC), Eosinophils, Eosinophil/CD5+, Neutrophils, Neutrophils/big, Neutrophils/CD5+, Neutrophils/small, B-cells, B-cells/CD5-CD27-, Monocytes/CD56+, Monocytes/CD56-, NK-cells, Dendritic cells (DC), T-cells, iNKT cells, gamma delta T-cells (total GD), Vd1 cells, Vd2 cells, Vdx cells, Mucosal-associated invariant T (MAIT) cells, TEMRA cells, CD4 naïve cells, T helper cells, CD4 effector memory cells, Treg cells, or any combination thereof.
29. The method of any one of embodiments 1 to 28, wherein the total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample is output as part of an immune profile in less than 24 hours, 12 hours, 10 hours, 8 hours, 6 hours, or 4 hours.
30. The method of any one of embodiments 1 to 29, wherein the immune profile is used to diagnose an immune-related disease or disorder, monitor progression of an immune-related disease or disorder, or monitor a response to treatment of an immune-related disease or disorder in the subject.

31. A computer-implemented method for generating an immune profile for a subject, the method comprising:

receiving fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from the subject;

providing at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

32. A system comprising:

one or more processors; and

a memory communicatively coupled to the one or more processors and configured to store instructions that, when executed by the one or more processors, cause the system to:

receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject;

provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

33. The system of embodiment 32, wherein the system further comprises a full spectrum flow cytometer.

34. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of a system, cause the system to:

receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject;

provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

[0142] It should be understood from the foregoing that, while particular implementations of the disclosed methods and systems have been illustrated and described, various modifications can be made thereto and are contemplated herein. It is also not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the preferable embodiments herein are not meant to be construed in a limiting sense. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. Various modifications in form and detail of the embodiments of the invention will be apparent to a person skilled in the art. It is therefore contemplated that the invention shall also cover any such modifications, variations and equivalents.

CLAIMS

What is claimed is:

1. A method for generating an immune profile for a subject, the method comprising:
 - contacting at least a first aliquot of a sample from the subject with at least a first immunophenotyping panel to fluorescently-label cells contained within the sample;
 - processing the fluorescently-labeled cells using a full spectrum flow cytometer to generate fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells from the sample;
 - providing at least a subset of the fluorescence intensity data, or data derived therefrom, for the plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and
 - outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.
2. The method of claim 1, wherein the ensemble machine learning model is organized in a cascading hierarchical tree structure comprising a plurality of nodes, and wherein each node comprises an individual machine learning model.
3. The method of claim 2, wherein each individual machine learning model comprises one input data set and from one to eight output data sets corresponding to branches of the cascading hierarchical tree structure.
4. The method of claim 2 or claim 3, wherein each individual machine learning model comprises a neural network model.
5. The method of claim 2 or claim 3, wherein each individual machine learning model comprises a gradient descent boosted tree model.
6. The method of any one of claims 2 to 5, wherein the plurality of nodes comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 nodes.

7. The method of any one of claims 2 to 6, wherein a number of individual machine learning models in the ensemble machine learning model is equal to a number of distinct immune cell sub-populations in the plurality of distinct immune cell sub-populations.
8. The method of any one of claims 2 to 7, wherein the design of the cascading hierarchical tree structure is based at least in part on an expert analysis of manually-gated fluorescence intensity data, or data derived therefrom, for one or more control samples.
9. The method of any one of claims 1 to 8, wherein individual cells are classified independently of all other cells in the plurality of fluorescently-labeled cells.
10. The method of any one of claims 1 to 8, wherein individual cells are classified recursively with all other cells of the plurality of fluorescently-labeled cells.
11. The method of any one of claims 1 to 10, wherein the ensemble machine learning model is trained using one or more labeled training data sets generated by an expert through manual gating of fluorescence intensity data, or data derived therefrom, for one or more control samples.
12. The method of claim 11, wherein the individual machine learning models in the ensemble machine learning model are trained individually using the one or more labeled training data sets.
13. The method of claim 11 or claim 12, wherein, during training, an individual model's predictions are used to validate the individual model but do not propagate forward through the ensemble machine learning model, thereby eliminating error propagation during training.
14. The method of claim 11, wherein the individual machine learning models in the ensemble machine learning model are trained collectively using a recursive training method.
15. The method of any one of claims 11 to 14, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are the same for every node in the cascading hierarchical tree structure.
16. The method of any one of claims 11 to 14, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are different for a subset of nodes in the cascading hierarchical tree structure.

17. The method of any one of claims 11 to 16, wherein the training of the ensemble machine learning model is controlled by one or more hyperparameter values that are determined by performing a random grid search of the value ranges for the one or more hyperparameters.
18. The method of any one of claims 1 to 17, further comprising performing a mathematical transformation of the fluorescence intensity data, or data derived therefrom, prior to using the transformed fluorescence intensity data as input for the ensemble machine learning model.
19. The method of any one of claims 1 to 18, wherein the fluorescence intensity data, or data derived therefrom, comprises fluorescence intensity data for at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, or 40 fluorescence detection channels.
20. The method of claim 19, wherein the fluorescence intensity data, or data derived therefrom, further comprises forward scatter height data, forward scatter area data, side scatter height data, side scatter area data, autofluorescence data, or any combination thereof.
21. The method of any one of claims 1 to 20, wherein the sample comprises a blood sample, a buffy coat sample, or a cell suspension.
22. The method of any one of claim 1 to 21, wherein the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with antigen-presenting cells (APCs).
23. The method of claim 22, wherein the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to IGM, CD5, CD62L, CD294, CD69, CD38, PD1, CD11C, CD3, CD8, HLADR, CD24, CD337, CD123, CD141, CD1C, CD4, TACI, CD319, CD335, PDL1, CD10, CD45, CD16, IGD, CD40, CD19_TCRGD, CD43, CD14, CD138, CD15, CD56, CD86, CD303, CD27, or any combination thereof.
24. The method of claim 23, wherein the panel of fluorescently-labeled antibodies further comprises fluorescently-labeled antibodies directed to cell surface markers that are indicative of live cells, dead cells, or both.
25. The method of any one of claims 1 to 24, wherein the at least one immunophenotyping panel comprises a panel of fluorescently-labeled antibodies directed to cell surface proteins associated with T cells.

26. The method of claim 25, wherein the panel of fluorescently-labeled antibodies comprises fluorescently-labeled antibodies directed to TIGIT, CD5, CD28, CXCR5, CD39, TIM3, CD38, PD1, TCRVA7_2_TCRVD1, CD95, CD3, CD8, HLADR, CD31, CCR4, CCR6, CCR7, CD57, ICOS, CD4, KLRG1, TCRVA24_JA18, CD122, CD103, CXCR3, TCRVD2, CD45, CCR10, CD16, CD25, CD161, CD19_TCRGD, LAG3, CD14, CD45RO, CD56, CD127, CD45RA, CD27, or any combination thereof.

27. The method of any one of claims 1 to 26, wherein the plurality of distinct immune cell sub-populations comprises at least 1000, 1200, 1400, 1600, 1800, 2000, 2200, or 2400 distinct immune cell sub-populations.

28. The method of any one of claims 1 to 27, wherein the plurality of distinct immune cell sub-populations includes white blood cells (WBC), Eosinophils, Eosinophil/CD5+, Neutrophils, Neutrophils/big, Neutrophils/CD5+, Neutrophils/small, B-cells, B-cells/CD5-CD27-, Monocytes/CD56+, Monocytes/CD56-, NK-cells, Dendritic cells (DC), T-cells, iNKT cells, gamma delta T-cells (total GD), Vd1 cells, Vd2 cells, Vdx cells, Mucosal-associated invariant T (MAIT) cells, TEMRA cells, CD4 naïve cells, T helper cells, CD4 effector memory cells, Treg cells, or any combination thereof.

29. The method of any one of claims 1 to 28, wherein the total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample is output as part of an immune profile in less than 24 hours, 12 hours, 10 hours, 8 hours, 6 hours, or 4 hours.

30. The method of any one of claims 1 to 29, wherein the immune profile is used to diagnose an immune-related disease or disorder, monitor progression of an immune-related disease or disorder, or monitor a response to treatment of an immune-related disease or disorder in the subject.

31. A computer-implemented method for generating an immune profile for a subject, the method comprising:

receiving fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from the subject;

providing at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine

learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

outputting a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

32. A system comprising:

one or more processors; and

a memory communicatively coupled to the one or more processors and configured to store instructions that, when executed by the one or more processors, cause the system to:

receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject;

provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.

33. The system of claim 32, wherein the system further comprises a full spectrum flow cytometer.

34. A non-transitory computer-readable storage medium storing one or more programs, the one or more programs comprising instructions, which when executed by one or more processors of a system, cause the system to:

receive fluorescence intensity data, or data derived therefrom, generated using a full spectrum flow cytometer to process a fluorescently-labeled cell sample collected from a subject;

provide at least a subset of the fluorescence intensity data, or data derived therefrom, for a plurality of fluorescently-labeled cells as input to an ensemble machine learning model

configured to process the fluorescence intensity data, or data derived therefrom, and classify individual cells of the plurality of fluorescently-labeled cells as belonging to one of a plurality of distinct immune cell sub-populations; and

output a total cell count or cell frequency for each of the plurality of distinct immune cell sub-populations in the sample as part of an immune profile for the subject.