IJWR

Enhancing Oncological Diagnosis by Single-Cell ATAC-seq Data for Internet of Medical Things

Hossein Haririmonfared[a]; Naser Elmi[b]; Kaveh Kavousi[b]; Babak Majidi[a]*

[a] *Department of Computer Engineering, Khatam University, Tehran, Iran*; {h.hariri, b.majidi}@khatam.ac.ir

[b] *Department of Bioinformatics, Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran*; {naser.elmi, kkavousi}@ut.ac.ir

ABSTRACT

Early cancer detection is crucial for improving patient survival rates, as timely intervention greatly enhances treatment efficacy. One promising method for early detection is identifying cancerous cells through the detection of protein-level modifications, which serve as early indicators of malignancy. These protein modifications often result from complex biochemical processes that occurs before visible cellular abnormalities, making them critical targets for diagnostic technologies. In recent years, wireless biomedical sensors have advanced significantly, enabling precisely detecting these protein-level changes. These sensors have the potential to detect cancer at its earliest stages by monitoring the subtle alterations in protein structures and functions that distinguish healthy cells from cancerous ones. As the costs of genetic analysis continue to decrease, the development of Medical Internet of Things (MIoT) devices has become increasingly feasible. These devices are designed to perform real-time analyses of biological specimens—such as blood and urine—by detecting protein-level changes indicative of cancer. In this paper, a new machine learning method based on Extreme Randomized Trees (ERT) is developed to increase the speed of classification of cancerous cells based on single-cell Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). The proposed method enhances the classification speed of the limited and noisy ATAC-seq data as it requires less computation to determine the best splits at each node of the decision trees. This method can significantly improve near real-time cancer risk assessment using samples collected by MIoT. Our proposed method achieves classification accuracy comparable to state of the art single-cell ATAC-seq data analysis techniques while reducing processing time by 259%, challenged by various low-data scenarios. This approach presents an efficient solution for rapid cancer monitoring within the MIoT framework.

Keywords*— *Single cell ATAC-seq, Machine Learning, Extremely Randomized Trees, Classification, Early cancer detection, Biomedical IoT devices.

1. Introduction

Gene expression, the complex process through which genes are translated into proteins, is fundamental for cellular growth, development, and normal physiological function. A critical component in regulating gene expression is the suite of proteins that bind to DNA and facilitate its transcription. These transcription factors exert regulatory control over gene expression at multiple levels, including the activation or control of gene expression, modulation of the rate of gene transcription, and alteration of the type of protein synthesized from a given gene. The understanding of transcription factors is pivotal for explaining the mechanisms governing gene expression under various conditions, thereby enabling the development of innovative therapeutic strategies for disease treatment and enhancement of human health.

Concurrently, chromatin accessibility, which concerns the structural configuration of DNA and its associated proteins, indicates the extent to which DNA is exposed to transcription factors. This metric is instrumental in identifying transcription factors and revealing gene regulatory mechanisms. The Assay for Transposase-Accessible Chromatin using

http://dx.doi.org/10.22133/ijwr.2024.459489.1222

Citation H. Haririmonfared; N. Elmi; K. Kavousi; B. Majidi, "Enhancing Oncological Diagnosis by Single-Cell ATAC-seq Data for Internet of Medical Things", *International Journal of Web Research*, vol.7, no.3,pp.1-13, 2024, doi: http://dx.doi.org/ 10.22133/ijwr.2024.459489.1222.

**Coressponding Author*

Article History: Received: 25 January 2024 ; Revised: 29 April 2024; Accepted: 29 May 2024.

sequencing (ATAC-seq) is a sophisticated molecular biology technique employed to evaluate chromatin accessibility across the genome of a cell. Furthermore, advancements in genetic research tools now permit the analysis of individual cells instead of bulk cell populations, thereby providing unprecedented granularity in genetic studies.

With the advent of machine learning methodologies, the voluminous data generated from molecular biology experiments can be analyzed fast and with high precision. This facilitates a more profound comprehension of various diseases and the development of novel therapeutic interventions. In this research, a novel classification method for single-cell ATAC-seq (scATAC-seq) data is proposed and a comparative analysis of its performance against existing classification methodologies is presented.

This is an emerging field of research that focus on the identification of different cell types including different cancerous cell types from different stages throuth the patterns in the chromatin accessibility of their DNA using scATAC-seq data and machine learning algorithms. These methods can help the diagnosis and prognosis process by early detection of malignancy and personalized therapy for the patient. The classification algorithm is a critical bottleneck in the ATAC-seq classification of cancerous cells. The recently presented state-of-the-art algorithms have achieved this goal with high accuracy, but the speed of classification is a significant problem in these algorithms along with keeping accuracy at high levels. The main novelty of the proposed method in this paper is improving the speed of classification by an average of 259% while maintaining the accuracy of the state-of-the-art methods. Moreover, this novel technique has been tested in four independent scATAC-seq datasets. Moreove the robustness of the proposed method on training data volumne examined.

2. Literature Review

The advent of biosensor technology is transforming the healthcare landscape by enhancing disease diagnostics and enabling continuous monitoring of patient's vital parameters in real time. Leveraging the nano-technological innovations, Ibrahim et al. [1] have introduced an innovative biosensor that employs a miniaturized, chip-integrated CRISPR/Cas system to interact with a patient's DNA sample. The resultant signal is subsequently digitized and securely stored in a cloud-based infrastructure. Advanced machine learning algorithms then process this data, whether in the images or numerical values, to discern patterns that can assist medical professionals in making informed clinical decisions.

Presenting an application of IoT's within the biological sciences, Parks et al. [2] present an IoT framework for cell biology investigations. Their approach involved the fabrication of specialized experimental apparatuses, including electrophysiology setups, microscopy tools, and microfluidics systems, all orchestrated through a cohesive control mechanism predicated on Raspberry Pi technology. This architecture facilitates real-time oversight and modulation of experiments, giving automation capabilities, and providing continuous updates on experimental integrity. The data harvested from these laboratory instruments undergoes a meticulously designed analytical pipeline. The analytical outcomes are subsequently rendered via Plotly Dash for visualization purposes and are made accessible to end-users through a web-based interface. Furthermore, the architecture is engineered to enable other researchers to access the analytical results in the cloud, utilizing platforms such as Nextflow.

Another significant issue in this domain is the safeguarding of user data privacy. To mitigate security challenges in communication, Ugandaran et al. [3] introduced an innovative cryptosystem inspired by DNA computing and the splicing system. This methodology capitalizes on biological information gathered by IoT devices such as Arduino. Before encryption through the contextual array splicing system, the input string is transmuted into a DNA sequence via a DNA sequencing mechanism. Upon the information's transfer to the cloud, the decryption process employs a binary sequencing system to revert the data from the DNA format, followed by a modified contextual array splicing system. Subsequently, predefined algorithms are utilized to analyze the cloud data for interpretation. Essentially, their approach harnesses the potential of biological data to fortify communication security.

Contemporary implementations of Internet of Things (IoT) technologies within the medical domain face many challenges. These challenges can be principally categorized into several domains: the capability for real-time data processing, the security of communication channels and data storage, the preservation of user rights, and the identification of optimal technologies for the collection and precise real-time analysis of diverse patient-related data.

In recent years, novel methodologies and software tools have been developed to mitigate the limitations of current DNA analysis. Chromatin accessibility analysis has become a potent technique for clarifying the epigenetic landscapes of diverse cell types. Initially, DNase-seq and FAIRE-seq were the leading technologies for chromatin accessibility analysis, yet substantial input requirements constrained their therapeutic application. The advent of ATAC-seq in 2013 [4] revolutionized chromatin accessibility analysis owing to its streamlined protocol involving the insertion of the Tn5

transposase via PCR, necessitating minimal sample volumes. Critically, the capacity to analyze single-cell data emerged as the quintessential advantage of this method. Independent research groups under the terms "scATAC-seq" and "sciATAC-seq" in 2015 [5, 6], showcased this capability in their studies. Augmenting its potential further, the recent innovation of droplet-based single-cell combinatorial indexing for ATAC-seq (dsciATAC-seq) enables extensive, high-throughput epigenomic profiling at the single-cell level [7].

In tandem with the advancement of single-cell ATAC-seq, a plethora of analytical methodologies have been developed to interrogate single-cell epigenomes with unprecedented granularity. Nevertheless, the intrinsic high dimensionality and sparsity characteristic of single-cell ATAC-seq data pose substantial computational challenges when contrasted with single-cell RNA-seq data. To mitigate the issue of sparse scATAC-seq data, researchers have introduced a variety of machine learning paradigms, encompassing both unsupervised and supervised strategies. Among these, chromVAR stands out by leveraging the spatial distribution patterns of transcription factor (TF) occurrences within open chromatin regions. It employs t-SNE for dimensionality reduction, projecting corrected deviation vectors of individual cells onto a two-dimensional plane. This facilitates the visualization of different cell types by their distinct TF binding profiles. Moreover, chromVAR's utility extends to the analysis of k-mer frequencies, enabling the de novo identification of novel and unannotated regulatory motifs [8]. Similar to chromVAR, BROCKMAN method also utilizes k-mer factorization to address this challenge [9]. SCARAT method capitalizes on established features such as TF motifs, ENCODE clusters, and MSigDB gene categories, thereby enabling the efficient and accurate extraction of biologically relevant information from single-cell data [10]. Distinctively, scABC exploits the inherent structure of read count patterns in genomic regions for unsupervised k-medoid clustering of cells, removing the necessity for unimportant information [11].

Natural Language Processing (NLP) methodologies offer sophisticated tools for probing single-cell chromatin accessibility datasets. An investigation by Cusanovich et al. used the capabilities of Latent Semantic Analysis (LSA) to discern discrete cellular clusters within an extensive single-cell atlas of murine organs, thereby underscoring the utility of NLP techniques in the clarification of intricate biological phenomena [12]. Additionally, the cisTopic framework shows an advanced probabilistic approach aimed at defining co-accessible enhancers and identifying robust cellular states. This framework integrates latent Dirichlet allocation (LDA) with a collapsed Gibbs sampling algorithm to infer and characterize distinct cis-regulatory motifs within the dataset [13].

Cicero et al. [14] have conceptualized a sophisticated machine-learning framework that incorporates a graphical lasso to enhance the precision of predicting cis-regulatory DNA interactions, thereby marking a substantial progression in clarifying genomic regulatory networks. By adopting a methodology that joins single-cell data aggregation with similarity-based sampling. Cicero effectively quantifies the co-variation among potential regulatory elements. These elements are subsequently associated with their respective target genes by applying unsupervised machine-learning algorithms. The innovative merit of this approach is underscored by its capacity to generate predicted interactions that exhibit a significant similarity with independent 3D chromatin conformation datasets, such as ChIA-PET and Hi-C. Extending its analytical purview, Cicero leverages single-cell ATAC-seq data not only to predict gene expression and 3D chromatin architecture but also to investigate chromatin accessibility [15].

In other researchs some user friendly tools developed to facilicate the analysis of scATAC-seq data. For instance Scasat [16] and Snap ATAC [17] developed to facilitate the analysis of ATAC-seq data. However, the present paper is a continuation of our previous work [18]. In this paper, we compare our novel method performance with a deep neural network approach and test our model feasibility in various input data amounts.

In recent years, the advent and pervasive application of deep learning methodologies have significantly impacted the analysis of ATAC-seq data, which is characterized by high dimensionality, sparsity, and a binary nature. To address mentionedchallenges, novel approaches leveraging deep learning have been proposed. Among these, the method known as SCALE, introduced by Xiong et al. [19], has gathered attention for its efficacy in analyzing single-cell ATAC-seq (scATAC-seq) data through latent feature extraction. SCALE mitigates extant issues by reducing data noise and imputing missing peak signals, thereby recovering incomplete data. Furthermore, this method excels in dimensionality reduction, offering superior representation to traditional techniques such as Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). Despite these advantages, SCALE has limitations; it presumes a constant read depth and neglects batch effects, necessitating the development of an enhanced method named SAILER [20]. SAILER aims to provide a more scalable and accurate learning framework, addressing the shortcomings identified in SCALE.

Tan et al.[21] introduced a multimodal deep learning model that combines one-dimensional

genome sequence data with three-dimensional chromatin structures to improve the prediction of non-coding variant effects on epigenetic profiles. The model, which integrates convolutional, dense, and graph neural networks, exceeds sequence models using long-range interactions and is very effective in predicting genetic expression and pathogenicity in various learning environments. Also, Jing [22] introduced the Single-Cell Deep Topology Embedded Characterization (scDTEC) model, which combines chromatin accessibility profiles and cellular topology with low-dimensional representation to improve the clustering of single-cell scATAC-seq data. Using a topology variation autoencoder and a joint optimization approach, scDTEC effectively addresses technical problems such as noise and data shortages, surpassing other advanced methods in the accurate partition of cell groups.

Ding et al. [23] present DeepSTF that is a complex learning model that uses a unique combination of convolutional neural networks, improved transformer encoders, and Bi-LSTM to predict the binding sites of transcription factors (TFBS) and integrate DNA sequence data with DNA shape profiles. Experiments with 165 ENCODE ChIP-seq datasets have shown that DeepSTF significantly outperforms existing models and demonstrates the critical role of DNA shape characteristics and transformer encoders in capturing complex dependencies and improving prediction accuracy. Ramakrishnan et al. [24] have developed a tool called DeepRegFinder, which is a customizable tool that automates the identification of regulatory elements such as enhancers and promoters by using histone-marking ChIP-seq data, with greater precision and recall than existing methods. DeepRegFinder uses convolutional and repeated neural networks and categorizes these elements into active and positioned states, rationalizing genomic analysis for multiple cell types.

Cellcano [25] is a new computational method that uses a two-stage supervised learning algorithm to accurately identify cell types from data on scATAC-seq and addresses the growing need for specialized tools in this field. Cellcano effectively manages distribution changes between reference and target data. It exhibits high accuracy, robustness, and efficiency for 50 standardization tasks, making it a valuable tool for epigenetic analysis in single-cell studies. Cellcano uses two-step supervision learning processes, first predicting cell types using multilayer perceptron (MLP) and identifying well-planned target cells (anchors) to form a new training set. Then, a self-knowledge distillation model is trained on these anchors to accurately predict the cell types of the other non-anchored cells, thus improving overall prediction accuracy.

2.1. ATAC-seq Applications

ATAC-seq is a widely approach used to explore the molecular mechanisms of cancer development, study immune cells for cancer immunotherapy, predict tumor stage and metastasis risks, and investigate targets for cancer treatment through drug studies. It plays a role in understanding the transcription factors involved in cancer progression and identifying potential therapeutic targets. Zhao et al [26] investigated applications of ATAC-seq on various types of cancer. Specific examples include acute myeloid leukemia, where different genetic mutations like CEBPA and FLT3-ITD have distinct prognostic implications.

Another research [26] analyzes single-cell chromatin accessibility and gene expression in human breast tumors and healthy tissue, identifying potential cells of origin for different tumor types. It introduces a new method to link regulatory elements with gene expression changes, revealing that some elements shift from silencing genes in normal cells to enhancing them in cancer cells, leading to the activation of key oncogenes. Additionally, ATAC-seq can be instrumental in characterizing the epigenetic features of CD8+ T cells (Tex), which are crucial in preventing bacterial infections and tumor development. Chen et al.'s work [27] examines the changes in chromatin accessibility and epigenetic characteristics of Tex following immune checkpoint blockade, as revealed by ATAC-seq. This research aims to uncover new therapeutic targets for persistent viral infections and cancer, while also offering fresh perspectives for designing effective immunotherapies to treat cancer and chronic infections.

Overall, we can categorize the applications of ATAC-seq into five groups:

- Cell type identification and heterogeneity: Examine chromatin accessibility profiles to distinguish cell types, investigate cellular heterogeneity, and detect rare cell populations.
- Gene regulation and enhancer activity: Explore the regulatory landscape of individual cells, pinpoint active enhancer regions, and uncover interactions between chromatin and gene-regulatory elements to gain insight into gene expression regulation.
- Cell differentiation and development: Examine how cells differentiate into various cell types, uncover the epigenetic mechanisms guiding cellular differentiation, and gain insights into cell fate determination.
- Mechanisms of disease and pathology: Investigate disease-specific changes in chromatin accessibility, identify regulatory elements linked to diseases, and understand

the epigenetic mechanisms driving disease progression.

- Integration with other omics data: Integrate single-cell ATAC-seq with techniques like single-cell RNA sequencing (scRNA-seq) for multi-omic analyses, providing deeper insights into gene regulation in cellular processes and disease.

2.2. CTCs and Omics Data

Circulating tumor cells (CTCs) facilitate the spread of cancer through the bloodstream, and their analysis can provide crucial insights into cancer metastasis. Ting et al. [28] used single-cell RNA sequencing to study CTCs in pancreatic cancer, identifying three distinct clusters with varying gene expression profiles. Notably, one cluster exhibited both epithelial and mesenchymal markers, stem cell-associated genes, and, unexpectedly, extracellular matrix proteins, suggesting CTCs may carry their microenvironment. Despite the promise of single-cell RNA sequencing, current methods face limitations like low throughput and antibody dependence. The novel Hydro-Seq; method overcomes these challenges using size-based capture, high capture efficiency, and contamination-free microfluidic chambers, making it suitable for large-scale CTC analysis and potentially improving targeting strategies.

2.3. Non-Invasive Sampling

In a recent study by Zieren et al. [29], it was shown that CTCs can be used as non-invasive diagnostic biomarkers. The clinical landscape of renal cell carcinoma (RCC) is evolving, with an increasing prevalence of incidental and early-stage detections, thereby introducing novel diagnostic complexities. The advent of innovative diagnostic biomarkers capable of differentiating benign from malignant small renal masses (SRMs) holds promise for mitigating the risks of overtreatment. Unlike traditional tissue biopsies, liquid biopsies—derived from a patient's blood or urine—are minimally invasive and facilitate longitudinal disease monitoring. The most promising liquid biopsy biomarkers for RCC diagnosis are circulating tumor cells, extracellular vesicles (EVs), and cell-free DNA. Circulating tumor cell assays exhibit the highest specificity, reduced processing time and cost-efficiency. Nevertheless, their application in SRM diagnostics is constrained by inherent biological characteristics and limited sensitivity.

In an alternative methodology for non-invasive biopsy delineated by Shi et al. [30], a groundbreaking integrated microfluidic chip has been developed to facilitate the sequential enrichment, isolation, and characterization of circulating tumor cells (CTCs) at the single-cell level. This innovative chip enables the analysis of individual CTCs within the same microfluidic platform. The chip is capable of blood clot filtration, single-cell isolation, identification, and the collection of lysates from targeted single cells. Validation experiments, wherein tumor cells were spiked into whole blood samples, demonstrated the chip's efficacy in performing RNA sequencing of single-cell CTCs. This approach establishes a robust foundation for comprehensively analyzing of RNA expression profiles in individual CTCs.

Xu et al. [31] have developed a novel protocol for ATAC-seq data analysis that efficiently profiles chromatin accessibility at the single-cell level.

This method combines bulk Tn5 transposase chromatin tagging with flow cytometric isolation of individual nuclei or cells, followed by the direct addition of sequencing library preparation reagents. The protocol produces high-complexity data with an excellent signal-to-noise ratio and supports comprehensive cell-type characterization via index sorting. The workflow, which takes one to two days, can process hundreds to thousands of nuclei and requires only basic molecular biology techniques and access to flow cytometry facilities.

3. ATAC-seq and Algorithm Description

The principal aim of this research is to introduce an innovative approach to ease the challenges associated with processing time , thereby advancing data analysis towards real-time applicability in edge devices within the realm of the Mobile Internet of Things (MIoT) in the foreseeable future. This study improves the computational time coupled with satisfactory accuracy in clustering single-cell ATAC sequencing data, leveraging a specific decision tree methodology known as Extreme Randomized Trees (ERT). The intrinsic noise and sparsity characteristic of scATAC-seq data poses formidable challenges for the precise extraction of biological signals and the formulation of robust hypotheses.

Mammalian DNA is compacted into three hierarchical structures: nucleosomes, chromatin, and chromosomes, with chromatin transitioning between euchromatin and heterochromatin to regulate gene expression. High-throughput sequencing technologies, mainly ATAC-seq, introduced in 2013, have revolutionized the study of epigenetic mechanisms by enabling the analysis of chromatin accessibility. ATAC-seq has become widely adopted, contributing significantly to research on enhancer landscapes, chromatin changes during hematopoiesis and leukemia, and chromatin states in diseases such as schizophrenia and various cancers cataloged in The Cancer Genome Atlas (TCGA).

ATAC-seq is a crucial technique for understanding chromatin architecture in specific cell types and its changes under pathological conditions,

it is notable for its efficiency with minimal cell numbers and lack of dependency on prior epigenetic knowledge. This study introduces a novel computational framework for classifying single-cell ATAC-seq data using supervised machine learning, achieving high accuracy with reduced computational costs compared to existing methods. Building on Guo et al.'s [36] comprehensive evaluation of six machine learning algorithms on scATAC-seq datasets, the proposed approach leverages the similarities in structure and feature selection between scRNA-seq and scATAC-seq data to enhance performance.

The analysis of single-cell ATAC-seq (scATAC-seq) data typically begins with unsupervised clustering to group cells based on chromatin accessibility profiles, focusing on identifying open chromatin regions, or "peaks". Cells are then labeled by cell type using marker genes within these regions, a process validated by prior research but hindered by the manual labor required for marker curation and cluster inspection, especially as datasets grow in size and complexity. To address these limitations, various machine learning classification methods like support vector machines, neural networks, and random forests have been adapted from scRNA-seq analysis for use in scATAC-seq.

This study evaluates the performance of such methodologies on four independent scATAC-seq datasets, finding that the proposed method achieves accuracy comparable to state-of-the-art techniques while requiring significantly less computational time. The results suggest that this approach is efficient, accurate, and adaptable to diverse scATAC-seq datasets, making it a promising tool for single-cell chromatin accessibility analysis.

In this investigation, we undertook a comprehensive performance assessment of six widely utilized machine learning algorithms—namely, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests (RF), Linear Discriminant Analysis (LDA), Neural Manifold Clustering (NMC), and Decision Trees (DT)—for the automated identification of cell types in scATAC-seq data. Furthermore, to assess the performance of novel deep approaches, we compared the results with a dense neural network method in addition to the last six algorithms. The evaluation framework incorporated four publicly accessible scATAC-seq datasets encompassing a range of biological contexts: human immune cells (Corces2016), the human hematopoietic cell system (Buenrostro2018), and mononuclear peripheral blood cells prepared using two distinct technologies (10X PBMCs v1 and 10X PBMCs Next Gem). The evaluation was divided into two distinctive stages:

- Intra-dataset (within-data) experiments: A five-fold cross-validation approach was employed on each dataset to measure the classification efficacy of each algorithm within its respective data setting.
- Inter-dataset (cross-data) experiments: The predictive ability of the algorithms was scrutinized by training models on one dataset (10X PBMCs v1) and subsequently applying them to predict cell types in a different dataset (10X PBMCs Next Gem), thereby simulating a more pragmatic scenario of deploying a trained model on an unannotated, novel dataset.

In addition, our evaluation framework was expanded to incorporate Extremely Randomized Trees (ERT) [37], facilitating comparative analysis of its classification accuracy relative to pre-existing methodologies within intra-dataset and inter-dataset contexts. Analogous to Random Forests, ERT constructs an ensemble of decision trees during the training phase. However, ERT distinguishes itself by employing random splits at each node and utilizing diverse subsets of features for each split, thereby enhancing robustness and potentially augmenting classification accuracy. Although bootstrapping (re-sampling with replacement) is not a default characteristic of ERT, certain implementations do accommodate this option. By integrating intra-dataset and inter-dataset evaluations across a spectrum of algorithms and datasets, this research offers a thorough appraisal of machine learning techniques for the automatic classification of cell types in scATAC-seq data. The crucial advantages that motivated the selection of ERT better classification are attributable to randomized node splitting.

Another recent method used for evaluation in this study is Multi Layered Perceptron (MLP). The MLP is one of the most commonly used architectures neural network. MLP neural network structure consists of an input layer, hidden layers, and an output layer. Each layer consists of a set of perceptron neurons.

Furthermore, we employed performance metrics such as the F1 score and accuracy to assess the efficacy of the various methods quantitatively. The equations for these metrics are delineated in Equ (1) to (4):

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}} \quad (4)$$

4. Experimental Results

The first phase of the data analysis starts with the alignment of the raw sequencing data for each cell, initially in fastq format, subsequently converting and preserving the output as a bam file. Thereafter, each bam file undergoes a sorting process, followed by the elimination of duplicate reads specific to each cell. Post aggregation, the files are converted to SAM file, from which cell-bin matrices are constructed. This construction process involves the definition of fixed-size genomic windows, and the subsequent employment of the peak file in conjunction with the BAM file to populate a cell-window matrix, formatted within the dgCMatrix structure.This is a sparse matrix format to store the count value of each peak for all cells. The evaluation of these predictions, compared against the ground truth cell types, is conducted through a spectrum of metrics, including the F1 score, confusion matrix, and accuracy, as visually represented in Figure 1.

The count matrices and corresponding ground truth annotations for the 10 × PBMCs Next Gem and 10 × PBMCs v1 datasets were procured from the publicly accessible repository maintained by 10x Genomics. Supplementary datasets were sourced from the National Center for Biotechnology Information (NCBI) repository, specifically from the Gene Expression Omnibus (GSE74310). Each dataset encapsulates the expression profiles of genes (features) across individual cellular entities (observations).

These data are organized within a compressed, sparse, columnar numerical matrix structure, wherein each column is indicative of a specific gene and each row corresponds to an individual cell. The matrices are preserved in the RDS file format, facilitating efficient data manipulation and storage.

This investigation employed four publicly accessible datasets for evaluation, specifically: Corces2016, which pertains to human immune cells; Buenrostro2018, associated with human hepatic cells; and the 10x PBMCs v1 alongside the Next Gem datasets, both of which concern peripheral blood mononuclear cells. These datasets exhibit variability in terms of tissue origin, the number of samples ranging from 575 to 4585 cells, and the sequencing methodologies utilized, namely Illumina and 10x Chromium platforms. Detailed information regarding data accessibility is encapsulated in Table 1.

In the preliminary phase of within-dataset evaluation, a 5 fold cross-validation methodology was implemented across all four datasets, employing seven distinct machine learning classification algorithms. Following this, a comparative inter-dataset assessment was executed exclusively on the 10x PBMCs datasets, selected for their uniformity in cell type and experimental protocol. Cell type annotations were assigned utilizing Seurat v3, which was selected for its demonstrated proficiency in annotating single-cell ATAC-seq data by leveraging corresponding single-cell RNA-seq data and their annotations (refer to Table 1).

Among the seven methodologies evaluated, Support Vector Machines (SVM) consistently demonstrated superior performance relative to the other techniques across all datasets. On the other hand, the K-Nearest Neighbors (KNN) algorithm exhibited the least effective performance, irrespective of whether 9 or 50 nearest neighbors were utilized. It is noteworthy that our proposed method, Extreme Randomized Trees (ERT), exhibited performance metrics comparable to those of SVM, while significantly reducing both training and testing computation time, in some instances by as much as 259%.

For instance, during the second phase of our experimental protocol, which entailed the comparative analysis of the 10× PBMCs Next Gem and 10× PBMCs v1 datasets, the median F1 score across the seven cell types within these datasets was approximately 0.786 for SVM and 0.729 for ERT. Importantly, SVM accomplished cell identification in 36 Sec, whereas ERT achieved this in 9 Sec. Additionally, an examination of the accuracy metrics for these two methods within the same experimental context revealed that SVM attained an accuracy of 0.866, compared to 0.831 for ERT, indicating a marginal accuracy differential of 0.03.

In examining the performance metrics of the SVM and ERT methods on a singular dataset, specifically the 10 × PBMCs v1, we observe that the median F1 score for the SVM approach is approximately 0.892, whereas the ERT approach yields a median F1 score of 0.806, indicating a disparity of 0.086. Additionally, the accuracy percentages for the SVM and ERT methods stand at 90.5% and 86.5%, respectively, delineating a 4% differential. When considering training times, the SVM method processes each data segment in approximately 26 Sec while the ERT method completes classification tasks in 6 Sec.

The results are compared with the MLP neural networks and the proposed ERT achives better results compared to MLP. It can be deducted that the MLP neural network is not compatible with this type of data and predicts less accurately. This is due to the fact that the amount of data in this field is usually low due to the high cost of sampling, production and labeling. Therefore, the amount of data required for training neural networks with high accuracy is very hard to collect.

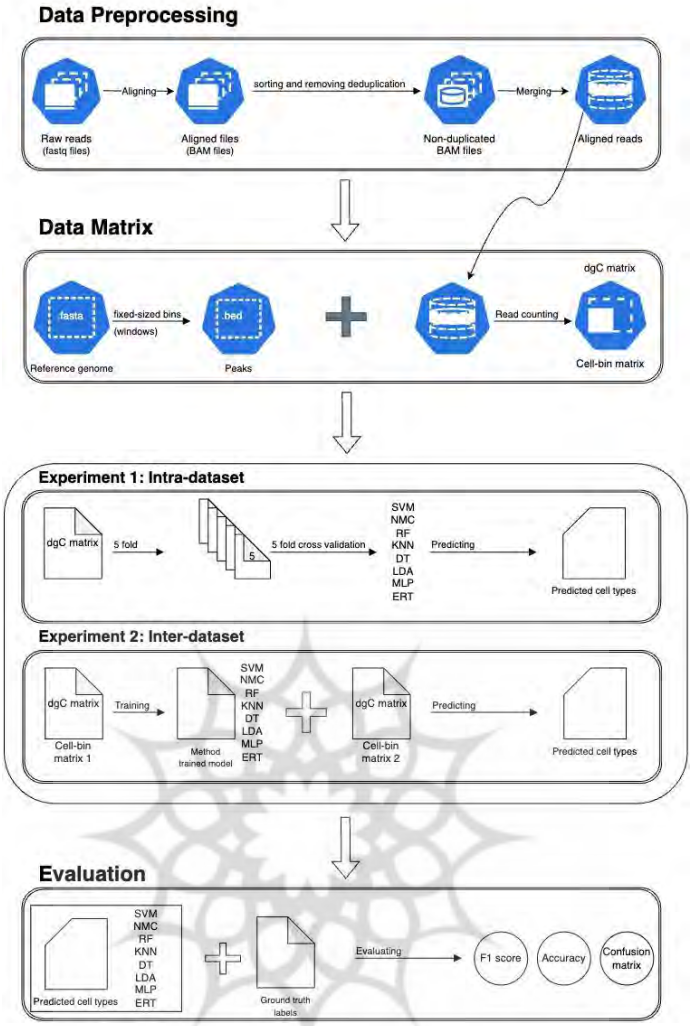

Figure. 1. The initial step involves aligning the raw sequencing reads to the reference genome utilizing the BWA software, resulting in the generation of a BAM file that contains the aligned reads. Subsequently, this BAM file undergoes sorting and deduplication processes to eliminate redundant reads, thereby enhancing the efficiency of subsequent analyses. 2) Following alignment, the BAM files from all samples are amalgamated into a single, sorted BAM file to streamline downstream analytical procedures. A list of fixed-size genomic windows, each centered on a transcription start site (TSS), is then generated. The construction of the cell-by-bin matrix ensues, achieved by enumerating the reads from each sample that fall within each predefined window. This matrix is subsequently stored in the dgCMatrix class. 3) The ensuing phase involves predictive tasks conducted through two distinct scenarios: A) Assessing the performance of various methods in inter-dataset experiments, which entail comparing the predictive capabilities of each method between 10X PBMCs v1 and 10X PBMCs Next Gem datasets. B) Conducting intra-dataset experiments through five-fold cross-validation within each dataset. 4) The efficacy of all seven methods will be evaluated by employing metrics such as the F1 score and the confusion matrix.

Table 1. This table summarizes the scATAC-seq dataset used to evaluate seven machine learning methods for cell type identification. The dataset includes approximately eleven thousand cells from up to ten different cell types, sequenced with three different technologies. Most of the cells are blood cells.

Dataset	***Number of cells***	***Number of populations***	***Description***	***Protocol***	***Ref.***
Corces2016	575	4	Human immune cell	Illumina NextSeq 500	[32]
Buenrostro2018	2034	10	Human hematopoietic system cell	Illumina NextSeq 500	[33]
10X PBMCs v1	3917(2927 labeled)	7	Peripheral blood mononuclear cell	10X Chromium Next GEM Single Cell ATAC Reagent Kits v1.1	[34]
10X PBMCs Next Gem	4585 (3670 labeled)	7	Peripheral blood mononuclear cell	10X Chromium Single Cell ATAC Reagent Kits	[35]

It is important to underscore that the objective of presenting these examples from the initial and subsequent stages of the experiment is to scrutinize the efficacy of these two methodologies when confronted with substantial data volumes, thereby subjecting their performance to rigorous evaluation. (Refer to Table 2 for detailed metrics).

In another scenario designed to test the classification robustness, we reduced the amount of input data for training data to three levels. In this experiment, the data was considered as training data at 90%, 80%, and 70% levels to measure the impact of data scarcity on different methods. This experiment was conducted to compare our proposed method, ERT, with the SVM method, which had the best performance when using all the data. The details are shown in Table 2. In summary, our proposed method showed better resistance to data reduction in classifying blood cells, maintaining its normal speed and accuracy. In contrast, the SVM method demonstrated its robustness in other cell data.

It is noteworthy that several methodologies employed in this study, specifically NMC, RF, and ERT, demonstrated performance metrics—namely, accuracy and F1 score—comparable to those achieved by SVM. In certain instances, these methods even surpassed the performance of SVM. Conversely, the methodologies DT, LDA, and KNN exhibited inferior performance. These experimental evaluations were conducted on the Google Collaboratory Pro platform with 355 GB of RAM and on the Google processor units.

5. Experimental Setup

The initial phase of our methodology encompasses preprocessing single-cell data and establishing a cross-validation framework tailored for machine learning analysis, executed through R scripting. The preliminary operations entail the ingestion of datasets, wherein cell population labels are sourced from a CSV file, and pre-processed single-cell data is retrieved from an R data file. To maintain data integrity, the script meticulously filters both datasets, ensuring the inclusion of only those cells that are present in both datasets. Subsequently, the script delineates a function designated as "Cross Validation". This function is pivotal in configuring a framework that partitions the data into training and testing subsets for the purpose of cross-validation.

It permits the user to designate the specific column within the label data that delineates the cell population classification level for subsequent analysis. Moreover, the function excludes cell populations comprising fewer than 10 cells, thereby omitting statistically insignificant groups from the analysis.

The principal operation of the function is predicated on the implementation of stratified K-Fold cross-validation. This method delineates five distinct folds, meticulously ensuring that each fold preserves an equivalent distribution of cell types as observed in the entire dataset. The function systematically iterates through each fold, generating discrete lists that encompass indices for both training and testing datasets, which are subsequently utilized in machine learning models. Ultimately, the cross-validation configuration specifics are archived in an R data file, intended for utilization in the classification phase. During the classification phase, requisite libraries for data manipulation, machine learning, and interfacing with R are imported. Additionally, a function is delineated to facilitate the conversion of sparse matrices from R format to Python format, thereby ensuring compatibility.

The described function is a sophisticated tool designed to facilitate a machine learning classification task on single-cell data. It requires paths to data files, label files, cross-validation configuration files, and an output directory. The function initiates by extracting cross-validation settings and subsequently reads and filters the data and labels according to configurations generated by R. Following this, the data undergoes normalization via a logarithmic transformation.

The machine learning model employed is the Extreme Randomized Trees (ERT) algorithm, an ensemble method that aggregates multiple decision trees. The learning process of these trees is modulated through specific hyper parameters, detailed in Table 3. This Python-based code serves as a wrapper, orchestrating the classification task by leveraging predefined configurations and systematically saving the results for subsequent analysis.

In the final stage, an R script is utilized to evaluate the performance of the machine learning classifier on the single-cell data. This script ingests true labels (ground truth) and predicted labels from separate CSV files. It also has the capability to focus on specific data subsets through provided indices. The core component, the "evaluate" function, processes the data to construct a confusion matrix, which clarifies the frequency of correct class predictions. Furthermore, it computes various performance metrics, manages unclassified predictions, and determines F1 scores for each class alongside overall accuracy.

Additionally, the script calculates the proportion of unlabeled cells and assesses the population size for each class. The results of this evaluation are meticulously organized and saved in distinct directories for confusion matrices, F1 scores, population sizes, and summary statistics. Each directory contains CSV files, ensuring that the evaluation data is readily accessible for further in-

depth analysis of the classifier's performance. Finally, some performance details of the three chosen methods (SVM, ERT, MLP) in all four datasets are gathered in Figure 2 to Figure 5. This platform will aim to facilitate the real-time analysis of a vast array of patient-related information from a single sampling event, serving both clinical and research purposes. Additionally, it is recommended that researchers engaged in chip development prioritize the creation of chips capable of receiving data from non-invasive samples (e.g., saliva), thereby advancing the development of these MIoT edge devices.

Table 2. The comparison between our suggested method and best performing method on inter and intra datasets.

Predicting Method	*Dataset Type*	*Dataset Name*	*Median F1 Score*	*Accuracy*	*Training Time (Per Fold)*
SVM	Inter-dataset	10X PBMCs Next Gem and 10X PBMCs v1	0.786	0.866	36s
	Intra-dataset	10X PBMCs v1	0.892	0.905	26s
ERT	Inter-dataset	10X PBMCs Next Gem and 10X PBMCs v1	0.729	0.831	9s
	Intra-dataset	10X PBMCs v1	0.806	0.865	~6s

Table 3. Used hyperparameters with their description and applied values for "ert"

Hyperparameter	*Value*	*Description*
n_estimators	10	This specifies the number of decision trees to be included in the ensemble. Increasing this value generally leads to a more complex model and potentially better performance, but also increases training time.
max_depth	None	This parameter controls the maximum depth of each individual decision tree. A deeper tree can potentially learn more complex patterns in the data, but also risks overfitting. Setting it to None allows the trees to grow freely until they reach a stopping criterion.
min_samples_split	2	This hyperparameter determines the minimum number of samples required to split a node in the decision tree. Higher values can prevent overfitting by avoiding splitting nodes with very few data points
random_state	0	This sets a random seed for the algorithm, ensuring reproducibility of the results when the code is run multiple times.

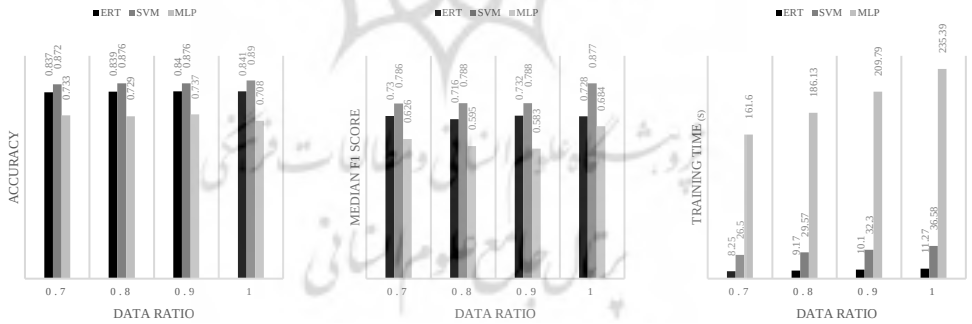

Figure. 2. Experimental results for Intra-10xPBMCsNextGem

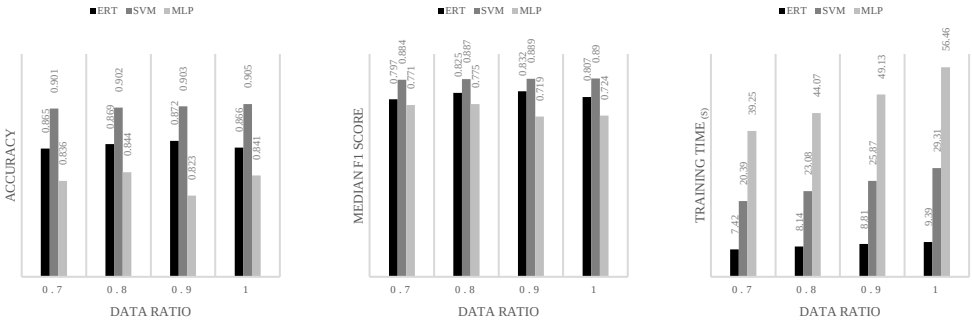

Figure. 3. Experimental results for Intra-10xPBMCsV1

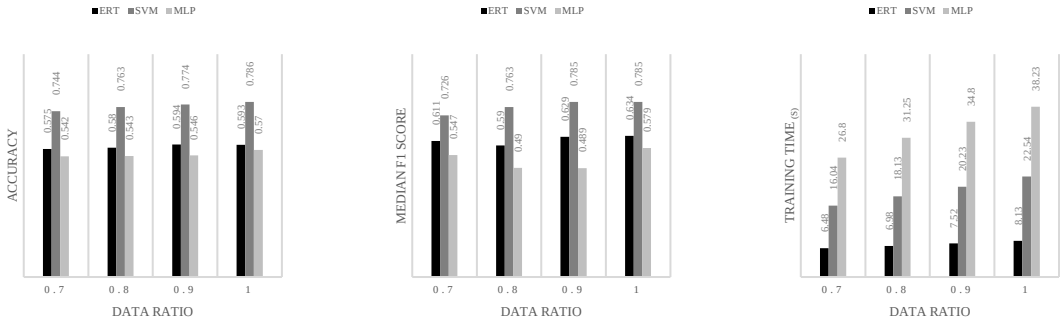

Figure. 4. Experimental results for Intra-Buenrostro2018

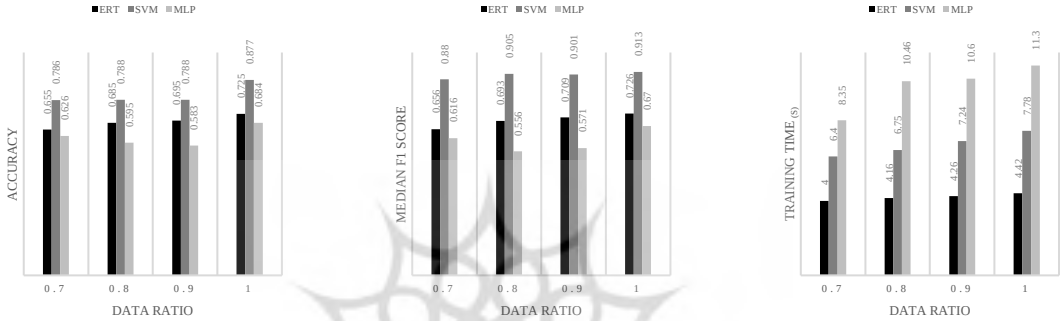

Figure. 5. Experimental results for Intra-Corces2016

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

HH: Study design, acquisition of data, interpretation of the results, statistical analysis, drafting the manuscript;

NE: Study design, interpretation of the results, drafting the manuscript, revision of the manuscript;

KK: Study design, drafting the manuscript;

BM: Study design, interpretation of the results, statistical analysis, drafting the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

[1] A. U. Ibrahim, F. Al-Turjman, Z. Sa'id, and M. Ozsoz, "Futuristic CRISPR-based biosensing in the cloud and internet of things era: an overview," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35143-35171, 2022/10/01 2022, https://doi.org/10.1007/s11042-020-09010-5.

[2] D. F. Parks *et al.*, "Internet of Things Architecture for Cellular Biology," *bioRxiv*, p. 2021.07.29.453595, 2022, https://doi.org/10.1101/2021.07.29.453595.

[3] I. Ugandran *et al.*, "A novel cryptosystem using DNA sequencing and contextual array splicing system for Medical Internet of Things," *Computers & Electrical Engineering*, vol. 96, p. 107429, 2021/12/01/ 2021, doi: https://doi.org/10.1016/j.compeleceng.2021.107429.

[4] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature methods*, vol. 10, no. 12, pp. 1213-1218, 2013, https://doi.org/10.1038/nmeth.2688.

[5] D. A. Cusanovich *et al.*, "Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing," *Science (New York, N.Y.)*, vol. 348, no. 6237, pp. 910-914, 2015, https://doi.org/10.1126/science.aab1601.

[6] J. D. Buenrostro *et al.*, "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, no. 7561, pp. 486-490, 2015, https://doi.org/10.1038/nature14590.

[7] C. A. Lareau *et al.*, "Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility," *Nature biotechnology*, vol. 37, no. 8, pp. 916-924, 2019, https://doi.org/10.1038/s41587-019-0147-6.

[8] A. N. Schep, B. Wu, J. D. Buenrostro, and W. J. Greenleaf, "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data," *Nature methods*, vol. 14, no. 10, pp. 975-978, 2017, https://doi.org/10.1038/nmeth.4401.

[9] C. G. de Boer and A. Regev, "BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization," *BMC bioinformatics*, vol. 19, no. 1, pp. 1-13, 2018, https://doi.org/10.1186/s12859-018-2255-6.

[10] Z. Ji, W. Zhou, and H. Ji, "Single-cell regulome data analysis by SCRAT," *Bioinformatics (Oxford, England)*, vol. 33, no. 18, pp. 2930-2932, 2017, https://doi.org/10.1093/bioinformatics/btx315.

[11] Zamanighomi *et al.*, "Unsupervised clustering and epigenetic classification of single cells," *Nature communications*, vol. 9, no. 1, p. 2410, 2018, https://doi.org/10.1038/s41467-018-04629-3.

[12] D. A. Cusanovich *et al.*, "A single-cell atlas of in vivo mammalian chromatin accessibility," *Cell*, vol. 174, no. 5, pp. 1309-1324. e18, 2018, https://www.cell.com/cell/fulltext/S0092-8674(18)30855-9

[13] C. Bravo González-Blas *et al.*, "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data," *Nature methods*, vol. 16, no. 5, pp. 397-400, 2019, https://doi.org/10.1038/s41592-019-0367-1.

[14] H. A. Pliner *et al.*, "Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data," *Molecular cell*, vol. 71, no. 5, pp. 858-871. e8, 2018, https://www.cell.com/molecular-cell/fulltext/S1097-2765(18)30547-1.

[15] J. R. Sinnamon *et al.*, "The accessible chromatin landscape of the murine hippocampus at single-cell resolution," *Genome research*, vol. 29, no. 5, pp. 857-869, 2019, https://doi.org/10.1101/gr.243725.118.

[16] S. M. Baker, C. Rogerson, A. Hayes, A. D. Sharrocks, and M. Rattray, "Classifying cells with Scasat, a single-cell ATAC-seq analysis tool," *Nucleic acids research*, vol. 47, no. 2, p. e10, 2019, https://doi.org/10.1093/nar/gky950.

[17] R. Fang *et al.*, "Comprehensive analysis of single cell ATAC-seq data with SnapATAC," *Nature communications*, vol. 12, no. 1, p. 1337, 2021, https://doi.org/10.1038/s41467-021-21583-9.

[18] H. Haririmonfared, N. Elmi, K. Kavousi, and B. Majidi, "Improving the Efficiency of Early Cancer Detection using Single-Cell ATAC-seq Data for Internet of Biomedical Things," in *2024 10th International Conference on Web Research (ICWR)*, Tehran, Iran, IEEE, 2024, pp. 105-111, https://doi.org/10.1109/ICWR61162.2024.10533327.

[19] L. Xiong *et al.*, "SCALE method for single-cell ATAC-seq analysis via latent feature extraction," *Nature communications*, vol. 10, no. 1, p. 4576, 2019, https://doi.org/10.1038/s41467-019-12630-7.

[20] Y. Cao *et al.*, "SAILER: scalable and accurate invariant representation learning for single-cell ATAC-seq processing and integration," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i317-i326, 2021, https://doi.org/10.1093/bioinformatics/btab303.

[21] W. Tan and Y. Shen, "Multimodal learning of noncoding variant effects using genome sequence and chromatin structure," *Bioinformatics*, vol. 39, no. 9, p. btad541, 2023, https://doi.org/10.1093/bioinformatics/btad541.

[22] T. Jing, "Unsupervised Deep Topology Embedded Characterization of Single-Cell Chromatin Accessibility Profiles," in *Proceedings of the 2024 3rd Asia Conference on Algorithms, Computing and Machine Learning*, 2024, pp. 315-322, https://doi.org/10.1145/3654823.3654881.

[23] P. Ding, Y. Wang, X. Zhang, X. Gao, G. Liu, and B. Yu, "DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape," *Briefings in bioinformatics*, vol. 24, no. 4, p. bbad231, 2023, https://doi.org/10.1093/bib/bbad231.

[24] A. Ramakrishnan, G. Wangensteen, S. Kim, E. J. Nestler, and L. Shen, "DeepRegFinder: deep learning-based regulatory elements finder," *Bioinformatics Advances*, vol. 4, no. 1, p. vbae007, 2024, https://doi.org/10.1093/bioadv/vbae007.

[25] W. Ma, J. Lu, and H. Wu, "Cellcano: supervised cell type identification for single cell ATAC-seq data," *Nature Communications*, vol. 14, no. 1, p. 1864, 2023/04/03 2023, https://doi.org/10.1038/s41467-023-37439-3.

[26] M. J. Regner *et al.*, "Defining the Regulatory Logic of Breast Cancer Using Single-Cell Epigenetic and Transcriptome Profiling," (in eng), *bioRxiv*, Jun 17 2024, https://doi.org/10.1101/2024.06.13.598858.

[27] C. Chen *et al.*, "Application of ATAC-seq in tumor-specific T cell exhaustion," *Cancer Gene Therapy*, vol. 30, no. 1, pp. 1-10, 2023/01/01 2023, https://doi.org/10.1038/s41417-022-00495-w.

[28] Y.-H. Cheng *et al.*, "Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells," *Nature Communications*, vol. 10, no. 1, p. 2163, 2019, https://doi.org/10.1038/s41467-019-10122-2.

[29] R. C. Zieren, P. J. Zondervan, K. J. Pienta, A. Bex, T. M. de Reijke, and A. D. Bins, "Diagnostic liquid biopsy biomarkers in renal cell cancer," *Nature Reviews Urology*, vol. 21, no. 3, pp. 133-157, 2024/03/01 2024, https://doi.org/10.1038/s41585-023-00818-y.

[30] F. Shi *et al.*, "A Microfluidic Chip for Efficient Circulating Tumor Cells Enrichment, Screening, and Single-Cell RNA Sequencing," (in eng), *Proteomics*, vol. 21, no. 3-4, p. e2000060, Feb 2021, https://doi.org/10.1002/pmic.202000060.

[31] W. Xu *et al.*, "A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility," *Nature Protocols*, vol. 16, no. 8, pp. 4084-4107, 2021, https://doi.org/10.1038/s41596-021-00583-5.

[32] M. R. Corces *et al.*, "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution," *Nature genetics*, vol. 48, no. 10, pp. 1193-1203, 2016, https://doi.org/10.1038/ng.3646.

[33] J. D. Buenrostro *et al.*, "Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation," *Cell*, vol. 173, no. 6, pp. 1535-1548, 2018, https://www.cell.com/cell/fulltext/S0092-8674(18)30446-X.

[34] 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (v1.0) [Online] Available: https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_pbmc_5k_v1

[35] 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor (Next GEM v1.1) [Online] Available: https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_pbmc_5k_nextgem?

[36] H. Guo, Z. Yang, T. Jiang, S. Liu, Y. Wang, and Z. Cui, "Evaluation of classification in single cell atac-seq data with machine learning methods," *BMC bioinformatics*, vol. 23, no. Suppl 5, p. 249, 2022, https://doi.org/10.1186/s12859-022-04774-z.

[37] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3-

42, 2006/04/01 2006, https://doi.org/10.1007/s10994-006-6226-1.

Hossein Haririmonfared was born in Tehran, Iran, in 1997. He received the B.Sc. from Islamic Azad University in chemical engineering and M.Sc. degrees in computer engineering from the Khatam University, Tehran, in 2021 and 2024, respectively. His main focus is on applications of machine learning in the Omics data analysis.

Nasser Elmi is a Ph.D. candidate in bioinformatics at the University of Tehran. He has over five years of experience in Next Generation Sequencing (NGS) data analysis, machine learning, and computational biology. He mainly focused on the applications of machine learning and computational biology approaches to studying the mechanism of complex diseases from different Omics perspectives.

Kaveh Kavousi received the BS degree in computer engineering, hardware from the Amirkabir University of Technology, Iran, in 1994, and the MS and PhD degrees in computer engineering, artificial intelligence, and robotics from the University of Tehran, in 2001 and 2012, respectively. Currently, he is an Associate Professor and founder of Complex Biological systems and Bioinformatics (CBB) in the Department of Bioinformatics, at the Institute of Biochemistry and Biophysics (IBB), University of Tehran. His research interests are mainly in machine learning, bioinformatics, and complex biological systems.

Babak Majidi was born in Tehran, Iran, in 1977. He received the B.Sc. and M.Sc. degrees in computer engineering from the University of Tehran, Tehran, in 2000 and 2003, respectively, and the Ph.D. degree in computer engineering from the Swinburne University of Technology, Melbourne, Australia, in 2013. From 2014 to 2021, he was an Assistant Professor with the Department of Computer Engineering, Khatam University, Tehran. Since 2021, he has been an Associate Professor with the Department of Computer Engineering, Khatam University, where he is also the Director of the Smart Digital Reality Laboratory. He is the author of more than 70 research articles.