and a wide variety of lymphomas all have different symptoms, prognoses, and require different treatment strategies.

Summary of the Invention

[0004] Due to the wide variety of hematologic malignancies, a rapid and precise classification is critical to the first step towards effective disease management and cure for patients.

[0005] Flow cytometry (FC) generates data from high-throughput streams of individual cells from bone marrow, blood or lymphoid tissue specimens for high quality hematologic malignancy screening, diagnosis, and monitoring. Antibodies labeled with fluorescent tags permit characterization of the complex expression of cell proteins (i.e. antigens) by flow cytometry. Thousands or millions of cells in a specimen sample are evaluated using a panel of multiple antibodies, that are distinguished from each other by channels specific for the emission of the different fluorescent tags, producing large amounts of data points. Physicians or inspectors rely on visualization tools to display the fluorescence that indicates the expression of pairs of antigens among cell populations on two-dimensional scatter plots, and then they perform a hierarchical gating procedure to identify the abnormal cell populations. After the physicians or inspectors interpret multiple pairs of antibody-fluorescence combinations, the type of hematologic malignancy of a patient can then be determined in conjunction with the evaluation of morphologic findings and other appropriate testing. The manual gating process for FC is laborious and suffers from inter-physician subjectivity. Current tools assist the gating process and mostly only offer unsupervised cell population clustering. The purpose of these tools is to quickly identify the cell populations that seem to form related clusters; however, the ultimate procedure to generate an interpretation itself remains unchanged.

inspectors efficiently to determine the type of hematologic malignancy of a patient. Additionally, the traditional methods of processing cytometric data cost too much computing power and memory space. The novel methods or devices provided in the present disclosure can save computing power, memory space, and the consumed electric power.

[0008] To deal with the bottleneck of the computing power and memory space, the operation of randomly downsampling the cytometric data is proposed. However, such random downsampling may create unwanted variability but also run the risk of discarding important cell data or residual tumor cell data. The novel methods or devices provided in the present disclosure can reduce the unwanted variability and the risk of discarding important cell data or residual tumor cell data.

[0009] The exemplary embodiments disclosed herein are directed to solving the issues relating to one or more of the problems presented in the prior art, as well as providing additional features that will become readily apparent by reference to the following detailed description when taken in conjunction with the accompanied drawings. In accordance with various embodiments, exemplary systems, methods, devices and computer program products are disclosed herein. It is understood, however, that these embodiments are presented by way of example and not limitation, and it will be apparent to those of ordinary skill in the art who read the present disclosure that various modifications to the disclosed embodiments can be made while remaining within the scope of the invention.

[0010] An embodiment of the present disclosure provides a method of processing cytometric data. The method comprises: dividing a first data matrix into a first plurality of first submatrices; encoding each of the first submatrices into one corresponding vector representation to acquire a first plurality of vector representations; and aggregating the first plurality of

communication interface 103, and memory 104. The input/output interface 102 is coupled with the processor 101. The input/output interface 102 allows the user to manipulate the computing device 100 to perform the procedures, operations, or methods of the present disclosure (e.g., the procedures, operations, or methods disclosed in FIGS. 2 and 3).

[0026] The communication interface 103 is coupled with the processor 101. The communication interface 103 allows the computing device 100 to communicate with data outside the computing device 100, for example, receiving data including cytometric data of a patient, the patient's information, the method, algorithm, program, or software to be performed, or the configurations of the method, algorithm, program, or software. The data received through the communication interface 103 may be stored in one or more databases outside the computing device 100.

[0027] A memory 104 may be a non-transitory computer readable storage medium. The memory 104 is coupled with the processor 101. The memory 104 has stored program instructions that can be executed by one or more processors (for example, the processor 101). Upon execution of the program instructions stored on the memory 104, the program instructions cause performance of the one or more procedures, operations, or methods disclosed in the present disclosure. For example, the program instructions may cause the computing device 100 to perform: receiving a first data matrix indicative of a first plurality of properties of a first set of cells; encoding the first brain image to generate a latent vector by the processor 101; dividing the first data matrix into a first plurality of first submatrices by the processor 101; encoding each of the first submatrices into one corresponding vector representation to acquire a first plurality of vector representations by the processor 101; and aggregating the first plurality of vector representations into a first ensemble representation by

approaches are effective in learning sample-level FC representation, they remain discriminatively suboptimal. That is, these approaches treat the learning of sample-level representation and classifier training as two detached and independent modules. Instead of implementing sample-level representation learning and the outcome classifier at different stages, building a discriminative network in an end-to-end manner can better optimize the overall performances. Previous SOTA distributional modeling, i.e., FV encoding, consists of non-differentiable functions making it infeasible for end-to-end learning. We extended the FV encoding approach by introducing NetFV, which reorganized the latent assumption to enable discriminative gradient propagation in the deep layers for FC representation learning. The framework of NetFV considers the label back-propagation as conditions in learning the latent cell clusters beyond purely relying on unsupervised data clustering.

[0032] While these end-to-end approaches are appealing, the parameterized maximum likelihood optimization usually suffers from the GPU memory bottleneck during actual implementation. This technical issue becomes increasingly challenging and relevant because the need to detect different antigens is constantly evolving and the increasing capacity to simultaneously detect more and more fluorescent markers in a single tube. The growing diversity of hematologic malignancies that are recognized requires development of more customized marker panels, producing even higher dimensional FC data. Consequently, downsampling cells may be inevitable in some FC representation learning implementations. Random downsampling not only creates unwanted variability but also runs the risk of discarding important cell data or residual tumor cell data.

[0033] The present disclosure addresses two major issues in FC data

properties of a set of cells, and the set of cells correspond to the cells of a sample or a tube. In some embodiments, the plurality of properties of the set of cells correspond to the properties of the antibodies shown in one of the rows of Tube 1 through Tube 5 for the UPMC dataset. In some other embodiments, the plurality of properties of the set of cells correspond to the properties of the antibodies shown in one of the rows of Tube 1 through Tube 3 for the hema.to dataset. The data matrix in operation 303 may correspond to the data matrix 230 in FIG. 2. The data matrix in operation 303 may have D columns and T rows, in which D is the number of properties inspected and T is the number of cells inspected (or the number of cells in the set of cells).

[0081] In operation 305, the data matrix is divided or chunked into a plurality of submatrices. The submatrices may correspond to chunks 241 in FIG. 2. Operation 305 may correspond to the operation of chunking 240 in FIG. 2.

[0082] In operation 307, each of the plurality of submatrices is encoded into one corresponding vector representation. The plurality of submatrices are encoded into corresponding vector representations. The number of the plurality of submatrices may be associated with or identical to the number of the vector representations. In operation 307, the plurality of vector representations are acquired. The vector representation may correspond to the chunk representations 252 in FIG. 2 or the aforementioned chunk-level embeddings. Operation 307 may correspond to the chunk-level pooling network 250 in FIG. 2.

[0083] In operation 307, each of the submatrices may be transposed (corresponding to the operation of transposing 251 in FIG. 2). In some embodiments, since the submatrices are transposed, the number of rows of the submatrices may be correspond to the number of rows of the data